

UNIVERSIDAD DE MURCIA

Escuela de Doctorado

TESIS DOCTORAL

Propuesta de un framework para la gestión de la deriva de concepto en problemas de Machine Learning desbalanceados

Proposal of a framework for concept drift management in Machine Learning imbalanced problems

AUTOR/A Antonio Guillén Teruel

DIRECTOR/ES Juan Antonio Botía Blaya

José Tomás Palma Méndez

2026



UNIVERSIDAD DE MURCIA Escuela de Doctorado

TESIS DOCTORAL

Propuesta de un framework para la gestión de la deriva de concepto en problemas de Machine Learning desbalanceados

Proposal of a framework for concept drift management in Machine Learning imbalanced problems

AUTOR/A Antonio Guillén Teruel

DIRECTOR/ES Juan Antonio Botía Blaya
José Tomás Palma Méndez

2026

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TITULO DE DOCTOR/A

Aprobado por la Comisión General de Doctorado el 19 de octubre de 2022.

Yo, D. Antonio Guillén Teruel, habiendo cursado el Programa de Doctorado en Informática de la Escuela Internacional de Doctorado de la Universidad de Murcia (EIDUM), como autor/a de la tesis presentada para la obtención del título de Doctor/a titulada:

Propuesta de un framework para la gestión de la deriva de concepto en problemas de Machine Learning desbalanceados

y dirigida por:

D.: Juan Antonio Botía Blaya D.: José Tomás Palma Méndez D.:

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como prevé el artículo 29.8 del reglamento, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

Murcia, a 18/06/2026

(firma)

Información básica sobre protección de sus datos personales aportados:	
Responsable	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad	Gestionar su declaración de autoría y originalidad
Destinatarios	No se prevén comunicaciones de datos
Derechos	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la quinta hoja, después de la portada de la tesis presentada para la obtención del título de Doctor/a.

Firmante: ANTONIO GUILLEN TERUEL; Fecha-hora: 18/06/2026 13:04:37; Emisor del certificado: CN=AC FNMT Usuarios,OU=Ceres,O=FNMT-RCM,C=ES;

@dministración electrónica

Código seguro de verificación: RUxFMlHR-Pu1ikkOk-CGwEpGe/-dUWsrmgW

Esta es una copia auténtica imprimible de un documento administrativo electrónico archivado por la Universidad de Murcia, según el artículo 27.3 c) de la Ley 39/2015, de 1 de octubre. Su autenticidad puede ser contrastada a través de la siguiente dirección: https://sede.um.es/validador/

This page intentionally left blank.

Agradecimientos

La realización de una tesis doctoral es, ante todo, un acto colectivo que acaba disfrazado de solitario. Detrás de cada trabajo, cada página escrita y cada idea hay personas que han puesto su tiempo, conocimiento y confianza para que esta tesis salga adelante. Por ello me gustaría dedicar unas palabras a quienes han contribuido de alguna forma a su realización.

Quiero comenzar agradeciendo a mis directores, Juan A. Botía y José Palma, la oportunidad que me brindaron de embarcarme en esta aventura. Vuestra combinación de exigencia y apoyo ha sido el motor de este trabajo. Gracias por vuestra paciencia, tiempo y cercanía; es algo que siempre tendré presente y guardaré en mi memoria con cariño.

Me gustaría agradecer a Mina Ryten, mi supervisora en la University College de Londres, por transmitirme, con mucha paciencia, conocimientos sobre un tema totalmente desconocido para mí. Su tutela despertó en mí una curiosidad genuina por las enfermedades neurodegenerativas que espero mantener en el futuro.

Quisiera agradecer también a Grzegorz J. Nalepa, mi supervisor en la Jagiellonian University de Cracovia, por su apoyo y ganas de impulsar mi creatividad y colaboración a pesar de que solo he podido estar unos pocos meses. Gracias por confiar en mí y por las risas juntos.

Agradezco al Servicio Murciano de Salud, especialmente a Ramón Rodríguez Iborra, por ayudarnos y proporcionarnos datos clínicos utilizados en esta tesis.

A todos mis coautores, gracias por el esfuerzo y el tiempo, ha sido un placer trabajar con cada uno de vosotros.

A mis compañeros de facultad, Comisionistas y Al Dente Lab, por hacer los días de trabajo más ligeros con risas, charlas y quedadas extraordinarias. No hubiera sido lo mismo sin vosotros: Alejandro Cisterna, Pilar, Enrique Pérez, Juan Sánchez, Joselu, José Adrián, Fran Aguilar, Fran Abad, Juandi, Alicia, Enrique Mármol, Jaime, Ade, Andrea...

A mis compañeros de estancia, gracias por hacerlas más sencillas y por ayudar en todo lo que os era posible, no hubiera sido lo mismo sin vosotros: Maciej, Szymon, Sabri,

Natalia, Sonia, Emil, Hannah, Áine, Jon...

A mis amigos de toda la vida, por no dejarme solo y apoyarme siempre en cualquier situación: Tomás, Antonio, Francisco David, Álvaro, Dieguito, Joserra, Alfonso, Diego, Virgilio, Escudero, Enrique, Pedro...

A Mauri, Marcelo y Érika, por ser mi segunda familia y tratarme como uno más.

A Gemma, mi pareja, por estar conmigo todos los días, nunca soltarme la mano y darme paz cuando más la necesitaba. Por hacer que los días pesados sean ligeros y que los retos parezcan más pequeños. Gracias por ser mi mayor apoyo.

A mis hermanos, Álvaro y María, por vuestra alegría y por motivarme a ser cada día un poco mejor. Estoy muy orgulloso de vosotros.

Por último, a mis padres, María y Toni, gracias por todo vuestro cariño, trabajo diario y por confiar siempre en mí. Soy consciente y agradezco de corazón todos vuestros esfuerzos por hacerme llegar hasta aquí.

Resumen

A lo largo de los últimos años, las soluciones de aprendizaje automático (Machine Learning, ML) se han integrado de manera creciente en entornos complejos y dinámicos del mundo real. En el capítulo 1 de esta tesis, introducimos y contextualizamos los dos retos más fundamentales que coexisten frecuentemente en estos despliegues prácticos. Por un lado, nos enfrentamos al hecho de que los eventos de mayor interés clínico o industrial suelen ser anómalos o extremadamente raros, lo que genera un severo desbalanceo de clases en los datos. Por otro lado, observamos que el proceso subyacente que genera dicha información no es estático, sino que evoluciona con el tiempo, introduciendo lo que en la literatura se conoce como deriva de concepto (concept drift). Un ejemplo de esta intersección de problemas, que hemos utilizado extensamente a lo largo de este trabajo, es la creación de sistemas de apoyo a las decisiones clínicas durante la reciente pandemia de COVID-19. En el ámbito clínico, hemos constatado que la predicción de resultados severos sufre intrínsecamente de este desbalanceo, ya que, afortunadamente, solo una minoría de pacientes desarrolla complicaciones graves. Aunque identificar correctamente a esta minoría es nuestra prioridad absoluta, las métricas de evaluación estándar y los procesos de entrenamiento tradicionales tienden a favorecer a la clase mayoritaria. Simultáneamente, el entorno pandémico ha demostrado ser altamente no estacionario, alterando incesantemente la relación entre las características de los pacientes y sus pronósticos. En estas condiciones, los modelos históricos se degradan, y los métodos de explicabilidad pueden producir atribuciones erróneas. Para resolver estos desafíos, proponemos FILM, un marco metodológico que unifica la evaluación robusta, el aprendizaje adaptativo y la explicabilidad confiable.

Para formalizar nuestra investigación, en el capítulo 2 establecemos los objetivos principales y la metodología a seguir. Nos propusimos no solo analizar sistemáticamente el estado del arte, sino también estudiar el efecto real del desbalanceo en las métricas de evaluación, desarrollar algoritmos robustos que no dependan exclusivamente de la generación de datos sintéticos, extender estas soluciones a flujos de datos dinámicos, e investigar en profundidad la fiabilidad de las herramientas de explicabilidad como SHAP. Hemos adoptado una metodología iterativa de desarrollo y validación, probando nuestras propuestas matemáticas tanto en bases de datos clínicas reales como en bancos de datos públicos (benchmarks) estáticos y dinámicos, garantizando así la reproducibilidad

y la validez externa de nuestros hallazgos.

Posteriormente, en el capítulo 3, llevamos a cabo una revisión sistemática de la literatura centrada en cómo los enfoques de Inteligencia Artificial han manejado la deriva de concepto en escenarios pandémicos. Siguiendo las directrices PRISMA, analizamos exhaustivamente cientos de artículos científicos para comprender las tendencias actuales. A través de este análisis, identificamos dominios analíticos clave, desde la previsión de curvas epidemiológicas y la estimación de la dinámica de transmisión, hasta la vigilancia genómica y la predicción de riesgos clínicos. Constatamos una progresiva transición desde los modelos compartimentales clásicos, como los enfoques SIR o SEIR, hacia técnicas predictivas avanzadas basadas en ensambles de árboles, redes neuronales profundas y métodos probabilísticos. En este capítulo, además, categorizamos las causas subyacentes de la deriva de concepto en diversas dimensiones: factores derivados del comportamiento humano (como las políticas de confinamiento), factores impulsados por el patógeno (nuevas mutaciones y pérdida de inmunidad), cambios estructurales en la recolección de datos, y variaciones espaciales o topológicas. Sin embargo, nuestra revisión evidenció una carencia crítica: la mayoría de las aplicaciones de ML se centran en la predicción a corto plazo sin incorporar mecanismos explícitos para adaptarse a los cambios de distribución de los datos a lo largo del tiempo, lo que nos confirmó la necesidad imperiosa de desarrollar sistemas de aprendizaje dinámico.

La fundamentación empírica de nuestra investigación se detalla minuciosamente en el capítulo 4, donde describimos las fuentes de datos utilizadas. La piedra angular de nuestro trabajo práctico es una extensa base de datos del mundo real proporcionada por el Servicio Murciano de Salud (SMS). A partir de esta fuente, que incluye registros de cientos de miles de pacientes, construimos y analizamos una cohorte específica para estudiar el Long COVID-19. Dado que la persistencia de los síntomas tras la fase aguda de la infección afecta a una proporción menor de individuos, este problema nos presentó un caso de estudio ideal con un drástico desbalanceo de clases. Aplicando modelos de regresión logística y calculando las Odds Ratios Ajustadas, determinamos que factores como pertenecer al sexo femenino, experimentar una alta carga sintomática en la fase aguda (destacando el dolor de pecho, la erupción cutánea, la hiposmia y la disnea), o haber requerido hospitalización previa, elevan de manera estadísticamente significativa el riesgo de padecer secuelas a largo plazo. Además, analizamos el efecto protector de las vacunas, integrando el historial de inmunización como un motor principal de los cambios poblacionales (deriva de concepto). Los datos evidenciaron que la vacunación previa reduce drásticamente el riesgo de Long COVID-19, destacando la superioridad de las vacunas de ARNm y demostrando empíricamente que los esquemas de vacunación con tres dosis proporcionan la máxima protección. Estos descubrimientos clínicos poseen un gran valor médico por sí mismos, pero en el contexto de nuestra tesis, nos proveyeron del escenario ideal, complejo y desbalanceado, para poner a prueba nuestros algoritmos. Asimismo, en este capítulo describimos los conjuntos de datos estándar (benchmarks como Mammography, Satimage, Forest, entre otros) que utilizamos posteriormente para

validar matemáticamente nuestros métodos frente a alternativas del estado del arte.

Con el contexto clínico y los datos definidos, en el capítulo 5 abordamos matemáticamente el problema estático del desbalanceo de clases a través de la formulación de los primeros componentes del framework FILM. Inicialmente, evaluamos cómo el desbalanceo distorsiona las métricas de rendimiento más utilizadas, como la Exactitud, el F1-score, el Kappa de Cohen o el Área bajo la curva ROC (rocAUC). Para nuestra sorpresa, los análisis de correlación revelaron que casi todas estas métricas sufren de una severa dependencia respecto a la proporción de la clase minoritaria en el conjunto de entrenamiento. Para solucionar esto, hemos diseñado e implementado una nueva métrica: el Coeficiente de Integración Ajustado por el Sesgo (BAIC). Esta métrica fusiona el poder discriminativo de múltiples indicadores estándar mediante una suma ponderada. La gran innovación matemática que introducimos radica en que los pesos se asignan de forma inversa a la correlación empírica de cada métrica con la proporción minoritaria. Utilizando una función gaussiana de penalización, el BAIC castiga aquellas métricas que fluctúan artificialmente por los cambios de prevalencia y recompensa las más estables. Nuestros experimentos confirmaron de manera concluyente que el BAIC proporciona una evaluación mucho menos sesgada. En paralelo, para la fase de entrenamiento, desarrollamos el algoritmo IPIP (Identical Partitions for Imbalance Problems). Frente a técnicas ampliamente utilizadas como SMOTE, que generan datos sintéticos asumiendo el riesgo de introducir ruido y solapamiento entre clases, IPIP adopta una estrategia de remuestreo basada en ensambles. Aprovechando las propiedades de la distribución binomial, IPIP divide la clase mayoritaria y garantiza, con una alta confianza estadística, que todas las instancias de la valiosa clase minoritaria estén representadas al menos una vez en los distintos subconjuntos balanceados creados. Demostramos empíricamente que IPIP, utilizando modelos base como Random Forest o Regresión Logística, iguala o supera a las técnicas tradicionales del estado del arte, siendo especialmente útil cuando los datos son escasos y no podemos permitirnos perder información ni inventarla.

Al comprender que la realidad clínica y los procesos de producción evolucionan de forma ininterrumpida, el paso lógico natural de nuestra investigación fue extender estos métodos estáticos a flujos de datos dinámicos continuos (data streams), lo cual presentamos en el capítulo 6. En este punto de la tesis, los datos ya no son un bloque estático, sino que llegan en forma de ventanas o bloques secuenciales a lo largo del tiempo. Para mantener una evaluación justa bajo estas condiciones cambiantes, formulamos el Adaptive BAIC. Esta versión dependiente del tiempo actualiza de forma secuencial la ponderación de las métricas aplicando un retardo de un paso temporal (t-1), lo que nos permite reestimar constantemente las correlaciones entre las métricas de rendimiento y la proporción de la clase minoritaria a medida que la distribución de la población fluctúa, proporcionando una señal de monitoreo extremadamente robusta. En cuanto al modelado predictivo, propusimos la extensión IPIP-CD. Hemos diseñado IPIP-CD como un ensamble dinámico de adaptación pasiva que no depende de detectores estadísticos de deriva, los cuales suelen fracasar estrepitosamente cuando las muestras positivas son muy es-

casas. En su lugar, guiado por la métrica Adaptive BAIC, IPIP-CD evalúa continuamente el ensamble sobre los bloques de datos más recientes, incorporando nuevos submodelos entrenados con la información actual y purgando activamente aquellos componentes históricos cuyo rendimiento ha decaído de manera irrecuperable. Evaluamos exhaustivamente IPIP-CD en ocho flujos de datos dinámicos, tanto sintéticos como reales, evidenciando que nuestra solución gestiona la combinación simultánea de un desbalanceo extremo y derivas de concepto abruptas de manera mucho más eficiente que otros enfoques reconocidos en la literatura, tales como Accuracy Weighted Ensemble o Learn++ CDS. Sin embargo, a lo largo de este proyecto comprendimos que diseñar un modelo predictivo matemáticamente robusto no es suficiente si no está acompañado de una explicabilidad transparente y confiable, un requisito indispensable para la adopción de la Inteligencia Artificial en el sector médico. Movidos por esta necesidad, en el capítulo 7 investigamos empíricamente una suposición habitualmente ignorada en la literatura: la presunta estabilidad de los métodos de explicabilidad post-hoc ante las variaciones en las distribuciones de los datos. Para ello, analizamos cómo la escasez de muestras impacta específicamente en los valores SHAP (SHapley Additive exPlanations), la herramienta de interpretabilidad más popular en la actualidad. Mediante simulaciones controladas de diferentes niveles de desbalanceo, y utilizando arquitecturas que abarcaron desde ensambles de árboles hasta redes neuronales, demostramos de manera categórica que el desbalanceo de clases distorsiona estructuralmente las atribuciones de importancia de las características otorgadas por SHAP. Detectamos que esta inestabilidad se manifiesta de diferentes maneras según la naturaleza del algoritmo subyacente. En los modelos neuronales y lineales, observamos que las magnitudes absolutas de atribución de los valores SHAP fluctúan masivamente a medida que la clase minoritaria se vuelve más escasa. Por su parte, en los modelos basados en árboles de decisión, las magnitudes promedio se mantienen engañosamente estables, pero la varianza de la atribución se dispara de forma crítica. Además, calculando la entropía de Shannon sobre los rankings de importancia de las variables, evidenciamos que en escenarios de alto desbalanceo los modelos alteran de forma casi caótica las variables en las que confían para emitir sus decisiones, haciendo que las explicaciones generadas para los clínicos carezcan de fiabilidad.

Para mitigar este grave problema y devolver la confianza a los usuarios finales en entornos altamente dinámicos, en el capítulo 8 introducimos el enfoque 3P (Tres Perspectivas). Hemos diseñado el enfoque 3P como una estrategia de visualización interactiva y monitorización continua que permite detectar, diagnosticar y explicar la deriva de concepto en flujos de datos desbalanceados. Nuestra premisa central aquí es que observar únicamente las caídas en el error predictivo agregado es completamente insuficiente. El enfoque 3P propone triangular simultáneamente tres ejes de información. En primer lugar, la perspectiva de los datos, monitorizando secuencialmente las estimaciones de densidad de kernel (KDE) de las variables de entrada para capturar desplazamientos poblacionales directos. En segundo lugar, la perspectiva del modelo, rastreando las fluctuaciones en la confianza estadística de las probabilidades de predicción emitidas por el

algoritmo, utilizando elementos visuales como la opacidad o transparencia de los puntos para alertar sobre la incertidumbre interna del sistema. Y en tercer lugar, la perspectiva de la explicación, que evalúa la evolución temporal de los valores SHAP sobre un conjunto de datos de referencia constante, lo que nos permite identificar con precisión algorítmica en qué momento temporal una variable clínica concreta pierde o gana peso predictivo. Al integrar estas tres dimensiones, proporcionamos a los analistas de datos, ingenieros y profesionales de la salud una herramienta de auditoría invaluable que permite abrir la caja negra de la adaptación automática, validando de manera humana si las alteraciones que experimentan las fronteras de decisión de la IA coinciden con fenómenos físicos o epidemiológicos reales.

Finalmente, en los capítulos 9 y 10 de esta tesis, recopilamos nuestras conclusiones generales, discutimos las limitaciones inherentes a nuestros análisis y delineamos las vías de trabajo futuro, resumiendo también nuestras contribuciones científicas publicadas. En conclusión, a lo largo de este extenso trabajo hemos comprobado y demostrado empíricamente que tratar el desbalanceo de clases y la deriva de concepto como problemas estancos e independientes conduce a soluciones frágiles y sesgadas. A través de la conceptualización y desarrollo del framework FILM, la invención metodológica de la métrica BAIC y su versión temporal, el diseño de los algoritmos de ensamblaje IPIP e IPIP-CD, y la integración crítica de la monitorización 3P para salvaguardar la explicabilidad, hemos consolidado un entorno algorítmico integral. Este conjunto de herramientas está capacitado para proveer predicciones exactas, una adaptación matemática fluida, y una interpretación honesta y auditable de los datos clínicos bajo condiciones sumamente volátiles. Con la publicación y liberación del código de estos algoritmos a la comunidad científica, aseguramos que los métodos aquí desarrollados trasciendan el ámbito académico y puedan ser aplicados de manera directa y efectiva en la gestión hospitalaria de pacientes, contribuyendo significativamente a la fiabilidad y robustez de la Inteligencia Artificial en las ciencias de la salud y en investigaciones futuras.

Abstract

Modern Machine Learning (ML) systems are increasingly deployed in complex environments where two major difficulties frequently co-occur: the events of interest are rare, and the underlying data-generating process evolves over time. A paradigmatic example is clinical decision support during a pandemic, such as anticipating severe COVID-19 outcomes or Long COVID-19 development. In these scenarios, severe class imbalance causes standard training procedures to favor the majority class and makes conventional evaluation metrics highly biased and misleading. Furthermore, phenomena such as new viral variants, changing health protocols, or vaccination rollouts introduce concept drift, leading models trained on historical data to degrade silently after deployment.

Beyond predictive performance, the reliability of post-hoc explanation methods, which are essential for clinical trust and decision-making, is also compromised. Explanation tools like SHAP values can produce unstable and biased feature attributions when the positive cases are scarce. To address the complex interplay of class imbalance, concept drift, and the reliability of model explanations, this thesis proposes a comprehensive methodological approach: the Framework for Imbalanced Learning Machines (FILM).

The empirical foundation of this research relies on an extensive real-world COVID-19 registry provided by the Servicio Murciano de Salud (SMS), which was used to derive a Long COVID-19 cohort. The clinical analysis identifies significant risk factors and highlights the protective effects of specific vaccination schedules against long-term complications, treating vaccination as a main driver of distributional changes. To ensure broad generalization, these real-world records are complemented with widely used public benchmark datasets and synthetic data stream generators for controlled experimentation.

The core methodological contributions are structured around evaluation, prediction, and monitoring. First, a systematic literature review categorizes how concept drift has been addressed in pandemic modeling, identifying main adaptation strategies. Second, to tackle evaluation biases, the thesis introduces the Bias-Adjusted Integration Coefficient (BAIC). This metric integrates complementary performance indicators through a weighted sum that explicitly reduces sensitivity to the minority class proportion. Third, the Identical Partitions for Imbalance Problems (IPIP) algorithm is proposed. IPIP is

an ensemble-based strategy that constructs balanced training subsets without relying on synthetic data generation, thus avoiding the introduction of artificial noise. Both BAIC and IPIP are subsequently extended to non-stationary environments. The Adaptive BAIC adjusts metric weights along chunk-based data streams, while IPIP-CD provides a passive adaptation strategy that continuously updates the ensemble to handle concept drift under severe imbalance.

Finally, the thesis investigates model monitoring and interpretability in dynamic settings. It quantifies the specific biases introduced by class imbalance on SHAP values and proposes the 3P approach, a multi-perspective strategy that integrates data distribution shifts, explanation dynamics, and model performance to support robust drift analysis. Ultimately, this thesis establishes a structured, end-to-end framework capable of ensuring reliable evaluation, adaptation, and trustworthy interpretability for machine learning models operating in highly imbalanced and non-stationary environments.

Contents

List of Figures xvii

List of Tables xxiii

Acronyms xxiv

1 Introduction **3**

1.1 Class imbalance in Machine Learning problems 4

1.2 Concept drift and non-stationary data streams 6

1.3 Imbalanced concept drift: challenges at the intersection 8

1.4 Monitoring and explainability under drift and imbalance 9

1.5 Organisation of the thesis 11

2 Objectives & Methodology **13**

2.1 Objectives 13

2.2 Methodology 14

3 Review of the state of the art **17**

3.1 Introduction 17

3.2 Background 19

3.2.1 What is concept drift and how is it related to pandemics 19

3.2.2 Machine learning approaches for pandemics 20

3.2.3 Pandemics 22

3.2.4 Explainability and drift visualisation in data streams 23

3.3 Methods . . . 25
3.3.1 Research questions . . . 25
3.3.2 Literature search methodology used . . . 26
3.4 Results . . . 26
3.4.1 Studies selected as relevant for the review . . . 26
3.4.2 RQ1: What are the modelling tasks in pandemics related to concept drift? . . . 28
3.4.2.1 Pandemic forecasting and curve modelling . . . 31
3.4.2.2 Transmission dynamics and parameter estimation . . . 31
3.4.2.3 Genomic and antigenic surveillance . . . 32
3.4.2.4 Patient screening, diagnosis and clinical risk prediction . . . 32
3.4.2.5 Early-event detection to rapid actions . . . 33
3.4.2.6 Intervention and policy evaluation . . . 34
3.4.3 RQ2: Which methods have been applied to deal with drift? . . . 34
3.4.3.1 Compartmental methods . . . 35
3.4.3.2 Time-series specific analysis models . . . 35
3.4.3.3 Probabilistic methods . . . 36
3.4.3.4 Tree-based Methods . . . 37
3.4.3.5 Neural Networks . . . 37
3.4.3.6 Support Vector Machines . . . 38
3.4.3.7 Clustering . . . 38
3.4.3.8 Other AI-based techniques . . . 39
3.4.4 Explainable AI methods . . . 39
3.4.5 RQ3: What are the causes of drift, and how does this influence the approach used? . . . 41
3.4.5.1 Human-derived concept drift . . . 41
3.4.5.2 Concept drifts derived from changes in the pathogen . . . 42
3.4.5.3 Drift caused by structural and data gathering changes . . . 42

3.4.5.4 Causes for drift of topological nature 43
3.4.5.5 Drift due to violation of model´s basic assumptions . . . 44
3.5 Discussion 44
3.5.1 Gaps and Future Directions 46

4 Data Sources and Datasets **49**
4.1 COVID-19 and Long COVID-19 Clinical Dataset 50
4.1.1 Methods 50
4.1.1.1 Study design, setting, and participants 50
4.1.1.2 Ethics approval and consent to participate 52
4.1.1.3 Variables and data source 53
4.1.1.4 Data description and preprocessing 54
4.1.1.5 Bias 54
4.1.1.6 Statistical analyses 55
4.1.2 Results 57
4.1.2.1 Participants 57
4.1.2.2 Analysis of sociodemographic factors, comorbidities and symptoms 59
4.1.2.3 Impact of COVID-19 hospitalisation on the development of LC-19 62
4.1.2.4 Effects of vaccination on LC-19 63
4.1.2.5 LC-19 calculator tool 66
4.1.3 Discussion 67
4.2 Benchmark datasets used throughout the thesis 72
4.2.1 Datasets from stationary problems 72
4.2.2 Concept Drift Benchmark Datasets 73
4.2.3 Stream-based Datasets 74
4.3 Data Availability and Reproducibility 75

5 A Framework for Learning in Highly Imbalanced Problems 77

5.1 Motivation of the Framework 78

5.2 Previous work 81

5.3 Materials and methods 84

5.3.1 Performance metrics 84

5.3.2 The Bias-Adjusted Integration Coefficient (BAIC) method 85

5.3.3 The IPIP algorithm 87

5.4 Experimental setup 92

5.4.1 Datasets 93

5.4.2 Implementation details 93

5.4.2.1 Concordance plots between metrics: Concordance plots . 93

5.4.2.2 Correlation between metrics and proportion of the minority class $p_{\min}$ 95

5.4.2.3 Comparison of metrics biases towards the proportion of minority class $p_{\min}$ 95

5.5 Results 97

5.5.1 No agreement among metrics in deciding on the best model . . . 97

5.5.2 The proportion of the minority class biases all binary classification metrics 105

5.5.3 A weighted integration of the metrics reduces bias 110

5.5.4 Results of BAIC metric 111

5.5.4.1 Execution times analysis. 113

5.6 Concluding the chapter 114

6 Extending the Framework to Dynamic Data Streams 117

6.1 Introduction 118

6.2 Background 119

6.3 Methods 121

6.3.1 Adaptive BAIC method for dynamic data streams 121

6.3.2 The IPIP passive adaptation to concept drift . . . 122
6.4 Experiments . . . 125
6.4.1 Datasets . . . 125
6.4.2 Implementation Details . . . 125
6.4.2.1 Area under the progression of the metrics along time . . 126
6.4.2.2 Correlation between the area under the progression and the minority class proportion . . . 127
6.5 Results . . . 128
6.5.1 The AUPC value of the Adaptive BAIC progression is less biased by the minority class proportion . . . 128
6.5.2 Results on the normalised AUPC of the Adaptive BAIC metric . . 128
6.6 Concluding the chapter . . . 130

7 Explanation biases in Imbalanced Problems 133

7.1 Introduction . . . 133
7.2 Related work on the reliability of SHAP-based post-hoc explanations . . . 135
7.3 Background . . . 136
7.3.1 SHapley Additive exPlanations (SHAP) values . . . 136
7.4 Experimental set-up . . . 137
7.4.1 The four families used in the context of imbalanced problems . . . 138
7.4.2 Sampling Strategy for IR Evaluation . . . 139
7.4.3 SHAP values analysis through basic statistical parameters . . . 139
7.4.4 Model stability indicators . . . 141
7.4.4.1 Predictive Complexity Score (PCS) . . . 141
7.4.4.2 Parameter-variance indicators . . . 143
7.4.5 SHAP consistency in producing feature rankings . . . 144
7.5 Results . . . 145
7.5.1 Datasets . . . 145
7.5.2 Implementation details . . . 145

7.5.3 Magnitude and variance of SHAP 147
7.5.4 Consistency of the most important features 149
7.6 Concluding the chapter 153

8 Explainability and Visualization in Dynamic Imbalanced Problems 157

8.1 Introduction 157
8.2 Integrating data, explanation, and model performance into a 3P approach 159
8.3 Empirical studies on concept drift analysis through the lens of the 3P approach 161
8.3.1 Synthetic dataset 161
8.3.2 Cold rolling mill simulator 162
8.3.3 SMS COVID-19 dataset 165
8.4 Concluding the chapter 167

9 Conclusions 169

9.1 Review, interpretation, and implications of the results 169
9.2 Limitations 173
9.3 Future work 175
9.4 Concluding remarks 177

10 Contributions of the Thesis 179

10.1 Publications and dissemination of results 179
10.1.1 Publication as first author 179
10.1.2 Publications as co-author 180
10.1.3 Conference paper 181
10.1.4 Collaboration with the University College London 181
10.1.5 Collaboration with the Jagiellonian University 182
10.1.6 Involvement in other research projects during the PhD 183

Bibliography 185

List of Figures

1.1 Illustrative example of misleading evaluation under class imbalance: two models applied to the same imbalanced dataset can achieve similar accuracy while exhibiting different behaviour on the minority class results. The two panels display confusion matrices, which compare the actual class labels (Real, y-axis) against the model's predictions (Predicted, x-axis). The main diagonal cells show correct predictions, whereas the off-diagonal cells represent classification errors. 6

1.2 Schematic illustration of performance degradation in a chunk-based streaming setting caused by a change in class proportions. The plot shows the evolution of model accuracy together with the proportion of Class 0 across sequential data chunks. When the class distribution shifts (vertical dashed line), a model trained on past data may no longer represent the current conditions, leading to a drop in predictive performance. 7

3.1 Overview of concept drift types, temporal patterns, and strategies to address them. The purple box summarises the main approaches: detection, adaptation, interpretability, and recalibration. Under an active concept drift approach, detection typically precedes the subsequent steps, while in passive approaches, the modules may operate in parallel, and detection might not be explicitly included. 21

3.2 Overview of the three RQs to their corresponding dimensions: medical tasks addressed by AI approaches (RQ1), the methods employed (RQ2), and the main causes of concept drift in pandemics (RQ3) with some examples. . . . 28

3.3 PRISMA flow diagram presenting the number of retrieved and excluded papers at each stage of the analysis. 29

3.4 Overview of the selected articles: (a) temporal distribution by publication type, and (b) journals with at least two published articles. 30

4.1 Patient Selection Flow diagram. Flow diagram of the enrollment of subjects, disposition status, and how they are analysed in the study. 58

4.2 Different types of Long COVID-19 are coded by colours. Odds ratios are adjusted for age, sex, and comorbidities. 60

4.3 A) Adjusted odds ratios for developing LC-19 based on different comorbidities. B) Adjusted odds ratios for developing LC-19 based on acute-phase COVID-19 symptoms. The female sex is coloured in red and the male sex in blue. Odds ratios are adjusted by age, the symptoms and the comorbidities. COPD is chronic obstructive pulmonary disease, AH is arterial hypertension, ICM is ischemic cardiomyopathy, DM is diabetes mellitus and RI is renal insufficiency. 61

4.4 Distribution of the number of symptoms for LC-19 (blue) vs. non-LC-19 (red) patients during primary infection. 63

4.5 Different types of Long COVID-19 are coded by colours. Odds ratios are adjusted for age, sex, and symptoms. 64

4.6 Odds ratio for suffering each categorisation of LC-19 adjusted by age with and without sex-disaggregated. 65

4.7 The dose comparison is represented as the first number vs. the second number; for example, 'Dose: 1vs0' shows the odds ratio of having 1 dose of any vaccine compared to 0 doses. Female patients are represented in red, and male patients in blue. 67

4.8 AORs for various vaccination schedules against LC-19. Vaccination schedules are represented by combinations of letters that reflects the order in which the vaccines were administered to participants. With A for AstraZeneca, P for Pfizer, M for Moderna, and J for Janssen vaccines. Blue squares represent greater protective behaviour indicated by lower AORs, when comparing the reference schedules (rows) with the comparison ones (columns), being then the comparison one more protective. In contrast, red squares signify a higher AOR, indicating less protective effect against the outcome. If the adjusted p-value for the AOR is less than 0.05, the comparison is flagged with a star (*). 68

5.1 Main modules of the FILM framework. The methodology integrates IPIP for static imbalanced learning, its dynamic extension IPIP-CD for data streams with concept drift, and the BAIC method being the model selection criterion. 78

5.2 Values of Accuracy, F1-score, and rocAUC when training a Random Forest model with upsampling on the SMS dataset, across both train and test partitions. The x-axis represents the proportion of the minority class (p_{min}), and the y-axis represents the corresponding metric value. Solid lines show the fitted linear regression for each metric, with the equation, adjusted R^2, and p-value reported in matching colour. 80

5.3 Taxonomy of the imbalance learning techniques studied in this section. . . . 83

5.4 Three different Gaussian functions. Red curve with parameters $(1, 0, 0.35)$, blue curve with parameters $(1, 0, 0.25)$ and green curve with parameters $(1, 0, 0.15)$. 87

5.5 FILM methodology for obtaining BAIC metric and choose the best model according to BAIC. In (1), n datasets are obtained from the original dataset d through sampling with several imbalance ratios. In (2), a set of algorithms is then trained on d and its n resamples, resulting in k metrics for each algorithm on each dataset. In (3), a three-dimensional matrix is formed to retain these values. In (4), we aggregate the metric results for each algorithm across all $n+1$ datasets, retaining the k metric values of all algorithms in d. This provides weights for each metric and algorithm. In (5), the weights and values of the metrics are finally aggregated over the original dataset to obtain the BAIC metric. Then it determines which algorithm has obtained the maximum BAIC value. 88

5.6 Synthetic examples of concordance plots illustrating two extreme evaluation scenarios. A) A scenario of perfect agreement forming two distinct metric clusters that show algorithms A1 and A2 as best models, respectively, with absolute discordance between them. B) A scenario of total lack of consensus, characterised by a high proportion of empty sections in the off-diagonal pie plots. 95

5.7 A) Concordance plot of the Diabetes dataset using Logistic Regression. B) Concordance plot of the Mammography dataset using Random Forest models. 98

5.8 A) Concordance plot of the Diabetes dataset using Random Forest. B) Concordance plot of the Mammography dataset using Logistic Regression models. 99

5.9 A) Concordance plot of the SMS dataset using Logistic Regression. B) Concordance plot of the SMS dataset using Random Forest models. 100

5.10 A) Concordance plot of the Forest dataset using Logistic Regression. B) Concordance plot of the Forest dataset using Random Forest models. 101

5.11 A) Concordance plot of the Satimage dataset using Logistic Regression. B) Concordance plot of the Satimage dataset using Random Forest models. . . . 102

5.12 A) Concordance plot of the Adult dataset using Logistic Regression. B) Concordance plot of the Adult dataset using Random Forest models. 103

5.13 A) Concordance plot of the Phoneme dataset using Logistic Regression. B) Concordance plot of the Phoneme dataset using Random Forest models. . . 104

5.14 Correlation heatmaps between metrics and p_{min} in Diabetes (A) and Mammography (B) datasets. 106

5.15 Correlation heatmaps between metrics and pmin in SMS dataset. 107

5.16 Correlation heatmaps between metrics and pmin in Forest dataset. 107

5.17 Correlation heatmaps between metrics and pmin in Satimage dataset. 108

5.18 Correlation heatmaps between metrics and pmin in Adult dataset. 108

5.19 Correlation heatmaps between metrics and pmin in Phoneme dataset. 109

5.20 Boxplots of correlations of the different metrics across all experiments ordered by median from highest to lowest. Above each boxplot is the p-value of the Wilcoxon's Test vs BAIC correlation values. 110

6.1 Overview of the Adaptive BAIC computation in a chunk-based data stream. At each time step t, a vector of performance measures $\mathbf{m}(t) = (m_1(t), \ldots, m_k(t))$ is obtained (bottom-left). For the same time step, metric values are compared across a set of fixed imbalance settings $\{p_1, \ldots, p_Q\}$ to estimate the correlation $r_j(t)$ between each metric $m_j(t)$ and $p_{\min}$ (middle). These correlations are mapped through a Gaussian function to produce metric weights $w_j(t)$ (right). The integrated score is then computed online using a one-step lag, combining current metrics with the previous weights via a dot product, yielding $\text{Adaptive BAIC}(t) = \sum_{j=1}^{k} w_j(t-1)\, m_j(t)$ (bottom-right). 123

6.2 Boxplots of Pearson's correlation values between the area under the temporal progression of different metrics and $\mathbf{p_{min}}$ across all streaming experiments. Metrics are ordered by median from highest to lowest. Above each boxplot is the p-value resulting from the Wilcoxon signed-rank test comparing that metric against the Adaptive BAIC correlation values. 129

7.1 Overview of the experimental framework that has been utilised in order to analyse the effect of class imbalance on SHAP values for each individual dataset. 138

7.2 SHAP distribution for each model and IR for the (a) SMS dataset and (b) diabetes dataset. 148

7.3 SHAP distribution for each model and IR for the Forest dataset 150

7.4 SHAP distribution for each model and IR for the Mammography dataset . . 150

7.5 SHAP distribution for each model and IR for the Satimage dataset 151

7.6 Entropy of the most important feature sets across different IRs for the (a) SMS dataset and (b) diabetes dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR. 152

7.7 Entropy of the most important feature sets across different IRs for the Forest dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR. 153

7.8 Entropy of the most important feature sets across different IRs for the Mammography dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR. 153

7.9 Entropy of the most important feature sets across different IRs for the Satimage dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR. 154

8.1 Flowchart of 3P approach. The data stream is divided into chunks, and each chunk is analyzed from three key perspectives—data, explanation, and model perspective—such that the data perspective focuses on tracking the Kernel Density Estimation of test data across chunks, the explanation perspective involves monitoring changes in feature importance across successive models while maintaining a fixed dataset, and the model perspective captures variations in prediction probabilities across successive models while using a fixed dataset. 158

8.2 Example of visualization of proposed 3P approach. 160

8.3 Proposed visualization for AGRAWAL data stream. The plots correspond to abrupt concept drift. 163

8.4 Proposed visualization for AGRAWAL data stream. The plots correspond to gradual concept drift. 164

8.5 Proposed visualization for tcm5 dataset. The top plot corresponds to feature force_1, and the bottom one to torque_1. 165

8.6 Proposed visualization for tcm5 dataset. The top plot corresponds to feature motor_power_3, and the bottom one to motor_power_4. 166

8.7 Proposed visualization for SMS COVID-19 dataset. The top plot corresponds to the chest pain symptom, and the bottom one to the number of systems affected by comorbidity. 167

8.8 Proposed visualization for SMS COVID-19 dataset. The top plot corresponds to the age feature, and the bottom one to the vomiting symptom. 168

List of Tables

3.1 Keywords categorised by topic. 27

3.2 Summary of studies depending on the medical task. 29

3.3 Summary of Compartmental methods. 35

3.4 Summary of time-series analysis methods. 36

3.5 Summary of probabilistic methods. 36

3.6 Summary of tree-based methods. 37

3.7 Summary of neural network methods. 38

3.8 Summary of Support Vector methods. 38

3.9 Summary of clustering methods. 39

3.10 Summary of Other AI-based techniques. 40

3.11 Summary of explainable AI methods. 41

3.12 Summary of studies depending on the cause of the CD. 41

4.1 Demographic characteristics and comorbidities for COVID-19 patients (Non-LC-19) and different subtypes of long COVID-19 patients. Continuous data are reported as median with interquartile range (Q3–Q1), and categorical data are expressed as percentages (%). COPD is chronic obstructive pulmonary disease and ICM is Ischemic cardiomyopathy. 59

4.2 Symptomatology for COVID-19 patients (Non-LC-19) and different subtypes of long COVID-19 patients. Categorical data are expressed as percentages (%). 62

4.3 Table of vaccination differences for non LC-19 and LC-19 patients. 66

4.4 Datasets from stationary systems used throughout the thesis, including the number of instances, IR, attributes, and detail on categorical labels. 73

4.5 Streaming datasets and generators used in the drift evaluations, detailing dimensions and chunk counts. 74

5.1 Equations of evaluation metrics for binary classification. 85

5.2 Values given to create the grid of hyperparameters for random forest models in all experiments for each dataset. 93

5.3 Distance matrix between correlation vectors for each metric and the unbiased correlation vector in each dataset. BAIC obtained the lowest distance value for all datasets in comparison to the remaining metrics. 106

5.4 BAIC values of the performance of the different models on the seven datasets. 112

5.5 Elapsed mean time in seconds taken to train a 5-fold cross-validation of each model ordered by column Mean, which is the mean time taken to train each model across all datasets. 113

6.1 Values used to create the hyperparameter grid for Random Forest base learners across all streaming datasets. 126

6.2 Global performance comparison using the normalised AUPC of the Adaptive BAIC progression curve. Values are reported with three decimal places followed by their relative rank within each dataset (1^{o} to 6^{o}). The highest performing model for each stream is highlighted in bold. 130

7.1 Balanced Accuracy (BalAcc [$CI_{95\%}$]) for each dataset, model, and IR. 146

7.2 Spearman correlation between IR and the mean/variance of |SHAP| for each dataset and model. Significance levels: *** for $p-value < 0.001$, ** for p-value < 0.01, and * for $p-value < 0.05$. 149

Acronyms

ACM	Association for Computing Machinery.
ADASYN	Adaptive Synthetic Sampling.
ADWIN	Adaptive Windowing.
AH	Arterial Hypertension.
AI	Artificial Intelligence.
AIDS	Acquired Immunodeficiency Syndrome.
AIRA	Artificial Intelligence Research and Applications.
ANN	Artificial Neural Network.
AOR	Adjusted Odds Ratio.
ARIMA	AutoRegressive Integrated Moving Average.
AUPC	Area under the Performance Curve.
AWE	Accuracy Weighted Ensemble.
BAIC	Bias-Adjusted Integration Coefficient.
CALM	Concept Drift-Aware Machine Learning.
CAM	Class Activation Mapping.
CD	Concept Drift.
CDC	Centers for Disease Control and Prevention.
CDS	Concept Drift with SMOTE (Learn++.CDS).
CI	Confidence Interval.
CNN	Convolutional Neural Network.
COBYLA	Constrained Optimization BY Linear Approximation.
COPD	Chronic Obstructive Pulmonary Disease.
COVID-19	Coronavirus Disease 2019.
CT	Computed Tomography.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise.
DDM	Drift Detection Method.
DL	Deep Learning.
DM	Diabetes Mellitus.
DNA	Deoxyribonucleic Acid.

DOI Digital Object Identifier.

ECML-PKDD European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.
EDDM Early Drift Detection Method.

FILM Framework for Imbalanced Learning Machines.
FILM-CD FILM for Concept Drift (R package).
FN False Negative.
FP False Positive.
FPR False Positive Rate.

GDPR General Data Protection Regulation.
GEIST Group for Engineering of Intelligent Systems Technologies.
GISAID Global Initiative on Sharing All Influenza Data.
GLM Generalised Linear Model.
GMean Geometric Mean.
GRU Gated Recurrent Unit.

H1N1 Influenza A virus subtype H1N1.
HICD Hybrid drift algorithm.
HIV Human Immunodeficiency Virus.

IAA Imbalance-Aware Approaches.
IASS Imbalance-Aware Stratified Sampling.
ICM Ischemic Cardiomyopathy.
ICU Intensive Care Unit.
IEEE Institute of Electrical and Electronics Engineers.
IPIP Identical Partitions for Imbalance Problems.
IPIP-CD Identical Partitions for Imbalance Problems for Concept Drift.
IQR Interquartile Range.
IR Imbalance Ratio.
ISS Instance Selection Strategies.

JAHCAI Jagiellonian Human-Centred AI Lab.

KDE Kernel Density Estimation.
KNN k-Nearest Neighbours.

LC-19 Long COVID-19.
LIME Local Interpretable Model-agnostic Explanations.
LSTM Long Short-Term Memory.

MCC Matthews Correlation Coefficient.
ML Machine Learning.
MLOps Machine Learning Operations.
MLP Multilayer Perceptron.
MOA Massive Online Analysis.
mRNA messenger Ribonucleic Acid.

OLA Online Learning Algorithms.
OR Odds Ratio.

PCR Polymerase Chain Reaction.
PCS Predictive Complexity Score.
PD Parkinson's Disease.
PH Page-Hinkley test.
PLR Polynomial Logistic Regression.
PPMI Parkinson's Progression Markers Initiative.
prAUC Area Under the Precision-Recall Curve.
PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
PROVIA Prognosis prediction in COVID-19 patients based on Artificial Intelligence.

RAM Random Access Memory.
RF Random Forest.
RGPD Reglamento General de Proteccion de Datos.
RI Renal Insufficiency.
RLOG Logistic Regression (base learner).
RNA Ribonucleic Acid.
RNN Recurrent Neural Network.
ROC Receiver Operating Characteristic.
rocAUC Area Under the Receiver Operating Characteristic Curve.
ROSE Random Over-Sampling Examples.
RQ Research Question.
RSMOTE Robust SMOTE.
RT-PCR Reverse Transcription Polymerase Chain Reaction.

SALSA Search, Appraisal, Synthesis, and Analysis.
SARIMAX Seasonal AutoRegressive Integrated Moving Average with eXogenous variables.
SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2.
SEA Streaming Ensemble Algorithm.
SEAIHRD Susceptible-Exposed-Asymptomatic-Infectious-Hospitalised-Recovered-Deceased.
SEIQR Susceptible-Exposed-Infectious-Quarantined-Recovered.
SEIR Susceptible-Exposed-Infectious-Recovered.

SEIRD Susceptible-Exposed-Infectious-Recovered-Deceased.
SEIRDV Susceptible-Exposed-Infectious-Recovered-Deceased-Vaccinated.
SEIRV Susceptible-Exposed-Infectious-Recovered-Vaccinated.
SEQIR Susceptible-Exposed-Quarantined-Infectious-Recovered.
SEQIRD Susceptible-Exposed-Quarantined-Infectious-Recovered-Deceased.
SERA SElectively Recursive Approach.
SGWT Spectral Graph Wavelet Transform.
SHAP SHapley Additive exPlanations.
SIR Susceptible-Infectious-Recovered.
SIRD Susceptible-Infectious-Recovered-Deceased.
SJR Scimago Journal Rank.
SMOTE Synthetic Minority Over-sampling Technique.
SMOTEBoost SMOTE with Boosting.
SMS Servicio Murciano de Salud.
STROBE Strengthening the Reporting of Observational Studies in Epidemiology.
SVIR Susceptible-Vaccinated-Infectious-Recovered.
SVM Support Vector Machine.
SVR Support Vector Regression.

TCM Tandem Cold Rolling Mills.
TN True Negative.
TP True Positive.
TWC Topological Weighted Centroid.

UB Uncorrelated Bagging.
UCI University of California, Irvine (Machine Learning Repository).

VMP Variance of the Model Parameters.

XAI Explainable Artificial Intelligence.
XGBoost Extreme Gradient Boosting.

1

Introduction

Modern Machine Learning (ML) solutions are increasingly deployed in settings where two practical difficulties co-occur: the events of interest are rare, and the process that generates the data evolves over time. A clear example of this is the development of clinical decision support systems during a pandemic, such as those used for COVID-19 triage. In this context, triage involves using predictive models to anticipate disease severity and allocate medical resources. When tackling this problem, fewer than 10% of patients may actually require hospitalisation, making these cases rare but a priority to classify correctly. Meanwhile, successive pandemic waves, changes in testing protocols, and the rollout of vaccination campaigns continuously alter both the patient population and the relationship between clinical features and outcomes [1, 2]. Under these conditions, standard predictive performance measures, such as overall accuracy, can be misleading, models trained on historical data can degrade silently, and explanation methods that are key to the trustworthiness of models and the guidance provided to healthcare workers' actions may yield wrong signals. This thesis addresses the interplay of these three fundamental challenges: (1) class imbalance, (2) concept drift, and (3) the reliability of post-hoc explanations.

Then, one of the common difficulties in applications of binary classification is that the outcomes of interest are rarely balanced [3, 4]. The minority class typically corresponds to critical events, such as severe clinical conditions or system failures, while the majority class represents normal conditions. Under this imbalance, model training can be dominated by the majority class, and performance assessment can become misleading if it relies on measures that are sensitive to the class distribution [5]. As a result, selecting an inadequate evaluation criterion may lead to models that appear reliable but fail on the cases of the minority class (positive cases), that in some cases are the ones that matter the most. In medical applications, this translates into missed diagnoses: a model with 99% accuracy on a dataset with 1% positive cases can achieve that score by predicting every patient as negative, thereby producing a 100% false-negative rate on the class of interest. This motivates the need for evaluation measures that remain informative regardless of

the class distribution [5, 6].

Another practical difficulty is that data collected over time do not remain invariant [7, 8]. Changes in protocols, interventions, population behaviour, environmental conditions, and acquisition processes can alter the patterns that models rely on. When this happens, models trained on historical data may degrade after deployment, as a consequence of alterations to the underlying processes [9]. This degradation can be difficult to interpret if monitoring is based solely on aggregated performance indicators. These issues are particularly challenging when changes interact with imbalance, since small shifts affecting the observations from the minority class may be masked by stable behaviour on the majority class instances.

Beyond prediction, this thesis also considers how models are analysed and monitored in practice. Modern machine learning workflows frequently use feature importance tools to support understanding, debugging, and communication of results. Among these, SHAP values have become widely adopted due to their flexibility across model families. However, explanation outputs are not immune to the structure of the data: when positive cases are scarce, attribution patterns can be unstable or biased, and their interpretation requires careful experimental validation. For this reason, the thesis combines methodological developments with empirical studies that examine the reliability of performance assessment and explanation tools under imbalance and changing conditions.

The empirical studies are supported by a real-world COVID-19 registry provided by the *Servicio Murciano de Salud* (SMS), available through a research project already finished, and complemented with widely used benchmark datasets and standard experimental generators. This combination makes it possible to ground the work in realistic case studies while also supporting controlled experiments that allow general conclusions beyond a single application domain. The specific objectives and the methodology guiding this research are detailed in Chapter 2, while the complete description of these data sources is provided in Chapter 4.

The remainder of this chapter is organised by topic. We first introduce class imbalance in machine learning problems and discuss its implications for model training and evaluation. We then present the background required to understand learning under changing data distributions and the challenges that arise when both issues co-occur. Finally, we motivate the scope of the thesis and summarise the organisation of the manuscript.

1.1 Class imbalance in Machine Learning problems

Class imbalance is a common characteristic of real-world classification tasks [3, 10], particularly in binary problems where one class represents normal or frequent conditions and the other corresponds to rare but critical events. In these settings, the number of ob-

servations from the minority class can be orders of magnitude smaller than the number of observations from the majority class, which affects both how models are trained and how their performance is assessed. Since many applications prioritise the correct identification of rare events, an adequate treatment of imbalance is required to avoid solutions that appear accurate while failing on the cases of interest.

From a learning perspective, imbalance modifies the effective training signal received by the model. Standard training procedures tend to favour hypotheses that optimise the global empirical risk, which can be dominated by the majority class simply due to frequency. As a result, a classifier may achieve low overall error by focusing on the observation patterns of the majority class, while providing limited sensitivity to the minority class examples. This behaviour is not necessarily caused by a lack of modelling capacity, but by an optimisation objective that is not aligned with the operational goal. The issue becomes more pronounced when the minority class is heterogeneous, when decision boundaries are complex, or when the available minority examples do not sufficiently represent the underlying variability.

Imbalance also has direct implications in model evaluation. Many widely used measures implicitly weight classes according to their prevalence, and therefore can produce high scores even when the minority class is poorly detected. A simple example is shown in Figure 1.1, where two models applied to the same imbalanced dataset can yield similar accuracy while behaving very differently on the minority class examples. Model A achieves a slightly higher overall accuracy by predicting almost all instances as the majority class, but completely fails to identify any minority-class cases. In contrast, Model B correctly identifies a portion of the instances from the minority class while maintaining a similar accuracy. Therefore, when the goal is to correctly predict both classes, and especially when detecting the minority class instances is important, Model B provides a much more informative and useful predictive behaviour. In practice, this can lead to conflicting conclusions when comparing models, as different measures emphasise different trade-offs between false positives and false negatives. Consequently, evaluation is not a mere reporting choice: it influences model selection, drives methodological conclusions, and may determine whether a model is considered suitable for deployment.

Several families of strategies have been proposed to address imbalance. At the data level, resampling techniques adjust the class distribution by oversampling the minority class, undersampling the majority class, or generating synthetic minority examples. At the algorithmic level, cost-sensitive learning and class-weighted objectives explicitly penalise minority errors more heavily. At the decision level, threshold moving and calibration-based approaches modify the operating point to prioritise minority detection according to application requirements. Although these approaches can be effective, their behaviour is closely tied to the chosen evaluation criteria and to the availability and representativeness of minority examples. This motivates the need for evaluation measures that remain meaningful under varying imbalance ratios and for learning strategies that

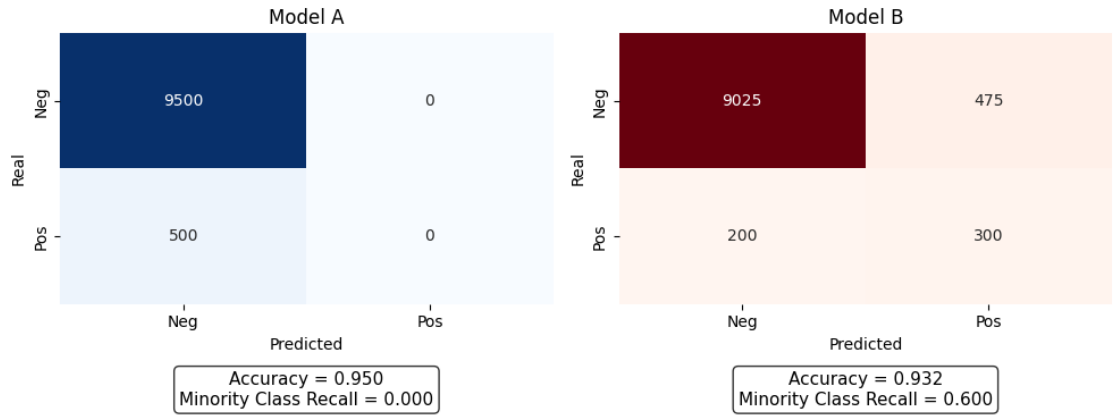


Figure 1.1: *Illustrative example of misleading evaluation under class imbalance: two models applied to the same imbalanced dataset can achieve similar accuracy while exhibiting different behaviour on the minority class results. The two panels display confusion matrices, which compare the actual class labels (Real, y-axis) against the model's predictions (Predicted, x-axis). The main diagonal cells show correct predictions, whereas the off-diagonal cells represent classification errors.*

improve the detection of examples from the minority class without introducing artefacts that compromise generalisation.

The remainder of the thesis builds upon this observation. Chapter 5 analyses the dependence of standard evaluation measures on the imbalance ratio and proposes a bias-adjusted performance measure together with an ensemble-based learning strategy designed for highly imbalanced data. In subsequent chapters, the thesis extends these ideas to settings where data evolve over time, showing that imbalance-related issues become even more critical when model validity must be assessed continuously under changing conditions.

1.2 Concept drift and non-stationary data streams

Many machine learning models are developed under the implicit assumption that training and deployment data are generated by the same underlying process. In practice, this assumption is frequently violated. Data collected over time may change due to external interventions, updates in acquisition protocols, population shifts, seasonal effects, or changes in operational conditions. When learning systems are deployed in such environments, their validity cannot be treated as a static property: performance can deteriorate after deployment, and the causes of degradation may not be obvious when monitoring relies on aggregated indicators. A schematic example is shown in Figure 1.2, where model accuracy drops sharply after a distributional change.

These scenarios are naturally described through the notion of a *data stream*, where observations arrive sequentially and the model may need to be evaluated and updated continuously. In streaming settings, storing all historical data is often impractical, and

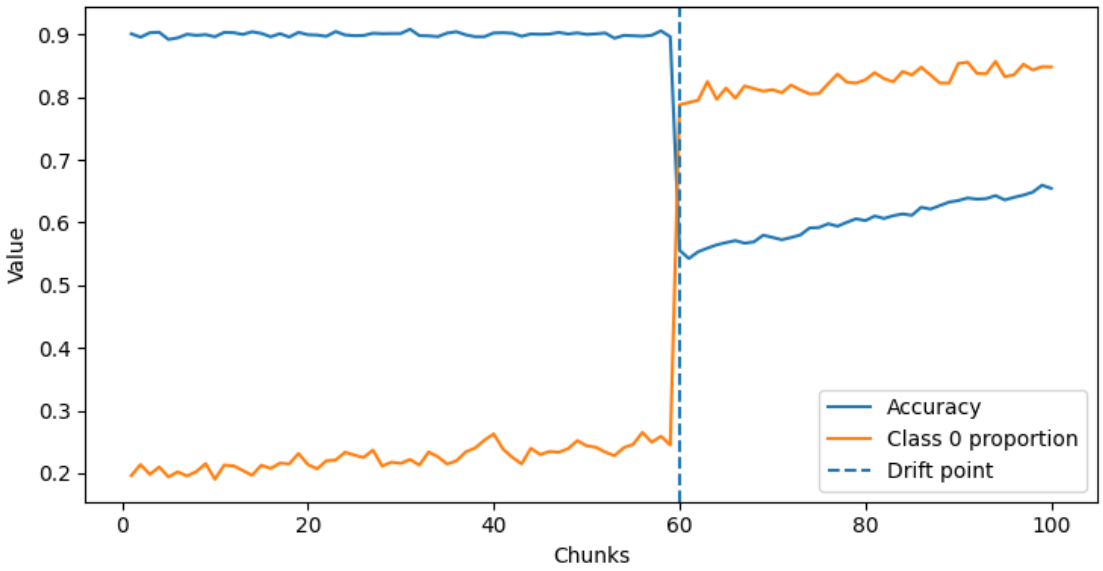


Figure 1.2: *Schematic illustration of performance degradation in a chunk-based streaming setting caused by a change in class proportions. The plot shows the evolution of model accuracy together with the proportion of Class 0 across sequential data chunks. When the class distribution shifts (vertical dashed line), a model trained on past data may no longer represent the current conditions, leading to a drop in predictive performance.*

learning must be performed under constraints such as limited memory, restricted computation, and strict requirements on reaction time. A common practical approach is to process the stream in *chunks* (i.e., sequential batches of incoming data) of consecutive instances, allowing the model to be assessed and updated periodically while preserving the temporal structure of the problem.

Within this context, *concept drift* refers to changes in the data distribution that alter how a previously learned model behaves over time [1]. A useful way to characterise these changes is by considering where the shift occurs: changes in the marginal distribution of inputs $P(X)$, changes in the class prior $P(Y)$, or changes in the conditional distribution $P(Y|X)$. These cases are often associated with different practical consequences. For instance, a shift in $P(X)$ may indicate that incoming data are drawn from a different population or operating regime, while a change in $P(Y|X)$ implies that the decision boundary itself may have evolved and that previously learned patterns are no longer predictive.

Concept drift can also be described by its temporal dynamics. Some changes occur abruptly, producing an immediate drop in performance, whereas others appear gradually or incrementally, potentially masking degradation when results are averaged over long periods. Recurring patterns may also emerge when regimes repeat over time, which makes it possible to reuse previously learned models or to design strategies that preserve knowledge across cycles. Because these behaviours affect how quickly a model should adapt and how monitoring should be interpreted, drift characterisation is a central ele-

ment of experimental design in streaming studies.

The presence of changing data distributions motivates learning strategies that go beyond a single training phase. In the concept drift literature, two complementary families of approaches are commonly distinguished, **active** strategies, which explicitly detect change and trigger an adaptation action, and **passive** strategies, which continuously adapt the model without declaring drift events [7, 11]. Active approaches incorporate a drift detector that monitors a signal (often the online error rate, a statistic of the input distribution, or the model confidence) and raises an alarm when a statistically significant change is detected. Once an alarm is triggered, the learner may be reset, retrained using recent data only, or updated with an adaptive window that discards outdated observations. Classical examples include DDM, which monitors the online error rate and signals warning and drift levels to guide model rebuilding, and ADWIN, which maintains a variable-length window and shrinks it when changes are detected [2, 12]. Passive approaches, in contrast, avoid explicit alarms and instead rely on mechanisms that gradually prioritise recent information, such as sliding windows, fading factors, or ensembles with dynamic weighting and replacement of weak members. These methods can be more stable in noisy environments and can adapt smoothly to gradual changes, at the cost of potentially reacting slower to abrupt shifts. A representative example is Dynamic Weighted Majority, an ensemble method that updates expert weights online and adds or removes experts as performance changes[13]. Both families are typically evaluated with sequential protocols that track performance over time (e.g., chunk-based or prequential evaluation), enabling the analysis of degradation and recovery patterns across the stream [7].

In this thesis, learning under changing data distributions provides the context in which static assumptions are relaxed and continuous assessment becomes necessary. Chapter 3 surveys how these issues have been addressed in pandemic-related modelling and identifies common sources of change and adaptation strategies reported in the literature. Later chapters build on these foundations by studying how evaluation and learning should be designed when data evolve over time, and by introducing monitoring and analysis tools to support the interpretation of model behaviour under changing conditions. The next section discusses why these challenges become particularly demanding when distributional change is combined with class imbalance, motivating the need for unified evaluation and adaptation principles.

1.3 Imbalanced concept drift: challenges at the intersection

Class imbalance and changing data distributions are often studied separately, yet they frequently co-occur in practical deployments. When the minority class corresponds to rare and critical events, even moderate changes in the data generating process can have a disproportionate impact on the operational value of the model (i.e., its practical use-

fulness for real-world decision-making). In this setting, the central difficulty is not only that performance may degrade over time, but also that the degradation can be difficult to detect and to interpret when the evaluation and monitoring procedures are sensitive to class prevalence.

A first challenge concerns performance assessment. In streaming evaluation, performance is typically tracked across time windows or chunks. However, when the class distribution is highly skewed, common measures may be dominated by the majority class, and short-term evaluations can become unstable due to the small number of minority examples available in each chunk. As a consequence, performance curves may appear stable while minority detection deteriorates, or they may fluctuate substantially due to sampling variability even when the underlying predictive ability does not change. This makes it difficult to distinguish between true distributional change and noise induced by rare events, and it complicates model comparison and selection over time.

A second challenge concerns adaptation. Many update strategies implicitly assume that recent data provide a representative view of the current concept. Under imbalance, recent chunks may contain very few positive examples, which can lead to updates that reinforce patterns from the overrepresented class while failing to incorporate the structure of the minority class instances. Conversely, aggressive resampling or weighting can overreact to a small number of minority observations, leading to unstable decision boundaries and poor generalisation. Therefore, updating a model under drift requires balancing two competing needs: responsiveness to change and stability when learning from underrepresented class observations. The extension of our proposed methodology to address these dynamic data streams is presented in Chapter 6.

A third challenge concerns monitoring and diagnosis. In non-stationary environments, practitioners often need to answer questions such as whether a performance drop is caused by a genuine change in the predictive relationship, by a shift in class prevalence, or by changes in the input distribution that affect the model indirectly. When imbalance is present, these questions become harder to address, because aggregate indicators may hide failures specific to observations from the minority class and because diagnostic signals can be distorted by the scarcity of positive cases. In this context, monitoring should integrate multiple views of the system, including performance measures that remain meaningful under skewed class distributions and complementary signals that help interpret how the model and the data evolve.

1.4 Monitoring and explainability under drift and imbalance

In applied machine learning, deployment is not the end of the modelling process. Once a model is integrated into a decision pipeline, its behaviour must be monitored to ensure that performance remains acceptable and that failures are detected early. This require-

ment becomes especially relevant in settings where data evolve over time and where the minority class corresponds to critical events. In such cases, monitoring based solely on a single aggregate measure can be insufficient, since stable global performance may hide a progressive deterioration on the minority class, and short-term fluctuations may be driven by sampling variability rather than by genuine changes.

Monitoring therefore involves more than periodically reporting a score. It requires tracking model behaviour over time, identifying when performance changes are likely to be meaningful, and providing diagnostic signals that support interpretation. In streaming settings, this typically includes performance curves computed on consecutive chunks, analysis of class prevalence, and inspection of changes in the distribution of relevant variables. However, even when such indicators are available, they may not fully explain *why* performance changes occur or which aspects of the data the model relies on in different periods.

This is where model explainability becomes relevant. Model explainability refers to the ability to describe how an ML model produces its outputs. In many practical workflows, practitioners complement predictive results with *feature importance* analyses to support understanding, debugging, and communication of model behaviour. Feature importance can be used to validate whether a model relies on plausible factors, to detect spurious associations, and to assess whether the model's reasoning remains consistent over time. When combined with temporal monitoring, explanation signals can also help diagnose changes in the predictive relationship by revealing shifts in which variables drive the model decisions.

Nevertheless, explanation tools are not immune to the structure of the data. Under class imbalance, feature importance estimates can be unstable or biased because the minority class is under-represented, and the model may optimise primarily for the majority class. In addition, when the data distribution changes, feature importance profiles may change for multiple reasons: the model may adapt, the input distribution may shift, or the relationship between predictors and the label may evolve. Without a careful protocol, these changes can be misinterpreted as meaningful insights, or conversely, real changes can remain unnoticed if explanations are aggregated without accounting for time and class prevalence.

For these reasons, monitoring and explainability should be treated as complementary components of a reliable learning pipeline in dynamic imbalanced settings. In this thesis, we study both aspects. Chapter 7 specifically investigates the impact of class imbalance on SHAP-based feature importance. Building on this, Chapter 8 introduces a multi-perspective approach to monitor changes over time by integrating signals from data, performance, and explanation dynamics. Together, these contributions aim to support not only accurate prediction, but also trustworthy monitoring and analysis when models operate under imbalance and changing conditions.

1.5 Organisation of the thesis

This thesis is structured into ten distinct chapters.

Chapter 2 states the objectives of the thesis and outlines the thesis-level methodological principles that guide the empirical studies, including the general evaluation perspective adopted across static and streaming settings.

Chapter 3 presents a systematic literature review of Artificial Intelligence approaches for dealing with concept drift in pandemics. This chapter synthesises the main medical problems addressed in the literature, the characteristics of the datasets used, the modelling families applied, and the sources of change and adaptation strategies reported across studies.

The results section commences in Chapter 4, which describes the data sources used throughout the thesis. It includes the clinical COVID-19 registry provided by the *Servicio Murciano de Salud* and the derived Long COVID-19 cohort, and it summarises the benchmark datasets and data streams used in the experimental chapters.

Chapter 5 analyses the effect of class imbalance on performance assessment and proposes a bias-adjusted evaluation approach designed to reduce sensitivity to the imbalance ratio. The chapter also introduces and evaluates an ensemble-based learning strategy for highly imbalanced classification, reporting results on clinical and benchmark datasets under static conditions.

Chapter 6 extends the framework to dynamic environments where data arrive sequentially and model validity may change over time. It studies evaluation under imbalanced streams, proposes robust assessment principles for this setting, and introduces an adaptation strategy designed for concept drift in imbalanced data streams, supported by experiments on real and synthetic streams.

Chapter 7 analyses how class imbalance can bias SHAP-based feature importance, discussing implications for interpretability practice in imbalanced learning.

Chapter 8 focuses on monitoring and analysis under imbalance and changing conditions. It presents a multi-perspective approach to study temporal change by integrating signals from data, performance, and feature importance dynamics.

In Chapter 9, the main conclusions of the thesis are summarised with respect to the stated objectives, discuss limitations, and outline directions for future research.

In Chapter 10, the main scientific contributions, publications, and dissemination activities derived from this thesis are detailed.

The final section is dedicated to the bibliography, which comprises all references cited throughout the thesis.

2

Objectives & Methodology

2.1 Objectives

The overall purpose of this thesis is to advance the development and evaluation of reliable machine learning systems in challenging real-world environments where two problems frequently co-occur: (i) severe class imbalance, which distorts performance assessment and training dynamics, and (ii) non-stationary data distributions, which lead to concept drift and progressive performance degradation over time. Pandemic scenarios are a paradigmatic example of this, as the events of interest are often critical yet rare (e.g., severe COVID-19 clinical outcomes) and the underlying data-generating mechanisms evolve continuously due to viral variants, external interventions, or shifts in healthcare protocols. In these medical contexts, model development cannot be separated from the way performance is assessed and monitored, as standard evaluation practices and static learning procedures lead to dangerously erroneous conclusions under imbalance, and fail entirely when concept drift occurs. Consequently, this study not only tackles predictive adaptation but also examines the reliability of model interpretation, focusing particularly on explanation methods based on feature attributions.

From this perspective, the research provides both methodological and empirical contributions. Methodologically, it analyses how imbalance influences model assessment and decision-making, proposes robust evaluation measures and learning strategies tailored to imbalanced settings, and extends these ideas to streaming environments affected by concept drift through a unified framework for evaluation, training, and monitoring. Empirically, it studies the behaviour of commonly used analysis tools in these scenarios, examining how imbalance can bias SHAP-based feature-attribution patterns and how explanation dynamics can support drift analysis. The proposed methodology is validated primarily using an extensive real-world COVID-19 registry (See Chapter 4), complemented with standard benchmark datasets and stream generators. This formulation ensures clinical relevance while allowing controlled experiments that support generali-

sation.

Ultimately, this work seeks to establish a structured, end-to-end framework capable of navigating imbalanced clinical streams affected by distributional change, and to validate this framework across real-world epidemiological registries and standard benchmarks. To achieve this goal, the thesis pursues the following specific objectives:

1. To systematically analyse how concept drift and non-stationary distributions have been addressed in medical and epidemiological modelling, identifying current limitations and adaptation strategies in the literature.
2. To study the effect of severe class imbalance on performance assessment, highlighting the limitations and disagreements of widely used evaluation measures.
3. To formulate and validate robust evaluation frameworks capable of mitigating the bias introduced by class imbalance in binary classification problems.
4. To contribute with new adaptive learning approaches for environments characterised by severe class imbalance and non-stationary data distributions for both static and streaming problems.
5. To assess the reliability of model interpretability and monitoring in dynamic imbalanced scenarios, identifying potential biases in feature importance methods and providing visual and analytical tools for robust drift analysis.
6. To disseminate the research findings and developed tools to the scientific community through peer-reviewed publications and international conferences.

Having defined the objectives of the thesis, the next section outlines the thesis-level methodology adopted to develop and validate the proposed contributions, and Chapter 4 describes the datasets used across the experimental studies.

2.2 Methodology

In the context of this thesis, the term *methodology* refers to the overarching research design shared across the different studies, rather than the detailed methods of each individual chapter. The thesis follows a development-and-validation approach: we first formalise the problem setting and the limitations of existing evaluation or learning practices, then propose methodological contributions, and finally validate them through extensive experimentation on both clinical and benchmark data.

To ensure consistency and rigour across all the empirical studies, this research design relies on four principles:

1. **Problem settings and scope:** The empirical studies in this thesis focus on binary classification problems, where class imbalance is explicit and performance assessment is particularly sensitive to minority prevalence. When benchmark datasets

are originally multi-class, they are transformed into binary settings following the protocols described in the corresponding experimental chapters. Two evaluation regimes are considered: a static regime, where data are assumed to be drawn from a stationary distribution, and a streaming regime, where data arrive sequentially in chunks and distributional change may occur over time.

2. **Experimental design and evaluation principles:** For static experiments, models are trained and evaluated following standard supervised learning protocols, while ensuring consistent preprocessing and comparable experimental conditions across datasets and baselines. For streaming experiments, models are evaluated sequentially using chunk-based processing, enabling the analysis of temporal performance evolution and the study of adaptation strategies under drift. Across all experiments, evaluation is based on multiple complementary measures, and results are summarised using comparable aggregation criteria to support fair method comparison.
3. **Statistical and analytical reporting:** Whenever appropriate, the thesis complements predictive results with statistical analyses suited to the study setting, including the reporting of effect sizes or odds ratios in clinical analyses and the use of non-parametric comparisons in experimental evaluations. These analyses are used to support the interpretation of observed differences and to reduce the risk of drawing conclusions from performance variations that may be explained by sampling effects.
4. **Reproducibility and research integrity:** All studies follow consistent data handling and experimental reporting practices, including explicit documentation of dataset variants, preprocessing decisions, and experimental protocols. Clinical data are handled in anonymised form, and ethical and legal aspects are addressed in the dataset description chapter.

In conclusion, after thoroughly reviewing the state-of-the-art and identifying the key research gaps that this thesis aims to address, it is first necessary to describe the data that will be utilised in our analyses. Therefore, the next chapter provides a detailed description of the various datasets employed across the different experiments of this thesis.

3

Review of the state of the art

3.1 Introduction

Machine Learning is a discipline within Artificial Intelligence (AI) that encompasses a family of computational techniques that learn patterns from data and generate predictions or decisions. Examples of these techniques include deep learning (DL), Bayesian machine learning and reinforcement learning. ML has become a central approach within software that needs to predict from complex, rapidly changing data [14]. The importance of ML is particularly evident in Biomedicine, where the timely acquisition and processing of relevant information can be critical. DL models are created from a subset of ML methods that use multiple layers of information processing through artificial neural networks to learn hierarchical representations from data automatically. For example, such models can analyse computed tomography (CT) images from patients with suspected or confirmed cases of SARS-CoV-2 infection and identify potentially malignant nodules [15]. Bayesian models are capable of forecasting hospital bed demand with an anticipation of several days [16]. Reinforcement-learning agents have demonstrated their capacity to design vaccine-distribution strategies that optimise both epidemiological impact and cost [17]. The nature of data surrounding a pandemic episode is inherently multimodal due to its intrinsic complexity. The capacity of ML approaches to integrate multimodal data, such as electronic health records, medical images, mobility traces, and even social media signals, makes it possible to convert noisy observations into probability-based predictions. Therefore, ML is a valuable tool to create continuously updated decision support tools for clinicians and policymakers [18].

This review focuses on how AI approaches, mainly ML methods, are applied to pandemic scenarios. Pandemics are characterised by the widespread transmission of infectious pathogens, such as viruses, leading to substantial increases in morbidity and mortality across extensive geographical regions. They usually impose considerable economic, societal and political strain on the affected populations and represent a signif-

icant application area for ML methods [19]. Pandemic scenarios often exhibit significant variability during the event in the number of affected individuals, transmission rates, and, consequently, in the utilisation of medical resources. This causes abrupt policy shifts, such as lockdowns or travel restrictions, potentially resulting in alterations to transmission dynamics [20]. Pandemic data refers to time-stamped, multimodal observations that resemble the circulation of pathogens in a target population. These observations encompass the level of infection within the population, along with demographic attributes, mobility patterns, information on healthcare resources, occupation, and intervention logs. Models built on these inputs are typically used to anticipate disease evolution by forecasting future incidence, hospital demand or mortality. The utilisation of AI, particularly ML-based prediction models, in the domain of an epidemiological event aims to facilitate the early detection of emerging threats [21, 22], forecasting resource requirements [23], optimising resource allocation in real time [24], and the refinement of vaccination schedules[25, 26]. A pandemic episode can occur over a potentially long period. And its rapidly changing features along the episode demand ML model production and maintenance pipelines capable of continuously adapting to change to control mortality and to avoid severe societal and economic disruption [27].

It is therefore evident that changes in pandemic features constitute a central challenge for the management of ML models throughout the course of a pandemic and beyond. Within ML, a change in the process being modelled that may affect model performance is named concept drift, as it was previously introduced in Chapter 1 (Section 1.2). Let X denote the inputs to the model and Y the outcome. In the modelling process, the aim is to characterise the conditional distribution $P(Y|X)$, which formally describes the probability of the outcome $P(Y)$ given the input variables. We use concept drift to refer to any temporal change in either the distribution of $P(X)$, $P(Y)$, or $P(X|Y)$[1] (see 3.2 for a detailed definition). For example, the emergence of a novel virus genetic variant has the potential to dramatically increase the transmissibility rate of the corresponding disease within a matter of days [9]. An appropriate management of vaccination campaigns has demonstrated to result in a steady decrease in the susceptible population, thereby ensuring that it remains below critical thresholds [28, 29]. Such examples of drift will likely affect the performance of models that were trained with data gathered before the drift happened. It is crucial to detect, quantify and correct any possible drift [8]. Practical responses to combat this, from an ML perspective, include the retraining of models using recent data after detecting the drift [30], the addition of adaptive layers to machine learning models to autonomously detect and correct the drift [31], and using ensemble-based ML models to dynamically vary the basic model population depending on data drifts [7, 32, 33]. Concrete measures that policymakers may take include the imposition of mask usage, the enforcement of stay-at-home orders, and the adjustment of political strategies to address outbreaks before hospitals reach maximum capacity [34]. In this context, it is crucial to move beyond static ML models towards truly adaptive decision-support systems capable of handling continuously evolving pandemic conditions, irrespective of the

origin of the underlying change.

The possible application of AI-based approaches, particularly ML, to obtain robust and lasting pandemic models has received little systematic attention. On the one hand, ML models work perfectly well in the generation of short-term predictions. On the other hand, the highly multi-factorial changing dynamics of pandemics compose high-impact scenarios that become challenging problems for classic ML approaches. By incorporating concept drift into the ML model production pipeline, we can interactively assess the validity of model predictions in the face of evolving conditions and react to change [35]. Notwithstanding, we can hardly find ML reviews on the topic. Epidemiological surveys addressing the use of computer science to manage pandemics usually include limited coverage of contemporary ML techniques, neglecting the consideration of drift. And studies specifically addressing drift rarely take into account the particular characteristics of infectious disease data found in pandemics. To the best of our knowledge, no previous work has combined current knowledge on concept drift within ML with pandemic data in a comprehensive manner. Consequently, analysing the use of ML approaches in pandemics, from a concept drift perspective, is necessary.

The contributions of this review are summarised together with the remaining contributions in Chapter 9.

The rest of the chapter is organised as follows. In Section 3.2, we introduce the background material of the review: we describe what concept drift is, its relation to pandemics, AI methods commonly used in pandemic problems, and pandemics. In Section 3.3, we explain the methodology we have used to select the papers included in this survey, the research questions we face and the survey's scope. In Section 3.4, we show the results obtained within the papers we have included in the review. The discussion and future directions of the chapter are shown in Section 3.5.

3.2 Background

3.2.1 What is concept drift and how is it related to pandemics

Formally, concept drift (CD) refers to the phenomenon where the statistical properties of the underlying data distributions within the target process change over time, impacting the reliability and accuracy of predictive models. It can be defined as any temporal alteration in (1) the joint probability distribution $P(X, Y)$, (2) in the distribution of the input features $P(X)$, (3) in the distribution of the target variable $P(Y)$ or (4) in the conditional distribution $P(Y \mid X)$[1]. Some authors distinguish between real drift (alterations in $P(X \mid Y)$) and virtual drift (shifts in $P(X)$ without changes in $P(Y \mid X)$) [7]. In other cases, the term concept drift is used interchangeably with data drift or distribution shift, referring more broadly to changes in data distributions that affect the performance of

machine learning models [36]. In a medical context, concept drift frequently emerges due to evolving pathogen characteristics, demographic shifts, or alterations in medical practices and interventions, which may be generated by decisions made by policymakers. For example, during a pandemic, the emergence of a new virus variant can modify transmissibility (the inherent ability of a pathogen to spread between hosts) and disease severity (how severe the outcomes are in infected individuals), altering the conditional distribution $P(Y|X)$, being Y patient prognosis and X clinical and demographic features. Concept drift can manifest itself through several distinct patterns, commonly classified as abrupt, incremental, gradual, and recurrent drift [1]. *Abrupt drift* occurs when the distribution changes suddenly and significantly. *Incremental drift* is characterised by a slow and steady evolution of the data distribution. Conversely, *gradual drift* denotes fluctuations between two or more data distributions over a prolonged period of time. *Recurrent drift* arises when previously observed data distributions reappear after somewhat regular time intervals. When facing a problem potentially affected by concept drift, several methodological strategies can be employed to address it. These strategies can be applied under either an active or a passive paradigm, depending on whether drift detection is explicitly performed before model adaptation or not. Figure 3.1 provides a comprehensive overview of the main types of concept drift, their temporal patterns, and the strategies commonly employed for their detection, adaptation, interpretability, and recalibration.

3.2.2 Machine learning approaches for pandemics

AI, and particularly ML, has become an essential tool to provide a better management and mitigation of the effects of pandemics. It is currently used to assist in tasks such as outbreak prediction, forecasting the level of infection, and decision-making. Across all these tasks, ML represents the most widely adopted group of techniques due to its robust predictive capabilities, adaptability, and ability to handle large, complex datasets [37]. Among these techniques, several families of algorithms have shown prominence, including compartmental models, neural networks, support vector machines (SVM), tree-based models, probabilistic methods, and clustering algorithms.

Compartmental models constitute one of the earliest and most widely used data analysis frameworks for describing infectious-disease dynamics. These types of models do not belong to the family of ML. However, they are included here as they represent an important branch of data-driven approaches to studying pandemics. In these models, the population is divided into mutually exclusive "compartments" that represent epidemiological states (e.g. Susceptible, Exposed, Infectious, Recovered), and the time-evolution of each state is described through a system of ordinary differential equations. The classical SIR formulation by Kermack and McKendrick splits the population into Susceptible (S), Infectious (I), and Removed/Recovered (R) groups, allowing key quantities such as the basic reproduction number R_0 to be expressed in closed form and estimated from data [38, 39]. For pathogens with a latent period, the SEIR extension inserts an Exposed

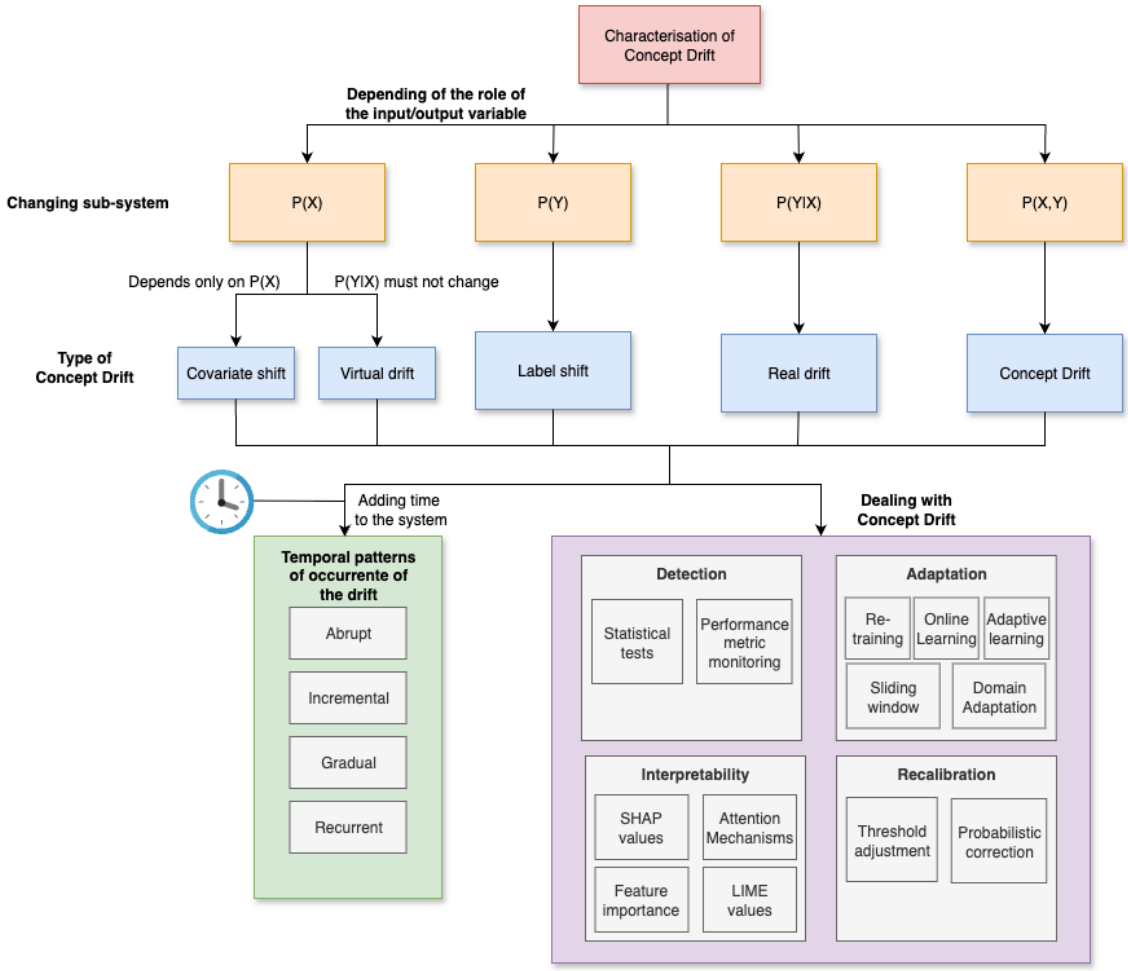


Figure 3.1: *Overview of concept drift types, temporal patterns, and strategies to address them. The purple box summarises the main approaches: detection, adaptation, interpretability, and recalibration. Under an active concept drift approach, detection typically precedes the subsequent steps, while in passive approaches, the modules may operate in parallel, and detection might not be explicitly included.*

(E) compartment between S and I. At the same time, further refinements, such as SEIQR (quarantine), or vaccine-aware models like SEIRV, have been deployed to capture features specific to COVID-19 and other recent outbreaks [40].

Neural networks, particularly deep learning techniques such as recurrent neural networks (RNN) [41], long short-term memory networks (LSTM) [42], gated recurrent units (GRUs) [43], transformers [44] and convolutional neural networks (CNN) [45], are extensively employed due to their exceptional ability to model complex, nonlinear temporal relationships prevalent in epidemiological data [46]. For instance, LSTM networks have effectively predicted future infection rates and identified temporal patterns during COVID-19 waves [47]. CNNs have also been utilised successfully in spatial analysis tasks, aiding in predicting geographic spread and identifying high-risk regions [48].

Support vector machines [49] represent another robust predictive tool widely utilised for classification tasks, such as distinguishing infected versus non-infected cases or predicting severe versus mild outcomes. SVMs operate by finding optimal hyperplanes

in high-dimensional feature spaces, providing reliable performance even when dealing with small sample sizes and imbalanced datasets, conditions frequently encountered in early pandemic stages [50].

Tree-based algorithms, including decision trees and ensemble methods like random forests [51] and gradient boosting [52], are extensively applied due to their interpretability and accuracy. Random forests, for instance, have demonstrated strong performance in predicting mortality rates and assessing patient risk based on clinical and demographic variables during pandemics [53]. Gradient boosting methods such as XGBoost [54] have similarly provided high predictive performance, particularly in contexts requiring rapid deployment and reliable risk assessments for resource allocation [55].

Probabilistic models, such as Bayesian networks [56], are well-suited for modelling uncertainty and capturing probabilistic transitions in disease states, as they have been effective in modelling and forecasting disease transmission under uncertainty, providing essential information for public health decision making under evolving pandemic conditions [57].

Lastly, clustering algorithms have been frequently adopted for exploratory analysis, grouping regions, patients, or viral variants based on similarity criteria. Methods like k-means clustering [58] and hierarchical clustering have proven valuable in identifying distinct epidemiological patterns, aiding the targeted application of public health interventions and enhancing understanding of disease transmission dynamics [59].

The aforementioned methods, when appropriately selected and implemented, significantly enhance our predictive capabilities and preparedness in facing pandemics. Their effectiveness is further amplified when integrated with advanced analytical techniques, including dimensionality reduction and feature extraction methods. It is worth noting that none of the compartmental and ML approaches mentioned here directly treat drift. This implies that the models created with them will likely have a limited useful life as an accurate mechanism for the prediction of features of interest.

3.2.3 Pandemics

Pandemics are large-scale outbreaks of infectious diseases characterised by extensive geographical spread, significantly increased morbidity and mortality, and profound social, economic, and political disruption [60]. Historically, pandemics such as influenza, HIV/AIDS, and most recently COVID-19, have underscored the necessity of effective surveillance, forecasting, and intervention strategies to mitigate their severe impacts on public health systems and economies worldwide [61]. The rapid spread and severity of pandemics are frequently driven by interconnected global populations, increased international mobility, urbanisation, and, in certain cases, novel pathogens for which population immunity is initially absent or limited. The COVID-19 pandemic, caused by the

novel coronavirus SARS-CoV-2, exemplifies the complex interplay of biological, environmental, and socioeconomic factors that characterise modern pandemics. First identified in December 2019, COVID-19 rapidly evolved into a global health crisis, infecting hundreds of millions and causing millions of deaths worldwide [62]. Its emergence triggered unprecedented public health responses, including stay-at-home orders, physical distancing measures, compulsory masking, and accelerated vaccine development programs [63]. These responses have profoundly altered transmission dynamics, underscoring the critical need for adaptive, real-time analytical capabilities to guide public health decision-making.

From an epidemiological modelling perspective, pandemics pose unique challenges due to their dynamic nature, nonlinear spread patterns, and rapidly evolving conditions. Traditional compartmental models (see section 3.2.2), such as the SIR and SEIR frameworks, have long been the foundation for mathematical modelling of infectious diseases. These models simplify population dynamics into discrete compartments and have been widely employed to estimate transmission rates, assess intervention strategies, and guide public health decisions in the early stages of outbreaks [64]. Their interpretability and mathematical tractability make them valuable tools for scenario analysis and policy evaluation. These analytical models typically rely on assumptions of homogeneous populations and stable transmission parameters, conditions rarely met during real-world pandemics due to changing policies, population behaviours, and pathogen evolution. Therefore, integrating advanced machine learning methods capable of capturing complex temporal-spatial dynamics and nonlinear interactions becomes essential for accurate and timely pandemic forecasting and decision support [65]. Given the profound implications of pandemics, the development and application of robust AI methods represent a priority for researchers, healthcare authorities, and policymakers. Effective use of these techniques has the potential to enhance the predictive accuracy of outbreak models, optimise resource allocation, and ultimately reduce morbidity and mortality rates associated with pandemics [66].

3.2.4 Explainability and drift visualisation in data streams

A data stream is an endless sequence of incoming instances that arrive at high speed, requiring rapid processing [67]. This presents unique challenges, particularly in terms of computational efficiency and limited prediction time. Beyond the need for fast processing, a major difficulty arises from the fact that stream learning algorithms operate in dynamic, non-stationary environments [68]. Such shifts in the data stream can have a significant impact on model performance, often leading to a decline in predictive accuracy and overall system reliability. To mitigate these effects, continuously monitoring and adapting to changes is crucial in real-world applications. Various drift detection techniques can be employed to identify and respond to concept drift, including the ADWIN (Adaptive Windowing) [69], PH (Page–Hinkley) [70], EDDM (Early Drift De-

tection Method) [71] and others[72]. Several methods have utilised SHapley Additive exPlanations (SHAP) to identify data drift. Seiffer et al. [73] proposed a technique for detecting concept drift through clustering based on SHAP values. This approach was empirically tested in a manufacturing scenario. Another domain where concept drift detection is critical is healthcare. In this context, the authors in [74] detected drift by evaluating feature importance in COVID-19 pandemic data.

Incorporating explainability into models operating on streaming data enhances transparency and aids in understanding model decisions over time. Explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations) [75], provide insights into feature contributions and model behaviour. Mozolewski et al. [76] discusses the evaluation of XAI methods within streaming data environments. They assess how XAI techniques can explain machine learning model dependencies on features in a streaming context. Authors in [77] provide an overview of existing XAI methods applied to time series data. It highlights the types of explanations these methods produce and reflects on their impact on building confidence and trust in AI systems operating on temporal data. Understanding the underlying factors contributing to concept drift is vital for developing robust adaptive systems. Recent approaches leverage XAI to elucidate the nature of drift. For example, Edakunni et al. [78] propose a novel framework called DBShap that uses Shapley values to identify the cause of model drift and measure the contribution of different components of the data to the overall deterioration of model predictions. Such methodologies enable practitioners to pinpoint specific features or external factors driving changes, facilitating targeted interventions to mitigate adverse effects on model performance. Another framework called iPDP [79] is an incremental version of Partial Dependence Plots that continuously updates to reflect changes in data distributions and model behaviour over time. This allows for the visualization of feature relationships in the context of streaming or evolving datasets. Authors in [80] proposed methods to understand drift in prediction distributions over time or between data segments. It unifies various feature importance measures under a common framework, providing a systematic approach to analysing prediction drift.

Existing methods for concept drift detection either lack interpretability or provide explanations only after drift has been identified. They often fail to pinpoint which features cause the drift and struggle to handle both abrupt and gradual changes effectively. Our approach integrates drift detection with explainability, using a three-perspective approach that tracks data distribution shifts, model confidence variations, and SHAP-based feature importance evolution. This enables real-time, feature-level insights into drift, making it more comprehensive, interpretable, and scalable for practical applications.

3.3 Methods

To ensure a systematic, thorough, and transparent approach to collecting and filtering relevant studies for this survey, we have followed the Preferred Reporting Items for Systematic Reviews (PRISMA) framework [81] through accomplishing each point in its checklist and by following its flow diagram for performing a systematic identification, selection, and evaluation of papers. We have complemented PRISMA with the Search, Appraisal, Synthesis, and Analysis (SALSA) framework [82] to guide us in the process of actually performing the review. SALSA is a structured methodology for synthesising all the collected literature into new and valuable knowledge. At the same time, PRISMA guided us on presenting all the insights in the form of a useful document.

3.3.1 Research questions

The primary objective of this survey is to provide a comprehensive understanding of the current state of research on addressing concept drift in pandemic-related data. Specifically, it examines the challenges arising from dynamically evolving data and the methods commonly employed to address the resulting distributional shifts. To achieve this, the survey focuses on answering the following research questions:

- RQ1: What are the modelling tasks in pandemics related to concept drift?
- RQ2: Which methods have been applied to deal with drift?
- RQ3: What are the causes of drift, and how does this influence the approach used?

The primary objective of RQ1 is to investigate the modelling challenges addressed in research focusing on concept drift during pandemics. By addressing RQ1, the survey aims to identify the key areas of data modelling related to pandemics where AI has been applied. Through the study of this RQ, we aim to detect research gaps where further research could contribute to better addressing pressing healthcare needs during pandemics.

RQ2 analyses the machine learning techniques employed to address concept drift in pandemic-related data. It aims to categorise the methods used, from conventional algorithms to advanced techniques such as deep learning models. By evaluating the strengths and limitations of these approaches, this survey seeks to provide insight into their effectiveness in handling the challenges posed by rapidly changing data and to identify promising avenues for future methodological improvements.

RQ3 explores where changes in the data occur and how they are treated in the context of concept drift. It analyses the strategies employed to detect and mitigate these changes, aiming to understand how existing methods adapt to the evolving characteristics of pan-

demic data while maintaining reliable model performance or accurately modelling the pandemic.

3.3.2 Literature search methodology used

We conducted a systematic literature search across four academic databases: PubMed, IEEE Xplore, Scopus, and ACM Digital Library. A structured keyword list was carefully developed and organised according to the thematic scope of our survey. We organised the keywords into three different groups, depending on the targeted area of knowledge. The first group of keywords included general terms associated with AI, including general terms (e.g., machine learning) and specific methods frequently applied within this domain (e.g., CNN). The second category included terminology specifically addressing the phenomenon of concept drift and its related aspects. The third category consisted of keywords relevant to pandemics and associated epidemiological contexts. The complete keyword list used for the literature search can be seen in Table 3.1. The objective of organising the search keywords into these groups was to ensure that each article identified in the search had at least one keyword from each group in the title or abstract. No restrictions were placed on the publication date of the articles. The topic of this survey is relatively recent, and the objective of the search was to identify as many relevant articles as possible. Only publications written in English were considered, and the search was conducted in August 2025.

3.4 Results

This section presents the findings of our review on AI approaches for addressing concept drift in pandemics. The analysis is structured in sections corresponding to the three key research questions stated in section 3.3.1 and, within each section, we divide subsections as represented in Figure 3.2. The categories used to answer RQ1–RQ3 were defined after reading the included studies, following an inductive process to select the labels that best encompassed each paper's medical problem, ML method and concept drift cause. In addition, for RQ2 we report explainable AI methods in a dedicated subsection, as this group of methods can accompany any underlying ML model, so grouping them under a specific method family would be misleading.

3.4.1 Studies selected as relevant for the review

In this section, the results of the selection process are presented, ensuring a transparent and systematic approach. The complete selection process can be viewed in Figure 3.3, from the identification of records to the final inclusion. The initial search yielded a total of 663 records. After removing duplicates and entries that were not single manuscripts

Category	Keywords
Artificial Intelligence	Machine Learning; Deep Learning; Artificial Intelligence; Neural Network; data-driven; data mining; pattern recognition; supervised learning; unsupervised learning; reinforcement learning; clustering; dimensionality reduction; feature extraction; ensemble methods; transfer learning; generative model; time series; data stream; support vector machine; decision tree; random forest; gradient boosting; bayesian network; generative adversarial network; autoencoder; action mining; survival analysis; decision support; prediction; predictive system; decision rule; rule-based system; expert system; SVM; k-means; k-nearest; XGBoost; XAI; explainability; knowledge discovery; Large Language Model; Recommender system; Intelligent system; Agent system; Genetic algorithm; Evolutionary algorithm; Heuristic algorithm; Heuristic search; DBSCAN; autonomous decision making; RNN; CNN; LSTM; transformer; linear regression; Regression; classification; Imbalanced problem; minority class; passive learning; active learning; online learning; offline learning; ensemble learning; domain adaptation; metaheuristic technique; bioinspired technique; data science; evolutionary computation; heuristic optimization; concept learning
Concept Drift	Concept Drift; data drift; distribution shift; model degradation; non-stationary data; drift detection; covariate shift; virtual drift; real drift; drift adaptation; dynamic model; adaptive learning; incremental learning; stream learning; time-varying data; model updating; evolving data stream; change detection; drift-aware algorithm; model retraining; windowing technique; ensemble methods for drift; drift robustness; gradual drift; sudden drift; feature drift; label drift; classification drift; prediction stability; concept stability; monitoring data stream; context-aware learning; drift mitigation; drift quantification; unsupervised drift detection; adaptive classifier; covariate shift correction; model recalibration; concept evolution; recurring concept; model performance degradation; time-varying distributions; anomaly drift detection; out-of-distribution data; long-term learning adaptation; continuous learning; context shift; adaptive thresholding; non-stationarity detection; domain shift
Pandemic	Pandemic; Pathogen spread; Quarantine; Social distancing; Disease and infection control; Public health emergency; Pandemic preparedness; Herd immunity; Vaccine development; Vaccine distribution; Public health response; Mortality rate; Morbidity rate; Case fatality rate; Basic reproduction number; R0; Asymptomatic transmission; Community transmission; Super-spreader event; Viral mutations; Variants of concern; Travel restriction; Lockdown measure; Healthcare capacity; Mass vaccination; Pandemic recovery; Economic and societal impact of pandemic; Pandemic modeling; Pandemic risk assessment; Pandemic fatigue; Post-viral syndrome; Flattening the curve; Pandemic governance

Table 3.1: *Keywords categorised by topic.*

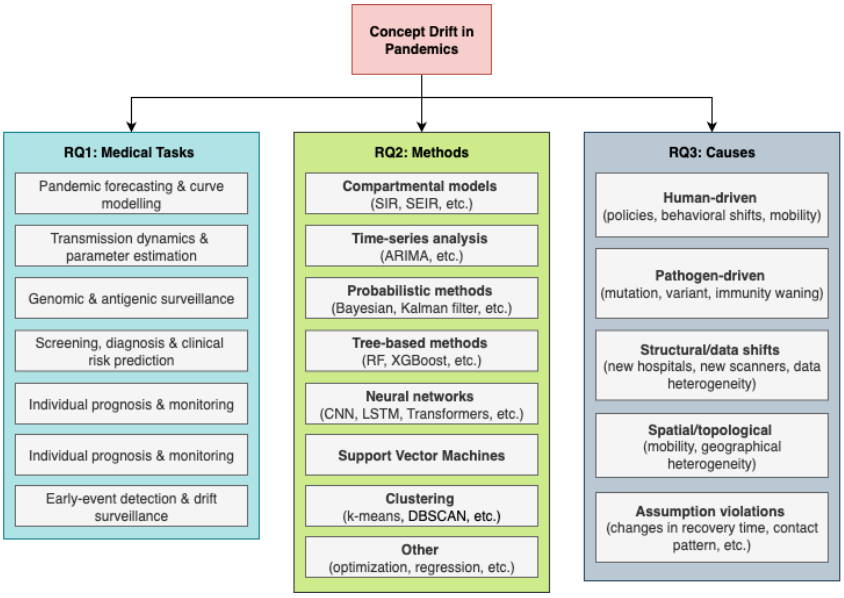


Figure 3.2: *Overview of the three RQs to their corresponding dimensions: medical tasks addressed by AI approaches (RQ1), the methods employed (RQ2), and the main causes of concept drift in pandemics (RQ3) with some examples.*

but proceedings from a conference published with a single DOI, 574 unique records remained. These were then screened based on their titles and abstracts, resulting in the exclusion of 307 records that did not meet the inclusion criteria. The full texts of the remaining 267 studies were reviewed, leading to the assessment of 236 papers for eligibility after some entries were excluded due to the impossibility of recovering them electronically. Then we performed an additional screening round to ensure eligibility. Those excluded did focus on non-pandemic applications. This resulted in 105 final studies in this survey. Figure 3.4 (a) illustrates the temporal distribution of the selected articles, representing that the number of publications addressing concept drift in the context of pandemics experienced a marked increase starting in 2020, coinciding with the onset of the COVID-19 pandemic and, while a decline in the number of articles can be observed in 2024, the trend appears to recover in 2025, suggesting a continued interest in this research area beyond the immediate pandemic context. Figure 3.4 (b) represents the journals with at least two publications, where it can be observed that the selected articles are distributed across a broad set of journals, with Scientific Reports and PLoS ONE being the most represented publication journals. Of the 105 papers selected for review, 10 focus on the influenza pandemic, while the remaining ones address medical problems related to COVID-19.

3.4.2 RQ1: What are the modelling tasks in pandemics related to concept drift?

The focus of RQ1 is on the identification of the data analysis-related problems found in pandemic-related clinical or public-health scenarios. To provide a systematic and com-

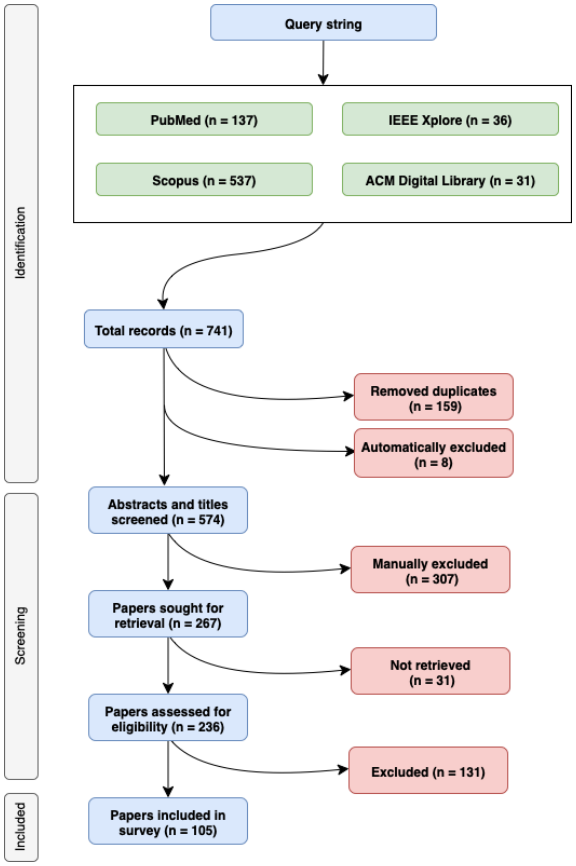
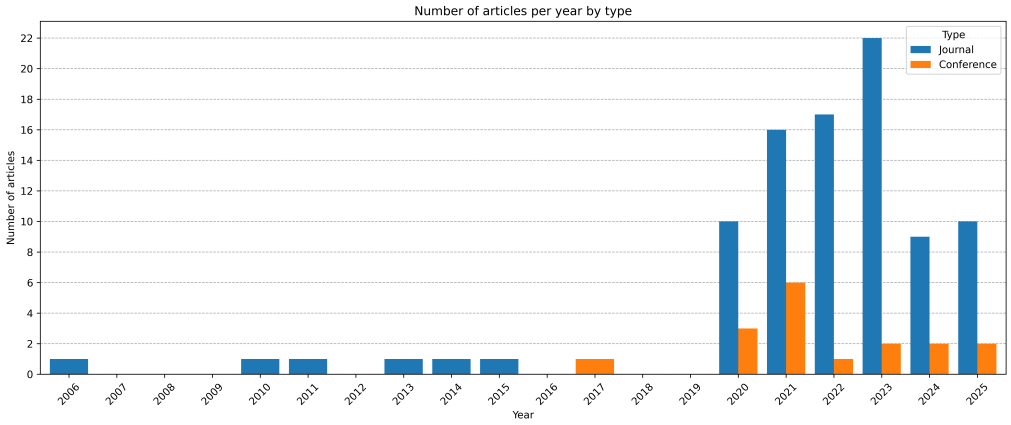


Figure 3.3: *PRISMA flow diagram presenting the number of retrieved and excluded papers at each stage of the analysis.*

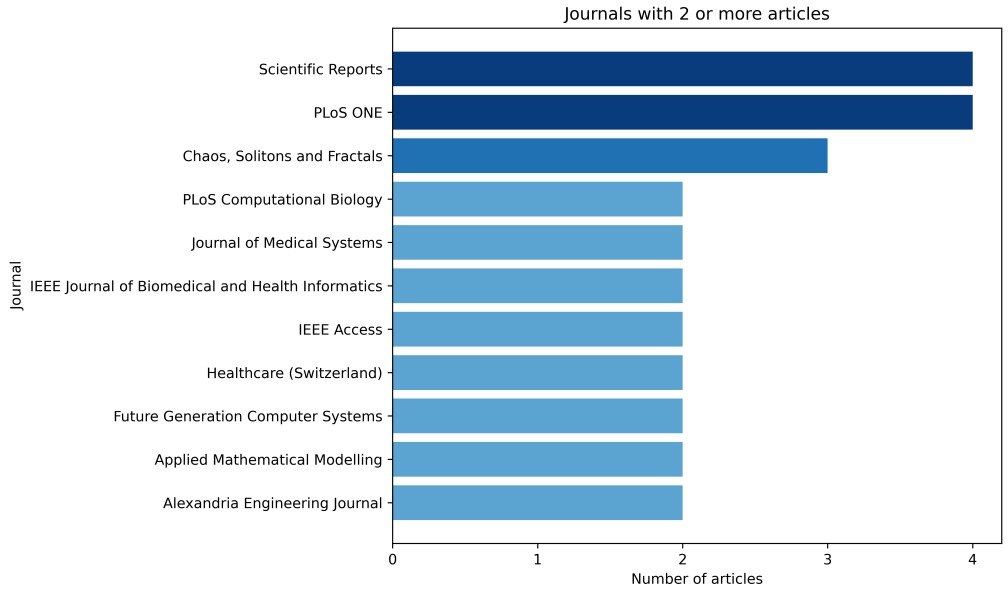
prehensive answer to RQ1, we have identified seven data analysis intensive domains that emerged during COVID-19 and earlier outbreaks: (1) pandemic forecasting and curve modelling, (2) transmission dynamics and parameter estimation, (3) genomics and antigenic surveillance, (4) screening, diagnosis and clinical risk prediction, (5) early-event detection and drift surveillance and (6) intervention and policy evaluation. Each domain is examined in an independent section, where we summarise the relevant studies, describe the pandemic data they use for the analyses, and explain how those data support the corresponding medical task. A detailed, study-by-study summary of all medical problems appears in Table 3.2.

Scope	**Papers**
Pandemic forecasting and curve modelling	[83–140]
Transmission dynamics and parameter estimation	[83, 89, 92, 94, 96, 104, 110, 115, 117, 121–123, 129, 131, 134, 139–146]
Genomic and antigenic surveillance	[147–149]
Patient screening, diagnosis and clinical risk prediction	[98, 101, 115, 120, 132, 150–169]
Early-event detection to rapid actions	[88, 108, 111, 113, 118, 120, 122, 123, 128, 145, 159, 161, 170–178]
Intervention and policy evaluation	[84, 90, 93, 103–105, 111, 112, 121, 125, 127, 129, 132, 141, 143, 144, 170, 172, 174, 179–186]

Table 3.2: *Summary of studies depending on the medical task.*



(**a**) *Number of articles per year separated by publication type.*



(**b**) *Journals with at least two published articles.*

Figure 3.4: *Overview of the selected articles: (a) temporal distribution by publication type, and (b) journals with at least two published articles.*

3.4.2.1 Pandemic forecasting and curve modelling

Pandemic analytics deals with the prediction of the number of susceptible, infected, recovered, and deceased individuals over different time horizons of interest, including daily, weekly, and monthly periods [102]. Commonly known as pandemic forecasting [187], this task draws from time-series published by public health agencies throughout an outbreak. The timely availability of data enormously eases the forecasting task. But, at the same time, it is hindered by hard-to-predict external events of both pandemic and political nature.

Every study considered in this chapter that addresses epidemic forecasting and curve modelling relies on surveillance counts, including variables such as deaths, new cases, and disease tests. Each of these counts can be represented in a curve, so that it is defined as curve modelling. These data are typically aggregated at national, state, provincial, or even smaller administrative levels, depending on the country. Most of the papers included in this section project or predict a single epidemiological variable, most commonly the daily or weekly number of confirmed cases [83, 84, 94, 106, 113, 126, 133, 138]. Many works also extend the scope of analysis with joint curves for cases, deaths and, when available, recoveries from disease [87, 91, 100, 105, 127, 130]. A subset of the studies incorporates additional data sources. Examples include information on tweets used to predict both incident cases and shifts through sentiment analysis [85], or data on human-mobility trends used to predict pandemic mortality in 26 different countries [86].

3.4.2.2 Transmission dynamics and parameter estimation

Quantifying the speed and acceleration of infection spread is essential during a pandemic. Most studies address this by relying on population-level metrics, which provide the basis for accurately assessing outbreak intensity and evaluating the effectiveness of public health interventions. One of these metrics is the effective contact rate, which represents the average number of adequate contacts per person per unit of time that can lead to a transmission. Closely related to it is the transmission rate (β). It measures the probability of infection per contact between a susceptible and an infectious individual. The basic reproduction number, R_0, indicates the average number of secondary cases generated by a single infected individual in a fully susceptible population. When immunity arises or interventions such as vaccination or restrictions are in place, R_0 is complemented with the effective reproduction number, R_t. It estimates the average number of cases under the current epidemiological conditions [188]. For example, the study [142] captures geographic particularities in Italy by predicting the effective reproduction number R_t in cases at the 2009 H1N1 pandemic. Two additional studies expanded upon the aforementioned approach by estimating cases, β and R_0 [117] or R_t concurrently [131]. A later study [96] develops a differential equations-based non-linear model to

estimate COVID-19 population-level metrics and simulate the pandemic in China. The study [110] trained a model to predict the R_0 parameter during the COVID-19 pandemic in Italy in three distinct time windows. Mobility patterns are essential in these scenarios. A mobility-based model demonstrates how it is possible to model COVID-19 reproduction parameters [115], and the spread of the BA.1 Omicron COVID-19 subvariant was simulated including the contact rate in Pakistan and Kazakhstan [92] at three different time frequencies. Several contributions estimate time-varying transmission parameters directly from daily surveillance curves [101, 120, 134, 146]. The study [132] in Thailand focuses on estimating β to reveal the immediate effects of actions imposed by policymakers. A further analysis in Beijing combines Baidu symptom-search trends with traditional influenza illness reports to improve real-time estimates of R_t during early 2020 [98].

3.4.2.3 Genomic and antigenic surveillance

Viruses are infectious agents defined by their genetic material, which can be either DNA (double-stranded sequences of nucleotides) or RNA (single-stranded sequences). This genetic information encodes the instructions needed to hijack a host cell's machinery, forcing it to copy this genetic material to produce more viruses to replicate themselves, spread within the body, and cause disease [189]. Each time the genetic code of a virus changes by mutating its nucleotides, we get a new strain of that virus (a genetic variant of the virus with distinct characteristics). New strains can accelerate transmissibility and, consequently, the impact that the virus has on the population.

Genomic surveillance focuses on monitoring the evolution and spread of viral genetic material within human populations. Its main concern is monitoring how viral mutations affect transmissibility, immune escape (the ability of the virus to evade recognition by the host's immune system), and vaccine match (the degree to which existing vaccines remain effective against circulating strains). Antigenic analysis, meanwhile, is a process of assessing whether circulating strains can bypass existing immunity. Two population-level studies in the corpus address these complementary objectives. One analyses how protein sequences of COVID-19 from the Global Initiative on Sharing All Influenza Data (GISAID) can be used to identify emerging lineages over time [147]. A separate study tries to relate amino acid substitutions in the H1N1 haemagglutinin and associated glycosylation changes to the rise of antigenic variants of H1N1 [148].

3.4.2.4 Patient screening, diagnosis and clinical risk prediction

The process of screening patients, together with the diagnosis and clinical risk assessment, is of paramount importance in the effort to maintain the containment of the pandemic and to facilitate optimal decision-making regarding patients presenting to pri-

mary care facilities or hospitals. It is crucial to make informed decisions regarding the most effective strategies for accurately detecting infections. One project proposed a dashboard that uses hospital records to monitor lab tests and scans, supporting teams that control infections in hospitals [152]. In the context of inpatient care, a prospective, dynamic prognostic model that incorporates vital signs and laboratory data continuously identified patients at high risk of short-term deterioration to facilitate efficient triage and resource allocation [161]. Another work in emergency departments used explainable machine learning algorithms for identifying risk of admission/overcrowding signals to maintain safe stratification [163]. Complementarily, a multicenter study delivered risk stratification for mortality, ICU admission, and invasive mechanical ventilation across waves, highlighting predictors such as SO2 and lymphopenia and documenting temporal drifts that affect clinical risk prediction [169].

From a screening perspective, different approaches have been followed. Some approaches work directly with imaging, blood panels or symptom checklists. Two studies utilised multi-centre CT collections for COVID-19 screening, despite variability across scanners and patient cohorts [150, 156]. Chest-X-ray pipelines that first segment the lung fields improve generalisability and maintain good performance predicting COVID-19 [151, 164]. In laboratory data, haematological markers are employed to predict whether a patient is affected by COVID-19 [153]. A further study utilised cough audio data from two datasets concerning the COVID-19 pandemic to diagnose patients [166]. Symptoms and lifestyle surveys provide rapid infection screening when PCR capacity is limited [159].

3.4.2.5 Early-event detection to rapid actions

Detecting sudden increases in viral transmission or shifts in patterns of human interaction is essential for enabling rapid public health responses. Most studies approach this detection task by analysing the evolution of case numbers by region, mobility data patterns, and symptom reports. They are used to predict incoming new waves or to identify localised outbreaks. For example, a study in Massachusetts used a method based on network analysis to find unusual routes of transmission [170], whereas mapping risk to each province in Iran is used to predict early outbreak probability [88], as in this example at Jammu and Kashmir in India [111]. In Colombia, a model based on similarity between areas gave early alerts for small towns [171]. Other studies applied methods that detect sudden changes in trends, without depending on any specific model, to flag shifts in pandemic behaviour [108, 118, 122, 123, 128, 173]. Additional approaches classify rising versus falling case trajectories to support outbreak prevention [175], exploit patient symptom profiles for surveillance [176], and incorporate event-based predictors [145] to enhance early-event detection.

3.4.2.6 Intervention and policy evaluation

Understanding the real-world effects of lockdowns, vaccination campaigns, and hospital protocols is key to managing both infection rates and pressure on healthcare systems. Most studies address this need by fitting or extending compartmental models and then adjusting the model as policies change, followed by the simulation of "what if" scenarios. For example, studies conducted in the United Kingdom [93], France [144], Canada [125], and a multi-site investigation [129] re-estimate transmission parameters as hypothetically introduced policies change, generating day-by-day predictions that policy-makers can visualise. Several papers also integrate optimisation capabilities into the models. For example, one model determines the optimal combination of vaccination and isolation to minimise both illness and costs [143], while another identifies vaccination strategies that maintain infections below a safe threshold [182]. One paper shows that replenishing antiviral stockpiles on demand outperforms one-off purchasing strategies for pandemic influenza [180]. Other research focuses on accelerating the exploration of epidemiological models, providing tools that help policymakers rapidly understand how different assumptions or interventions could influence epidemic outcomes [179]. COVID-19 specific works evaluate the joint impact of vaccination and non-pharmaceutical measures [104], compare intervention strategies across multiple regions [90] and combine epidemiological and economic indicators, allowing authorities to balance health benefits against gross domestic product losses when planning how and when to resume regular activity in urban areas [105].

3.4.3 RQ2: Which methods have been applied to deal with drift?

This section reviews the AI techniques employed to perform predictive tasks in pandemic contexts subject to concept drift. We categorise the approaches into two groups, distinguishing between AI-based and non-AI-based methods. Including non-AI approaches allows for a more comprehensive analysis of the literature. Among non-AI methods, we identify compartmental models, statistical techniques, and probabilistic approaches. The AI methods considered include tree-based algorithms, neural networks, clustering techniques, support vector machines, less frequently represented approaches, and selected methods for explainable AI.

In the following subsections, we discuss each category of AI methods in detail, focusing primarily on the specific pandemic-related problems for which these techniques have been applied. Rather than delving into the mathematical or algorithmic foundations of each method, our emphasis is placed on their practical use cases, such as forecasting case numbers, predicting emerging variants, or evaluating intervention policies under concept drift scenarios. For readers interested in the technical or theoretical underpinnings of these models, we encourage consultation of the original or more specialised publications [190, 191].

3.4.3.1 Compartmental methods

Although compartmental models are not usually classified as AI-based, they remain fundamental tools for pandemic analysis. Built on differential equations and mechanistic assumptions, they serve as standard benchmarks and complements in studies that confront concept drift in epidemiological data. Their strengths include interpretability, i.e. how their outcomes are linked to domain knowledge, and the flexibility to incorporate changes to essential system variables like new enforcing policies, emerging variants, and population behavioural shifts over time [192]. The classic SIR (Susceptible–Infectious–Recovered) framework supports many early-wave or short-term policy analyses [90, 93, 94, 96, 134, 142]. Extending SIR to SEIR (Susceptible–Exposed–Infectious–Recovered) introduces a latent stage that captures the impact of distancing amongst individuals or mobility restrictions on presymptomatic transmission [85, 92, 102, 103, 141]. Adding a mortality compartment yields SEIRD (Susceptible–Exposed-Infectious–Recovered-Deceased) models, which allow policymakers to evaluate interventions in terms of deaths averted [91, 100, 107, 108, 146]. Some studies further stratify populations into low and high contact groups to track distinct reproduction numbers over time [97], and some others implement new compartments such as vaccination or individuals in quarantine [112, 131, 139]. Moreover, hybrid approaches combine compartmental frameworks with machine-learning modules to improve predictive accuracy or adapt to drift [87, 100, 124]. A study-by-study overview of the compartmental methods appears in Table 3.3.

Method	**Papers**
SIR	[90, 93, 94, 96, 101, 102, 106, 110, 111, 117, 121, 129, 130, 134, 142, 144, 185]
Contact-SIR	[145]
SIRD	[110, 111, 130]
SEIR	[83, 85, 92, 98, 102–104, 114, 115, 117, 123, 132, 135, 141, 181, 184]
SEIRD	[87, 91, 100, 107, 108, 119, 127, 136, 146, 182]
SVIR	[104, 139]
SEIRDV	[105]
SEQIR	[124, 131, 140, 143]
SEQIRD	[112, 117]
SEAIHRD	[89, 97]

Table 3.3: *Summary of Compartmental methods.*

3.4.3.2 Time-series specific analysis models

Time-series analysis studies phenomena whose features are captured through variables that are measured in regular time intervals. These can be univariate or multivariate time series. The most used technique in time-series modelling is the AutoRegressive

Integrated Moving Average (ARIMA). It remains useful when analysts require a low-dimensional representation of long-memory patterns and no reconstruction of a complete probabilistic model. In Italy and Iran ARIMA and exponentially weighted variants have been utilised to formulate forecasts of COVID-19 daily cases, thereby providing policymakers with a baseline against which to assess the efficacy of newly implemented restrictions [84, 88]. One study also showed that ARIMA can support other models; for example, adding ARIMA results to a SEIRD epidemic model helped reveal trends that the standard compartmental approach alone could not capture [100]. Other methods such as Prophet and SARIMAX are utilised to deal with time-series data to predict confirmed, deceased and recovered cases of COVID-19 [99, 172]. Detailed study-wise summaries of the time-series specific analysis models can be found in Table 3.4.

Method	**Papers**
ARIMA	[84, 86, 88, 99, 100, 119, 162]
SARIMAX	[172]
Continuous Markov Chain	[83, 127]
Interrupted Time Series	[84]
Prophet	[99, 162]
Random Walk Method	[120]

Table 3.4: *Summary of time-series analysis methods.*

3.4.3.3 Probabilistic methods

Many studies combine epidemic models with Bayesian filtering techniques to update predictions as new data arrive. A common approach is to use a Kalman filter [193], which estimates the hidden state of the outbreak in real time [83, 144]. Another method, based on Gaussian process regression, avoids fixed model assumptions and instead uses data patterns to make forecasts. Gaussian process regression is used to forecast cases and deaths with probabilistic uncertainty estimates, enabling state-level projections during early 2020 [95]. Dynamic Bayesian networks have also been applied, first in influenza surveillance [177], and more recently to explore different COVID-19 scenarios [94]. For a complete study-by-study account of the probabilistic methods reviewed, see Table 3.5.

Method	**Papers**
Kalman Filter	[83, 144, 185]
Gaussian Process	[95, 116]
Naive Bayes	[175]
Mean Change Model	[173]
Bayesian Models	[86, 94, 109, 120–122, 159, 171, 177]

Table 3.5: *Summary of probabilistic methods.*

3.4.3.4 Tree-based Methods

Decision trees, random forests, and gradient-boosted tree ensembles, such as XGBoost, are among the most popular methods in pandemic studies that confront concept drift. One study used XGBoost to train an event-centric predictor that uses Twitter activity with official case reports to flag local surges [85], while Hoeffding trees are used to predict case growth across Californian counties [133]. Random forests accelerate policy searches when coupled to a SIR model [134], and decision-tree stratification informs age-specific antiviral stockpiling for influenza [180]. Additional work compares tree-based models for emergency-department crowding [174], other ones trained them in haematology-based COVID-19 classifiers [153], national case counts [109, 163], and forecasts next-day hospital flows [162]. The popularity of these methods stems in part from their capacity for incremental updating and online splitting, which makes them well-suited to environments where data distributions are constantly shifting. Tree-based statistics have also served as a variable-selection method in a prospective model of rapid clinical deterioration using TreeScan method [161]. Study-level details for all tree-based methods are summarised in Table 3.6.

Method	**Papers**
Gradient Boosting	[85, 91, 109, 153, 154, 160, 163, 169, 172]
RF	[91, 109, 134, 153, 160, 162, 172, 174, 175]
Decision Tree	[109, 118, 172, 180]
Hoeffding tree	[109, 133, 175]
TreeScan	[161]

Table 3.6: *Summary of tree-based methods.*

3.4.3.5 Neural Networks

Neural networks are widely used in pandemic research because they combine high predictive accuracy with the ability to model complex nonlinear relationships in heterogeneous data. We group the principal architectures into five broad families. CNNs dominate imaging applications, supporting cross-site CT or chest-X-ray classifiers that remain robust to scanner and cohort variation [150, 151, 157, 164]. Recurrent networks handle temporal signals: vanilla RNNs appear in short, univariate forecasts [91]; LSTM models are the mainstay for pandemic curves and hospital patient-flow prediction [99, 109, 123, 126, 138]; and Gated-Recurrent-Unit (GRU) variants provide comparable accuracy with fewer parameters, including bidirectional extensions that exploit both past and future context during training [98, 126, 146]. Graph neural networks capture spatial transmission patterns at the city or provincial scale [170], while Transformer architectures support drift-aware monitoring of symptom streams and other high-dimensional inputs [140, 158, 176]. Several studies benchmark multiple neural-network types within automated machine-learning workflows to identify the configuration that yields the most accurate COVID-19 forecasts [86]. A study-by-study overview of the neural network

methods appears in Table 3.7.

Method	Papers
MLP	[90, 97, 123, 128, 130, 153, 164]
ANN	[111, 125, 126, 134]
CNN	[95, 116, 133, 150–152, 156, 157, 164–166, 168]
AdaNet	[151]
LSTM	[86, 99, 109, 123, 126, 133, 138, 167]
GRU	[86, 98, 126, 146, 167]
RNN	[91, 137, 167]
Transformer	[140, 158, 176]
Graph RNN	[123]
Graph Attention Network	[170]
Adaptive Neuro-Fuzzy Inference System	[123]

Table 3.7: *Summary of neural network methods.*

3.4.3.6 Support Vector Machines

SVMs are relatively infrequent in the reviewed literature. One Iranian study employs a support-vector classifier to estimate the probability of an outbreak in each province [88]. Other investigations use linear-kernel SVMs as baseline models, valuing their simplicity and numerical stability when benchmarking more complex algorithms [134]. Moreover, in a haematology-based diagnostic study, a SVM is trained to predict COVID-19 mortality from blood tests [153]. Support Vector Machines can also be utilised as regression task models (SVR), as two studies used SVR to model the progression of pandemics over time [134, 162]. A comprehensive summary of the support vector machine methods, organised study by study, is provided in Table 3.8.

Method	Papers
Support Vector Classifier	[88, 134, 147, 153, 172]
Support Vector Regressor	[109, 162]

Table 3.8: *Summary of Support Vector methods.*

3.4.3.7 Clustering

Clustering techniques help uncover spatial and temporal patterns that are not apparent in raw pandemic data, precisely because of their unsupervised nature, in contrast to the other AI-related techniques reviewed above. One analysis based on real-world events is performed according to Twitter posts about COVID-19 to identify relevant events related to the pandemic for different batches of data using DBSCAN, and, after this clustering technique is performed, a COVID-19 event trend is performed [85]. A second study on the spread of COVID-19 in Italy groups consecutive time windows by analysing how the epidemic evolved. It uses hierarchical clustering to identify similarities and differences in how the pandemic evolves across different regions [110]. Finally, a study applies a Topological Weighted Centroid (TWC) clustering method to US case data to find hotspots and

likely spread paths of COVID-19, showing patterns that line up with major airport routes [178]. Study-level details for all clustering methods are compiled in 3.9.

Method	Papers
DBSCAN	[85]
Hierarchical clustering	[110]
Topological Weighted Centroid	[178]

Table 3.9: *Summary of clustering methods.*

3.4.3.8 Other AI-based techniques

This section accounts for methods which do not fit into the earlier categories. It predominantly includes optimisation approaches used to refine particular aspects of compartmental models. In a first example, a particle-swarm optimisation scheme is used in the work [141] to fine-tune a SEIR model and to design equitable vaccine administration in China. In this second example, [143], a Pareto optimal distribution of vaccination resources and automated design of an isolation strategy is performed on a reaction-diffusion model while trying to control infection rate and intervention cost simultaneously. In the third example of this group of works, least-squares inverse optimisation extracts SIR coefficients directly from observed case curves [101].

Linear models are also an option, either as baseline approaches to compare with or as interpretable components to identify the effect size and direction of each variable considered in the model. For example, we found five studies relying on logistic regression [110, 117]. Interestingly, we identified one incremental approach that remains continuously up to date by incorporating new information from each successive batch of viral strains [147]. Two further examples combine a logistic regression and a Lasso model with k-nearest-neighbour classifiers to enhance robustness under domain shifts [153, 174].

A system-dynamics dashboard together with a fuzzy inference iteratively searches reopening policies that maximise gross domestic product and minimise risk [105]. In [115], a dynamic model, specifically used to predict the transition from Delta to Omicron virus strains, is formulated as a non-linear optimisation problem and solved by the COBYLA algorithm based on decomposing the problem into multiple linear problems. Finally, in [182], a genetic algorithm-based approach is applied to the optimisation of a fuzzy neural network that predicts the vaccination rate and the isolation strategy. All methods in this category are summarised in Table 3.10.

3.4.4 Explainable AI methods

The performance of AI models alone is not sufficient in all contexts. In safety-critical domains, it is equally important to gain a transparent understanding of how models operate and how their predictions are generated. The underlying machine learning and

Method	Papers
Symbolic Regression	[179]
Particle Swarm Optimisation	[107, 138, 141]
Domain-Specific Batch Normalization	[150, 152]
Polynomial Regression	[88, 183]
Ridge	[88, 162]
Lasso	[104, 148, 174]
ElasticNet	[161]
Lattice Gas Model	[142]
Linear Regression	[91, 110, 113, 117, 147, 153, 160, 172]
KNN	[153, 162, 172, 174, 175]
Cox Hazard Model	[155]
Fuzzy Inference System	[105, 182]
Genetic Algorithm	[135, 182, 186]
Emerging pattern model	[149]
Signal Processing	[140]
Radial Basis Function Networks	[123]
Reptile Search Algorithm	[126]
Paretto Optimisation	[143]
Least Squares Optimisation	[101]
COBYLA	[115]

Table 3.10: *Summary of Other AI-based techniques.*

deep learning techniques often produce highly complex models, whose decision processes are not easily accessible to humans. In healthcare, where interpretability and trust are paramount, the demand for XAI is particularly evident, as it enables understanding of how predictions are generated [194]. Moreover, the utilisation of XAI methodologies in the analysis of fluctuations in feature importance trends has the potential to identify changes in the data, thereby facilitating the identification of concept drift. Once concept drift has been identified, it can also facilitate an understanding of the underlying causes of the change, thereby aiding decision-making processes. XAI is addressed almost entirely with SHAP feature importance techniques [195]. SHAP was used to explain a voting ensemble for forecasting patient flows, showing which features (e.g., alert level, search trends, pedestrian traffic, autoregressive lags) drove predictions and how these drivers shifted during pandemic-related concept drift [162]. In an emergency departments admission model, SHAP tracked changing feature importance over time to monitor data and concept drift, flagging when retraining might be needed and helping identify emergent clinical risks [163]. A multicenter study used SHAP to identify important variables to stratify clinical risk (mortality or ICU) across waves [169]. In imaging, a novel method produces attention heat maps that indicate the lung regions most responsible for each diagnostic decision, linking visual cues to model output [158]. A CT study uses class activation mapping (CAM) to produce heatmaps that highlight the lung regions driving each prediction, giving clinicians a simple visual check of suspected lesions and improving trust in model outputs [157]. At the city level, a spectral graph wavelet transform (SGWT) on a learned mobility graph helps spot anomalous spread patterns and rank influential locations over time, offering an interpretable signal of where and when risk is rising [170]. Table 3.11 summarises all methods related to this category.

Method	Papers
SHAP	[162, 163, 169]
Class activation mapping	[157]
Attention maps	[158]
Spectral graph wavelet transform	[170]

Table 3.11: *Summary of explainable AI methods.*

3.4.5 RQ3: What are the causes of drift, and how does this influence the approach used?

Pandemics are inherently dynamic; consequently, their associated data are dynamic in nature. Therefore, when concept drift happens, identifying its cause is crucial for a correct treatment of the change. Within a pandemic context, abrupt drifts are usually due to sudden and strict lockdown enforcement, whereas gradual drifts may arise during the transition between viral variants within the infected population. In this section, we identify the possible causes behind each observed pandemic drift, as reported in the reviewed studies, together with the corresponding adaptation strategies. Whenever possible, we differentiate between causes of biological nature (e.g., the emergence of new genetic variants or waning immunity) and those that are driven by the action of policymakers (e.g., changes in non-pharmaceutical interventions or testing protocols). For each of them, it is easy to infer the type of concept drift, as appears in Table 3.12.

Causes of CD	Papers
Human-derived concept drift	[83–85, 87–92, 94, 95, 97, 101–108, 110–114, 116–134, 136–138, 140, 141, 143–146, 154, 155, 159–163, 169, 170, 172, 173, 175–177, 180–183, 185, 186]
Changes in the pathogen	[86, 92, 115, 116, 118, 122, 123, 129, 131, 133–135, 138, 140, 141, 144, 146–149, 168, 177, 182, 186]
Structural and data gathering changes	[86, 88, 89, 109, 118, 126, 132, 134, 137, 139, 150–153, 155–158, 163–168, 171, 175, 184, 196]
Topological nature	[88, 93, 96, 111, 116, 118, 127, 132, 133, 145, 170, 171, 174, 178, 183]
Violation of model´s basic assumptions	[84, 89, 97–102, 107, 112, 114, 117–119, 121, 130, 132, 134, 142, 155, 158, 171, 172, 179–181]

Table 3.12: *Summary of studies depending on the cause of the CD.*

3.4.5.1 Human-derived concept drift

Changes in human behaviour, often driven by policy enforcement measures, can induce concept drift in pandemic data. Such a type of drift is frequently called *behaviour and policy-driven concept drift*. The study by Capistrán et al. [83] estimates time-varying changes in the effective participating population and the observation probability, capturing rapid behavioural responses across 27 EU countries. This other study employs an interrupted time series design to evaluate the initial and secondary 2020 waves in Italy [84]. The observed shifts are attributed to policy-induced behavioural changes, namely the contrast between a nationwide lockdown and subsequent region-based measures. The study ad-

vocates a learning-from-outcomes approach and adaptive policy updating to mitigate future waves. In this other work [162], they treat government restrictions as a driver of concept drift by adding a real-time COVID-19 alert system to patient arrival forecasting models, and uses SHAP plus moving windowing retraining of an ensemble to monitor and adapt. The study by Feng et al. [160] introduces a performance-monitoring framework for clinical risk models during the pandemic that addresses a model performance dynamic control limits to detect calibration drift. Another example of using governmental measures as causes of drift, in combination with other factors such as seasonal weather or viral variants, is [144]. It relies on a dynamic model to predict infection rates. In the same vein, an analysis of how vaccination and isolation strategies affect COVID-19 transmission is performed in [143] through a model that studies COVID-19 prevention and control. Finally, an interesting factor that can be considered as driving concept drift is based on interaction patterns among humans, considered in this study [145] and based on a *dynamic contact network model*.

3.4.5.2 Concept drifts derived from changes in the pathogen

Small changes in the DNA or RNA of a virus driving a pandemic can alter its properties and function, and may ultimately give rise to concept drift. In this review, we classify papers under the category *being caused by changes in the pathogen* when the sudden emergence of a new viral lineage disrupts model performance. Such an event occurs when a virus, although genetically similar to its ancestor, acquires sufficient functional differences to modify its transmissibility. The paper [115] proposes a two-strain pandemic model and shows its ability to accommodate mutation-driven changes in transmission. The dynamic model in [92] studies the Omicron variant across three periods by re-estimating key rates (contact, quarantine, testing), capturing stage-specific shifts. A modelling of successive COVID-19 waves with a non-autonomous SIR is also performed in [129], re-training the transmission-rate function when needed. COVID-19 data were used to train a Transformer that learns time-varying parameters and forecasts trends across waves and mutations [140]. Finally, a Particle Swarm Optimisation together with a LSTM predicts confirmed cases over multiple waves, adjusting to evolving epidemic patterns [138]. When there are fluctuations in the virus transmission rate, if we know the possible factors that influence such variation, e.g., changes in weather or season, it is possible to use that knowledge about the virus's behaviour to improve model adaptation and decision-making.

3.4.5.3 Drift caused by structural and data gathering changes

Domain drift arises when the underlying feature space of the model is altered. This type of shift can invalidate learned representations, leading to severe degradation of predictive performance or, in extreme cases, complete model failure. Examples of such domain

drifts include adding a new CT scanner to the lab equipment used to monitor the pandemic, or including a new hospital in the study after a while. The framework [150] deals with COVID-19 diagnosis from CT images when new data arrives from different hospitals, creating a heterogeneous dataset and adapting to those different data distributions, conducting a separate data normalisation in a latent space. A similar example performs independent feature normalisation in a latent space to improve learning from multi-site CT datasets from hospital information systems while mitigating inter-site distribution differences [152]. A federated semi-supervised ML approach is followed to be able to learn from CT images in three different countries, reducing performance loss from domain shift and uneven annotation quality [156]. Another study aggregates several CT datasets and creates a weakly supervised classifier robust to the domain shift, adding tools that help handle heterogeneous scanners and understand cross-site differences [157]. A transformer neural network model trained on image contours was adapted to CT scans from ten different countries, improving robustness under distribution shift while preserving clinically meaningful attention [158]. A privacy-preserving, domain adaptation approach enables diagnosis across centers and prioritizes learning from the most informative target samples [165, 196]. Finally, the study [153] that models mortality risk based on blood-test samples, detects domain shift by detecting drops in performance, and reacts by re-training the model to restore its performance.

3.4.5.4 Causes for drift of topological nature

An important pandemic problem dimension that can drive drift is spatial information, in various forms, primarily as the mobility patterns of individuals. The study [170] used a model based on mobility connections between cities in the state of Massachusetts, USA, to detect areas with unusual infection patterns, and showing how the model structure drives changes in observed COVID-19 spread. In China, researchers built a model that reflects how COVID-19 spreads differently across regions as conditions change, and captures how spatial connectivity changes COVID-19 transmission [96]. In Portugal, an analysis across 278 municipalities found that a single model is not able to explain the spread of the virus equally well at all locations. Therefore, they created a flexible model that adjusts to the specific infection patterns of each area by taking advantage of information from similar neighbouring regions [171]. In the Jammu and Kashmir region, the model was adapted separately for each area, utilising updated local data. This helped improve accuracy in capturing how the virus was spreading in each place [111]. Moreover, in the US, a study combined the central positions of counties on a map with time-based analysis to identify locations where outbreaks were likely to appear next [178]. All these works demonstrate that regional information and local views of the modelling problem are crucial for global success.

3.4.5.5 Drift due to violation of model´s basic assumptions

Unexpected external events, such as measures applied to the population by policymakers or the abrupt emergence of new virus strains, are relatively easy to identify as linked to drift. However, they are not the only causes of changes in the pandemic. Usually, more subtle changes in the form of violations of the model's original assumptions may also lead to drift. Models are designed by considering basic assumptions on the target problem. Assumptions usually simplify the modelling problem by enabling abstraction from unnecessary low-level details. Therefore, assumptions are thought to remain during the model's lifetime. However, in highly changing environments like a pandemic, this might not be the best strategy. And usually, when the scenario no longer fulfils an assumption, the quality of the corresponding model´s predictions will get worse.

Pandemic models commonly rely on assumptions regarding contact patterns and the time to recovery from the disease. Due to better informative policies, human contact patterns may evolve. Patient recovery times may also shorten with improved treatment. In that sense, an international analysis of the COVID-19 pandemic is performed in [101], estimating time-dependent SIR parameters and showing that they need readjustment when conditions change, while also noting that the assumption of the model often fails. A drift-aware approach is used in the study [118] to predict daily infections and deaths per million people. They combine incremental and batch learning to detect changes in data and make the model adapt when its assumptions fail. This other study [117] uses logistic regression, SIR, and SEIR-based models and re-estimates parameters via nonlinear least squares on new data to generate time-varying dynamics for forecasting outbreaks. In [146], violations of fixed-parameter assumptions in a compartmental model are addressed by using a GRU to produce time-varying parameter estimates, allowing adaptation to policy shifts and viral mutations. In a related study [91], a SEIRD model is developed through an incremental learning approach, adding or updating models as performance degrades to follow changing data. Another study assumes that in pandemics, fixed-model assumptions may not hold indefinitely. Therefore, an online process for estimating parameters of a Bayesian model based on particle filtering is performed. This model is the basic building block of a decision-support system to assist in real-time [107]. Also based on a Bayesian approach, this other study performs a dynamic updating, or even re-fitting, of a probabilistic model according to the knowledge gained with previous models [155].

3.5 Discussion

This systematic review provides a comprehensive analysis of how AI and ML approaches have been adapted to handle concept drift during pandemics. Pandemics inherently create dynamic environments. Therefore, predictive models based on pandemic data

must continually adapt to changes in underlying data patterns. Such data shifts occur due to multiple factors, including the biological evolution of pathogens, changes in human behaviour, and the implementation of new public health policies in response to the disease. The necessity to incorporate adaptive mechanisms into predictive models, regardless of whether such models are produced with statistics, differential equations or machine learning, is consistently emphasised throughout the reviewed literature, aligning with similar claims from previous surveys in related fields that highlight drift as a critical challenge for predictive analytics [197, 198]. Recent publications have presented some measures to address concept drift, offering more complex approaches than the conventional updating of compartmental model parameters, which is the common strategy employed when a pandemic emerges, with a primary focus on the number of cases and deaths, as reported by the governments of respective countries, utilising time series data while adapting to different waves and mutations of the virus. Furthermore, ten articles out of all those obtained for the survey deal with Influenza pandemic, while the rest work with COVID-19 virus data. Interestingly, the selected papers are published in multidisciplinary journals, together with specialised journals in computer sciences, applied mathematics, and biomedical informatics. This dispersion indicates that research on concept drift in pandemics is still scattered across diverse publication journals, which further reflects the interdisciplinary nature of this topic.

This survey reveals that research has primarily focused on forecasting and curve modelling, parameter estimation, and clinical prediction tasks. These areas have historically been dominated by compartmental models such as SIR, SEIR, and SEIRD, which remain a cornerstone due to their interpretability and policy relevance. However, their reliance on fixed assumptions and homogeneous populations makes them highly vulnerable to drift, requiring constant recalibration. Machine learning methods are increasingly integrated into these tasks, offering improved predictive performance and flexibility. Yet, the review shows that most ML applications have prioritised short-term prediction (e.g., incidence curves, hospital demand, or diagnostic support) rather than directly addressing drift adaptation. This imbalance underscores an opportunity for future work to better align ML predictive power with explicit drift-aware strategies.

Compartmental models continue to play a predominant role in pandemic analysis, particularly during the early phases of an outbreak, yet the survey reveals a progressive shift towards machine learning approaches. Tree-based ensembles (RF, XGBoost), neural networks (LSTM, CNN), and hybrid frameworks have been increasingly applied to pandemic data streams. However, only a subset of these explicitly incorporates drift adaptation mechanisms, often through incremental learning or ensembles that can reweight or replace outdated base models. A particularly promising approach is the application of Transformer architectures, which recent studies have exploited for their capacity to model long-range temporal dependencies and integrate multimodal inputs. Unlike RNNs or LSTMs, Transformers do not rely on sequential state propagation, allowing more efficient modelling of heterogeneous pandemic signals such as symptom streams,

genomic sequences, and clinical records. Despite their potential, their application is still limited compared to classical models, suggesting that their role in drift-aware pandemic modelling is at an early but highly promising stage.

The categorisation of drift into biological, behavioural, structural, and spatial domains proved useful for systematising the diverse causes reported. Although few studies explicitly adopt this classification, our framework highlights the implicit variety of drift drivers more clearly. For example, biological drift arises with the emergence of new variants or waning immunity; behavioural and policy-driven drift stems from lockdowns, vaccination strategies, or changes in mobility; structural drift emerges from new diagnostic protocols or heterogeneous data sources; and spatial drift is linked to regional heterogeneity and mobility networks.

Adaptive strategies vary accordingly: compartmental models are often re-fitted as new policies or variants appear, ML ensembles are retrained or extended to cope with distributional shifts, and privacy-preserving federated approaches have been tested to handle domain drift across hospitals. Despite these advancements, several limitations within the reviewed literature must be considered. Many studies lacked thorough external validation, limiting their generalizability across different geographic and demographic contexts. Explicit drift detection and quantification methods were often inadequately described or absent, leaving significant uncertainty regarding optimal timing and frequency of model updates. Additionally, comparative evaluations of multiple adaptation strategies within standardised environments remain almost unexplored, hindering the identification of effective methods.

Methodological limitations of this systematic review itself also warrant discussion. Although comprehensive searches were conducted across major databases (PubMed, IEEE Xplore, Scopus, ACM Digital Library), the possibility of publication or indexing biases cannot be entirely eliminated. Furthermore, the interdisciplinary nature of the topic posed challenges in standardising terminologies and criteria during the screening process. Notably, the lack of standardised terminology and unified notation related to concept drift complicates the identification and comparison of relevant studies. Moreover, explicit classification of the drift type, such as sudden, gradual, incremental, or recurring, is frequently absent, with most articles focusing only on drift causes rather than their temporal characteristics. Addressing these issues will facilitate more transparent communication within the research community and enable more effective systematic reviews and meta-analyses.

3.5.1 Gaps and Future Directions

Although significant progress has been made in the application of adaptive artificial intelligence methods to address concept drift in pandemic-related datasets, several important research gaps remain. Identifying these gaps is crucial for guiding future work,

enhancing the robustness and applicability of predictive models, and ultimately improving public health responses in future pandemics. Below, we outline critical limitations in current methodologies and practices, offering clear recommendations for future research directions to address these challenges systematically.

1. Standardisation of drift detection and quantification methods: Future research should focus on developing standardised methods for reliably detecting and quantifying drift, providing clearer guidelines for model retraining and adaptation. This thesis addresses the necessity of detecting drift in Chapter 8.
2. Comparative benchmarking of adaptive methods: Systematic comparative studies evaluating multiple drift-handling techniques under consistent epidemiological scenarios are urgently needed to establish best practices clearly. This necessity is tackled in Chapter 6, where an extensive comparison of state-of-the-art adaptive algorithms is presented for a COVID-19 dataset.
3. External validation and generalisation: Most current studies have a localised scope; further work should emphasise external validation across diverse geographic and demographic contexts to ensure robustness and generalizability.
4. Integration of Explainable AI: Incorporation of explainable methodologies (e.g., SHAP values) into adaptive models should be prioritised to enhance transparency, build trust among end-users, and improve clinical and policy decision-making. This thesis aims to answer specifically to this necessity by proposing an XAI-based monitoring system in Chapter 8.
5. Real-time adaptive systems in clinical practice: Future research should explore the integration of drift-aware AI systems within clinical environments, focusing on real-world feasibility, usability by healthcare professionals, and quantifiable improvements in healthcare outcomes. This thesis contributes to this field in Chapter 5 by developing an ML method capable of automatically adapting to changes in clinical data.
6. Secure and effective data-sharing techniques: Advances in federated learning and privacy-preserving algorithms should be further developed and systematically evaluated to enhance collaborative data-sharing efforts in public health emergencies.
7. Standardised notation and explicit drift classification: There is a clear need for unified terminology and standardised notation to describe concept drift consistently across studies. Furthermore, explicit identification and reporting of drift types (e.g., abrupt, gradual, incremental, or recurring) beyond merely stating their causes would significantly enhance the clarity and comparability of research findings.
8. Interdisciplinary collaboration: There is a clear need for greater interdisciplinary collaboration among epidemiologists, clinicians, data scientists, and policymakers to ensure the continuous and effective integration of domain expertise into adaptive AI systems for pandemics.
9. Simulation-based approaches: Future research should explore simulation frame-

works that enable the controlled study of concept drift under realistic pandemic scenarios. Simulation environments can facilitate reproducibility, allow systematic stress-testing of adaptive models, and support the development of drift-aware policies before their deployment in real-world settings. Despite their potential, such approaches are virtually absent in the current literature.

10. Integration of MLOps pipelines: The long-term sustainability of drift-aware AI systems requires robust operational frameworks. Incorporating MLOps practices, including automated retraining, continuous monitoring, version control, and deployment pipelines, would provide the necessary infrastructure to maintain predictive performance as pandemic conditions evolve. This integration remains unexplored in the reviewed works and represents a critical step toward real-world applicability.

4

Data Sources and Datasets

This chapter describes the datasets used throughout the thesis, together with the corresponding cohort construction criteria, variable definitions, and preprocessing decisions. Since the proposed contributions are evaluated under different aspects of interest, including clinical phenotype prediction, learning under imbalanced conditions and learning under concept drift, the data sources are organised into four groups. This organisation allows subsequent chapters to focus on methodological aspects and results, while referring back to this chapter for dataset-specific details.

First, we present the clinical registry from the Regional Health System of Murcia (Spain) used for COVID-19 related analyses and for the study of the Long COVID-19 (LC-19) cohort (Section 4.1). LC-19, is defined as a multisystemic condition characterised by the persistence of symptoms or the emergence of new clinical manifestations following the acute phase of a SARS-CoV-2 infection, typically lasting for at least four weeks. Due to its role in the clinical motivation of this thesis, this data source is not merely described. Instead, an in-depth descriptive and statistical analysis is performed on it. This comprehensive analysis is necessary to formally characterise the population, extract epidemiological insights (such as risk factors and the impact of vaccination), and clinically justify the variables that will be used for training the predictive models. Second, we describe the benchmark datasets employed in the imbalanced data experiments that support the evaluation of the proposed methods under varying imbalance ratios (Section 4.2.1). Third, we introduce the datasets and stream generators used in the concept drift experiments, including both real-world and synthetic streams (Section 4.2.2). Finally, we present specific stream-based datasets and industrial simulators explicitly utilised for the visual monitoring of feature importance under changing environments (Section 4.2.3).

4.1 COVID-19 and Long COVID-19 Clinical Dataset

The studies presented throughout this thesis are motivated by a recurrent difficulty in real-world decision-support, as the outcomes of greatest clinical interest are often infrequent, while the data-generating process evolves due to external changes. Long COVID provides a clear example of this setting. Only a minority of infected individuals develop persistent post-acute symptoms, yet identifying higher risk profiles is essential for follow-up planning, early intervention, and resource allocation. In parallel, the pandemic context introduces multiple sources of non-stationarity and successive epidemic waves, shifts in testing and clinical practice, and, critically, vaccination rollouts, making both the clinical analysis and the modelling task inherently dynamic.

Within this context, this chapter introduces the LC-19 cohort and the associated study published in *Scientific Reports* on 28 March 2025 [199]. Beyond characterising demographic and clinical risk factors, a central focus is placed on vaccination variables, including vaccination status and schedule patterns, to understand how immunisation history interacts with post-acute persistence. This emphasis is particularly relevant because vaccines changed not only population-level risk profiles but also the observable clinical landscape, meaning that any realistic analysis must treat vaccination as a main driver of distributional changes and heterogeneous patient trajectories.

In addition to the epidemiological findings, the work provides a practical predictive component designed with clinical usability in mind. The resulting model explicitly accounts for the strong class imbalance typical of long-term complications as a publicly accessible online risk calculator, enabling clinicians to obtain LC-19 risk estimates from routinely available patient information. In the broader scope of the thesis, this study therefore plays a dual role, as it supplies a real dataset that will be reused to illustrate methodological decisions under imbalance, and it grounds the framework's motivation in an outcome that connects research contributions to clinical practice.

4.1.1 Methods

4.1.1.1 Study design, setting, and participants

In this study, we followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist for reporting [200]. The study protocol was approved by the Bioethics Committee of the Universidad de Murcia. Anonymised individual-level data from all patients diagnosed with COVID-19 within the Public Health Care System, collected in a prospective Regional Registry, were provided by the Murcia Health Care Information Department. A waiver for the need of informed consent was obtained in accordance with the General Data Protection Regulation (GDPR).

We conducted a retrospective cohort analysis using a population selected from a database provided by the Servicio Murciano de Salud (SMS). This database includes patients who tested positive for COVID-19 between January 4, 2020, and May 3, 2022, across 10 hospitals in the Region of Murcia, with both LC-19 and non-Long COVID-19 cases. Diagnosis of COVID-19 was confirmed by antigen testing or real-time reverse transcription-polymerase chain reaction (RT-PCR) assays using nasopharyngeal swab samples. RT-PCRs were performed in hospitals by clinicians and technicians, while antigen tests were conducted by nurses and clinicians in primary healthcare settings. The first inclusion criterion was being part of the COVID-19 database and having records related to symptoms, comorbidities, or vaccination status.

According to the Centers for Disease Control and Prevention (CDC) [201], a set of symptoms are more commonly associated with a case of LC-19, which we will refer to as LC-19 symptoms. The available LC-19 symptoms in our database include chest pain, muscle pain, cough, sleep problems, dyspnea, dizziness, hyposmia, hypogeusia, fever, headache, rash and tiredness. The symptoms included in our study were recorded, evaluated, and diagnosed by clinicians in primary healthcare settings or during hospitalisation. Using this information we distinguished three subtypes of LC-19 patients [202] from non-Long COVID-19 patients following an operational definition. The time at which each patient symptom was recorded, evaluated, diagnosed or reported was considered the moment of appearance of the symptoms.

To identify non-Long COVID-19 patients, we selected individuals from the aforementioned database whose symptoms resolved within 28 days of testing positive for COVID-19. These patients' medical records were reviewed to confirm that their symptoms or reasons for follow-up appointments were not consistent with LC-19 symptoms, or that they had been diagnosed with other conditions. Additionally, patients who did not return for further medical attention at primary care or emergency services after their initial recovery were also classified as non-Long COVID-19.

For all patients, we assessed whether they exhibited LC-19 symptoms at least 28 days after their initial COVID-19 infection to identify those with LC-19. Acute post-COVID patients were defined as individuals whose LC-19 symptoms persisted for more than 28 days but less than 12 weeks after testing positive. Long post-COVID patients were those whose symptoms lasted between 12 and 24 weeks, while persistent post-COVID patients had symptoms that extended beyond 24 weeks. We followed the same process of reviewing medical records and appointments for these patients as we did for non-Long COVID-19 patients. All subtypes of LC-19 patients were considered part of the broader LC-19 group.

The study also considered the vaccination status of the population from December 21, 2020, to May 3, 2022. Available data included specific vaccine brands, such as the mRNA vaccines BNT162b2 (Pfizer/BioNTech) and mRNA-1273 (Moderna), as well as the non-replicating viral vector vaccines ChAdOx1 (Oxford/AstraZeneca) and Ad26.COV2.S (Janssen).

For a vaccination schedule to be validated, it had to be recorded at least 21 days prior to the patient's COVID-19 diagnosis.

After identifying each patient category, we assessed risk factors for developing LC-19, including symptoms at the time of the positive test, pre-existing comorbidities, sociodemographic factors, and vaccination status. These characteristics were compared between non-Long COVID-19 patients, LC-19 patients, and across the subtypes of LC-19 patients.

4.1.1.2 Ethics approval and consent to participate

All the analyses of the data reported in this section was done in accordance with the following criteria:

1. Article 157 of the General Data Protection Regulation (RGPD) recognises the benefit that access to records can provide to research on diseases, so that the results of these studies could be more solid, based on a larger population.
2. Article 89.1 of the RGPD confirms this principle as long as the appropriate measures are adopted, in particular respect for the principle of minimisation of personal data. These measures may also include the use of pseudonymised or anonymised data.
3. In turn, article 14.5 b) of the RGPD allows the data controller not to inform the interested parties when "said information is impossible or involves a disproportionate effort in particular for the treatment for purposes of public interest, scientific research purposes or historical or statistical purposes subject to the conditions and guarantees indicated in article 89, paragraph 1, or to the extent that the obligation mentioned in paragraph 1 of this article may make it impossible or seriously impede the achievement of the objectives of such treatment".
4. The seventeenth additional provision of Organic Law 3/2018, of December 5, on the Protection of Personal Data and guarantee of digital rights allows the use of health data for scientific research purposes as long as these data have been submitted to prior treatment of pseudonymisation or anonymisation.

In accordance with the exposed precepts, consent is not necessary for the following reasons:

1. Access to clinical records for scientific research purposes is lawful, as long as certain guarantees are adopted, including respect for the principle of data minimisation and that the information obtained is adequately pseudonymised or anonymised.
2. Obtaining consent would not be a legal requirement, since it would not only entail disproportionate efforts but could also hinder the development of the study due to the large amount of information that has been analysed.

3. The Murcia Health Service, in accordance with the provisions of the aforementioned regulations, has submitted this clinical information to an anonymisation process, excluding any identifying information of the patients, making their identification impossible.

4.1.1.3 Variables and data source

The data used for this study is similar to that explained in the data description and preprocessing section of Cisterna et al. work [203]. However, there is a new database called the 'vaccination database' that contains information on all patients in the Region of Murcia (Spain) during the period under review in this study previously specified. We examined demographic factors, comorbidities, symptoms, hospitalization status, and vaccine information (including the number of doses, administration schedules, and vaccine types) recorded in the COVID-19 database provided by the Regional Health System of Murcia. Data regarding medication and length of stay in each hospital department were available, they were excluded from the analysis as they were deemed to be outside the scope of the study and would have significantly reduced the number of patients included due to low patient availability.

Regarding demographic factors, we have information about biological sex information (male or female) and the age of the patient (discrete quantitative variable). No other sociodemographic information was extracted or used in the analysis.

Regarding comorbidities, we have information from previous comorbidities associated with each patient. These comorbidities were diagnosed before COVID-19 by a clinician. We have information about diabetes mellitus, heart failure, chronic obstructive pulmonary disease, arterial hypertension, depression, ischemic cardiomyopathy, stroke, renal insufficiency, cirrhosis, osteoporosis, osteoarthritis, arthritis, obesity and asthma. All of these comorbidities are dichotomous qualitative variables. Additionally, we have information about the number of chronic diseases and the number of systems affected, which are discrete quantitative variables.

Regarding symptoms, we have information from symptoms at the initial diagnosis of the COVID-19 infection and up to 28 days after the positive test for COVID-19. The symptoms included in our study were recorded, evaluated, and diagnosed by clinicians in primary healthcare settings or during hospitalisation. We have information about chest pain, muscle pain, cough, lack of appetite, sleep problems, nasal congestion, dyspnea, rhinorrhoea, low grade fever, dizziness, hyposmia, hypogeusia, headache, tiredness, eye symptoms, expectorate, fever, sore throat, stomach pain, vomit, nasal discharge, chills, rash and malaise. All of these symptoms are dichotomous qualitative variables.

Furthermore, we have information regarding the severity of patients, as indicated by whether the patient has been hospitalised and, if so, the date of hospitalisation. This

information is recorded by a clinician at the time the patient enters a hospital area.

Regarding vaccines status, we have information about the dosage, the brand of the vaccine (Astra Zeneca, Janssen, Moderna or Pfizer) and the schedule, which is recorded concurrently with the administration of the dose to each individual.

4.1.1.4 Data description and preprocessing

The demographic information and comorbidities of all patients in the population study were obtained from a database used for stratification purposes during the COVID-19 pandemic. Available characteristics include age, sex, number of chronic diseases, number of affected systems, and various medical conditions such as diabetes mellitus, heart failure, chronic obstructive pulmonary disease, arterial hypertension, depression, ischaemic cardiomyopathy, stroke, renal insufficiency, cirrhosis, osteoporosis, osteoarthritis, arthritis, obesity, and asthma. A total of 322303 patients have been studied; however, symptomatology records are available for 297737 patients who have sought care in primary healthcare settings. This data includes features such as lack of appetite, nasal congestion, rhinorrhoea, low grade fever, eye symptoms, expectorate, sore throat, stomach pain, vomit, nasal discharge, chills and malaise. To perform the analyses that are based on the vaccination database, only those patients who received a vaccination 21 days prior to a confirmed diagnosis of COVID-19 infection were included in the study. Additionally, we eliminated patients whose schedule was inconclusive, for example patients that had received several doses on the same day. This left us with vaccine schedule information for 320141 patients. Other data, such as medication (30311 patients) or length of stay in each hospital department (15829 patients), were not included in the analysis as they were deemed to be outside the scope of the study and would have significantly reduced the number of patients included.

4.1.1.5 Bias

To ensure data quality and minimise bias, we designed a rigorous selection process to include individuals with comprehensive descriptions and complete data for all relevant features. The large sample size, derived from a retrospective cohort analysis using a health database, allowed for the exclusion of patients lacking essential information, such as comorbidities or demographic data. Cases with inconclusive or implausible records, such as two vaccine doses administered on the same day, negative age, or undetermined biological sex, were removed to maximize data accuracy and completeness. No NA values were recorded in the database. Therefore, any patient with at least one record in each of the symptom, comorbidity, and vaccine datasets, without any erroneous data, was considered part of the complete sample.

We applied methodologies to mitigate potential confounders and address timing-related

variables. These adjustments were based on the timing of patients' positive COVID-19 test results and the onset of symptoms. We considered the time points when symptoms were first recorded, evaluated, diagnosed, or reported to clinicians to ensure accurate temporal alignment in the analysis.

A potential source of bias was sex, as prior studies suggest females may be at a higher risk of developing LC-19 [204–209]. To address this, we adjusted for sex in our analyses and performed sex-disaggregated assessments to reduce the influence of this bias.

Another potential bias is the lack of detail regarding the progression of symptoms over time. We only considered the presence or absence of symptoms within the first 28 days after a positive COVID-19 test, without accounting for the specific days on which symptoms appeared. We acknowledge that this limitation may affect the accuracy of our model in capturing the temporal dynamics of symptom development.

One more possible source of bias is the time interval between vaccination and the development of immunity. To ensure valid immunization data, we only considered vaccination schedules where the final dose was administered at least 21 days prior to the patient's COVID-19 diagnosis. This criterion was set to account for the time required for the immune response to develop fully, thereby reducing the risk of misclassifying individuals who had not yet achieved full immunity.

4.1.1.6 Statistical analyses

Continuous data are presented as the median with interquartile range (IQR), while categorical data are expressed in percentages (%). Odds ratios (OR) along with 95% confidence intervals (CIs) were also employed. Following a pipeline to detect erroneous data and quantify them, we treated those values as NAs. Given their low proportion (2.54% of the total patients), we excluded these patients from the study, opting for a complete-case analysis. In various analyses, the ORs were adjusted for sex and age (as Adjusted Odds Ratios, AOR), and specifically for all comorbidities and symptoms within the respective analyses of comorbidities and symptomatology. Statistical analyses were conducted using R software (version 4.3.1), with a significance threshold for p-values set at 0.05. The Mann-Whitney U test is employed for the purpose of analysing differences in continuous data between two sets of patients. Differences between groups or conditions were evaluated by computing AOR, the associated 95% CI and p-value using the 'logistic.display' function of the R package *epiDisplay* [210] (version 3.5.0.2) on a generalised linear model (GLM), trained through the 'glm' function using R package *caret* [211] (version 6.0-94), employing logistic regression. This approach employs logistic regression to predict a binary outcome from a set of predictors. The 'logistic.display' function is used to determine the weights associated with each variable, which in turn allows the AOR [95% CI] and p-values for all predictors to be calculated.

AORs were also employed to examine differences among various vaccination schedules in the cohort, assessing whether specific schedules offer greater or lesser protection compared to others. We generated a heatmap that allows for pairwise comparisons of vaccination schedules, enabling the identification of statistically significant differences. We used p-values with Bonferroni correction to denote these differences and indicated which schedules are statistically significant (*) more protective or less protective. Additionally, an OR study was conducted to assess the comparative efficacy of mRNA-based vaccines and viral vector-based ones for the prevention of LC-19. This analysis was adjusted by age, sex, and the number of doses received. A further analysis was conducted to determine whether a combination of an mRNA-based vaccine and a COVID-19 viral vector-based vaccine is more protective than two separate vaccines for suffering LC-19. Two further analyses were conducted to ascertain whether a combination of one COVID-19 viral vector-based and one mRNA-based vaccine is more efficacious than two mRNA-based vaccines or two COVID-19 viral vector-based vaccines for suffering LC-19, with both analyses adjusted for sex and age.

In order to obtain AORs in the study, we used logistic regression models. It is calculated as follows:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

,

where p is the probability of being a LC-19 patient, X_j any predictor and β_j its associated weight. The following analyses were conducted:

- To obtain adjusted odds ratios for all comorbidities (diabetes mellitus, heart failure, chronic obstructive pulmonary disease, arterial hypertension, depression, ischaemic cardiomyopathy, stroke, renal insufficiency, cirrhosis, osteoporosis, osteoarthritis, arthritis, obesity, and asthma) and for sex, we employed a single logistic regression, where every comorbidity and the sex variable is adjusted by age, sex and the rest of comorbidities. In this model, each X_j is a comorbidity, age, and sex. This resulted in a model with 17 predictors, 15 of which were comorbidities.
- To obtain adjusted odds ratios for the symptomatology data (chest pain, muscle pain, cough, lack of appetite, sleep problems, nasal congestion, dyspnea, rhinorrhoea, low grade fever, dizziness, hyposmia, hypogeusia, headache, tired, eye symptoms, expectorate, fever, sore throat, stomach pain, vomit, nasal discharge, chills, rash and malaise), we also employed a single logistic regression, where every symptom is adjusted by age, sex and the rest of symptoms. In this model, each X_j is a symptom, age, and sex. This resulted in a model with 26 predictors, 24 of which were symptoms.
- When analysing the risk of suffering LC-19 if a patient is hospitalised, a logistic regression is also utilised. The present study has been conducted on the basis of

disaggregation by sex and with both sex included, so that here we trained 3 logistic regression models. In the first two models (males - females), X_1 is the hospitalisation condition, and X_2 is the age variable. In the last model, X_1 represents the hospitalisation condition, and X_2, X_3 represent the age and sex variables.

- A logistic regression model is constructed for each AOR reported in Table 4.3 (Table of vaccination differences for non LC-19 and LC-19 patients), with X_1 representing each row analysis, and X_2 , X_3 represent the age and sex variables. So here we trained 10 logistic regressions. Then, in these experiments, we have adjusted by age and sex.
- A logistic regression model is also constructed for each AOR reported in Figure 4.8. In each model, X_1 represents the vaccination schedule administered in each patient between those two that are being compared, while X_2, X_3 represent the age and sex variables. So here we trained 210 logistic regressions. Then, in these experiments we have adjusted by age and sex.

4.1.2 Results

4.1.2.1 Participants

A total of 675126 patients who tested positive for COVID-19 between January 4th, 2020 and May 3rd, 2022 were available in the study (figure 4.1). The study excluded patients for whom basic information was unavailable, such as comorbidities or demographic data (number of patients 'n' = 329300), as well as those for whom erroneous values were present in any of the available records (n = 17163). Those erroneous values were identified by searching patients with at least one of the following constraints: negative age values, age values greater than 130, unknown sex, or multiple administrations of the vaccine dose in a single day. Furthermore, patients who have experienced reinfection are excluded from this study, as they are not the primary focus of interest (n = 6360). For all patients remaining (n = 322303), we checked if they exhibited LC-19 symptoms at least 28 days after the initial infection date to determine which patients had LC-19 (n = 4360). According to LC-19 categorisation definitions, a patient is assessed as an acute post-COVID case if any LC-19 symptom has persisted for more than 28 days but less than 12 weeks after the COVID-19 infection date (n = 1058). They are considered a long post-COVID case if they exhibit any LC-19 symptoms persistence between 12 and 24 weeks after the infection (n = 590), and a persistent post-COVID case if LC-19 symptoms persist beyond 24 weeks from the infection (n = 2807). The mean follow-up time for LC-19 patients is 358 days, whereas for those non-LC-19 patients, the mean follow-up time is 160 days.

Since this study is a retrospective cohort analysis, the availability of patient data depended on the completeness of electronic health records, rather than on predefined data collection. Our inclusion criteria required patients to have a confirmed positive COVID-

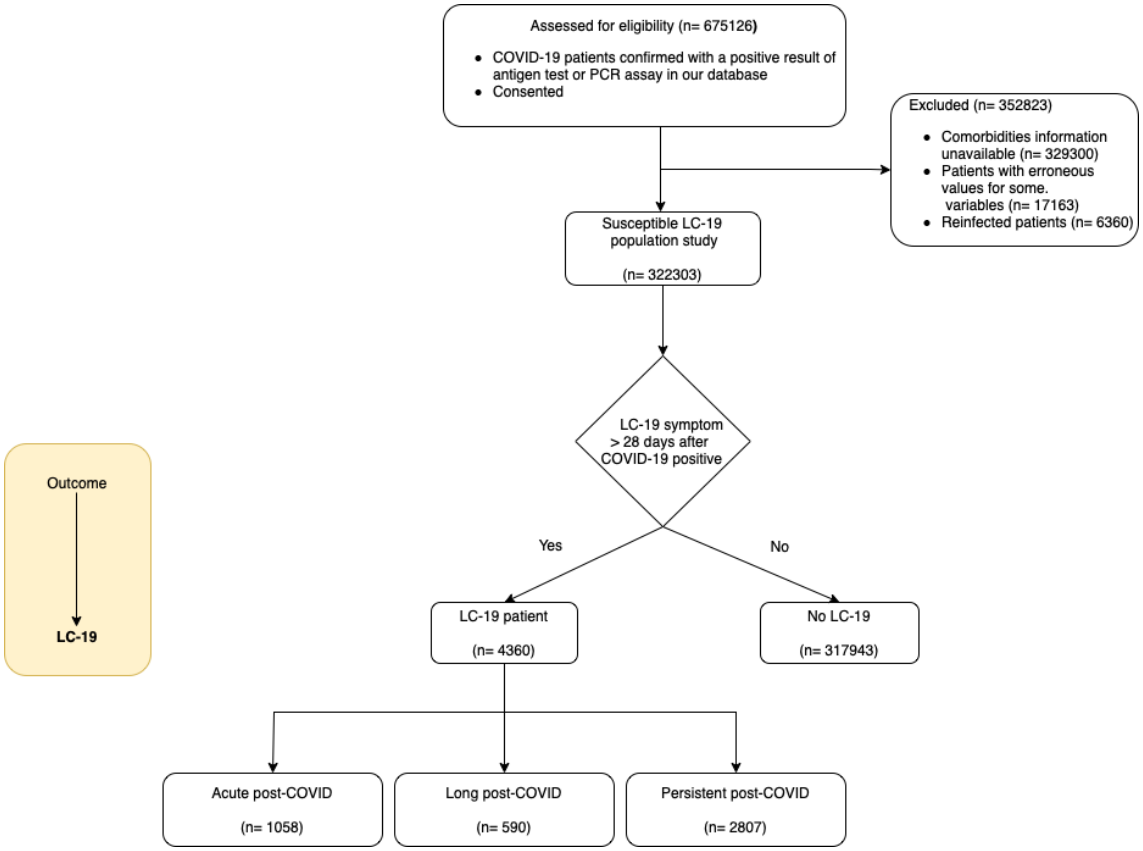


Figure 4.1: *Patient Selection Flow diagram. Flow diagram of the enrollment of subjects, disposition status, and how they are analysed in the study.*

19 test, along with documented information on demographics, symptoms, comorbidities, and vaccination status. From the initial dataset of 675,126 patients with confirmed COVID-19 diagnoses, we identified that 329,300 patients lacked all data on comorbidities (16 variables) and demographic information (e.g., age and sex). Given that these variables were critical for our analyses, patients without this information did not meet the inclusion criteria and were therefore excluded. This exclusion was not a standard procedure for handling missing data within an established study population, but rather a necessary step to ensure that all included patients had sufficient data for valid statistical comparisons aligned with the study's objectives. Due to the large size of the dataset, the exclusion of these cases did not affect statistical power, but was essential to maintain the integrity and reliability of our analyses. The missing data pattern indicates that the lack of information was not missing at random or missing completely at random [212], but rather a consequence of structural gaps in the health records where the covariates did not exist. Including these cases would have introduced significant bias due to the absence of essential covariate data.

4.1.2.2 Analysis of sociodemographic factors, comorbidities and symptoms

Following the classification of LC-19 patients proposed by Fernández-de-las-Peñas et al. [202] we conducted an exploratory analysis of sociodemographic factors, comorbidities and symptoms to seek for differences between non-Long COVID-19 and LC-19 patients (Table 4.1 and 4.2). The main difference between non-Long COVID-19 and LC-19 patients regarding sociodemographic factors was sex (Female %: 52.5% v. 61.1 % respectively). Being female increased the likelihood of developing LC-19 (AOR [95% CI] = 1.39 [1.31, 1.48], adjusted p <0.001). No differences in sex composition were observed between the subtypes of LC-19 patients. The age median was very similar for non-Long COVID-19 and those with the condition (38 v. 37, respectively). However, we observed age differences among the subtypes of LC-19 patients; acute post-COVID patients were older than both long post-COVID and persistent post-COVID patients (adjusted p <0.001).

Table 4.1: *Demographic characteristics and comorbidities for COVID-19 patients (Non-LC-19) and different subtypes of long COVID-19 patients. Continuous data are reported as median with interquartile range (Q3–Q1), and categorical data are expressed as percentages (%). COPD is chronic obstructive pulmonary disease and ICM is Ischemic cardiomyopathy.*

Characteristics	Non-LC-19	LC-19	Acute post-COVID	Long post-COVID	Persistent post-COVID
No. of individuals (N)	317943	4360	1058	590	2807
Age median (IQR)	38 (21, 52)	37 (23, 50)	48 (36, 58)	34 (19.25, 48)	33 (22, 45)
Sex					
Male (%)	151174 (47.5%)	1695 (38.9%)	413 (39%)	237 (40.2%)	1072 (38.2%)
Female (%)	166769 (52.5%)	2665 (61.1%)	645 (61%)	353 (59.8%)	1735 (61.8%)
Comorbidities					
Number of Chronic diseases median (IQR)	2 (1, 4)	3 (1, 5)	4 (2, 6)	3 (1, 4)	2 (1, 4)
Number of systems affected median (IQR)	2 (1, 3)	2 (1, 4)	3 (2, 4)	2 (1, 3)	2 (1, 4)
Diabetes mellitus (%)	20585 (6.5%)	241 (5.5%)	112 (10.6%)	27 (4.6%)	107 (3.8%)
Heart failure (%)	3231 (1%)	26 (0.6%)	15 (1.4%)	4 (0.7%)	7 (0.2%)
COPD (%)	4869 (1.5%)	47 (1.1%)	19 (1.8%)	8 (1.4%)	21 (0.7%)
Arterial hypertension (%)	47036 (14.8%)	550 (12.6%)	233 (22%)	65 (11%)	263 (9.4%)
Depression (%)	29289 (9.2%)	524 (12%)	182 (17.2%)	74 (12.5%)	290 (10.3%)
ICM (%)	5823 (1.8%)	58 (1.3%)	24 (2.3%)	8 (1.4%)	27 (1%)
Stroke (%)	4019 (1.3%)	37 (0.8%)	12 (1.1%)	3 (0.5%)	23 (0.8%)
Renal insufficiency (%)	5950 (1.9%)	69 (1.6%)	24 (2.3%)	12 (2%)	34 (1.2%)
Cirrhosis (%)	6259 (2%)	93 (2.1%)	37 (3.5%)	8 (1.4%)	48 (1.7%)
Osteoporosis (%)	8765 (2.8%)	103 (2.4%)	46 (4.3%)	9 (1.5%)	49 (1.7%)
Osteoarthritis (%)	18534 (5.8%)	284 (6.5%)	106 (10%)	34 (5.8%)	151 (5.4%)
Arthritis (%)	4471 (1.4%)	73 (1.7%)	26 (2.5%)	8 (1.4%)	39 (1.4%)
Obesity (%)	28890 (9.1%)	461 (10.6%)	165 (15.6%)	40 (6.8%)	262 (9.3%)
Asthma (%)	30302 (9.5%)	466 (10.7%)	116 (11%)	70 (11.9%)	292 (10.4%)

Considering sex as an important factor in the development of LC-19 prompted us to analyse comorbidities and symptoms, disaggregated by sex, as illustrated in figure 4.3. In female patients, obesity (AOR [95% CI] = 1.28 [1.13, 1.45]) and depression (AOR [95% CI] = 1.38 [1.23, 1.54]) were identified as comorbidities associated with a statistically significant increased risk for LC-19. Conversely, dementia (AOR [95% CI] = 0.23 [0.1,

0.52]), arterial hypertension (AOR [95% CI] = 0.72 [0.62, 0.83]) and osteoporosis (AOR [95% CI] = 0.75 [0.61, 0.93]) were associated with a reduced risk of LC-19 (figure 4.3A). In male patients, no comorbidities were found to be associated with an increased risk of LC-19, while only stroke (AOR [95% CI] = 0.54 [0.29,0.98]) was associated with a reduced risk of LC-19. We performed the same analysis disaggregated by LC-19 type of patient and no differences were observed (See Figure 4.2).

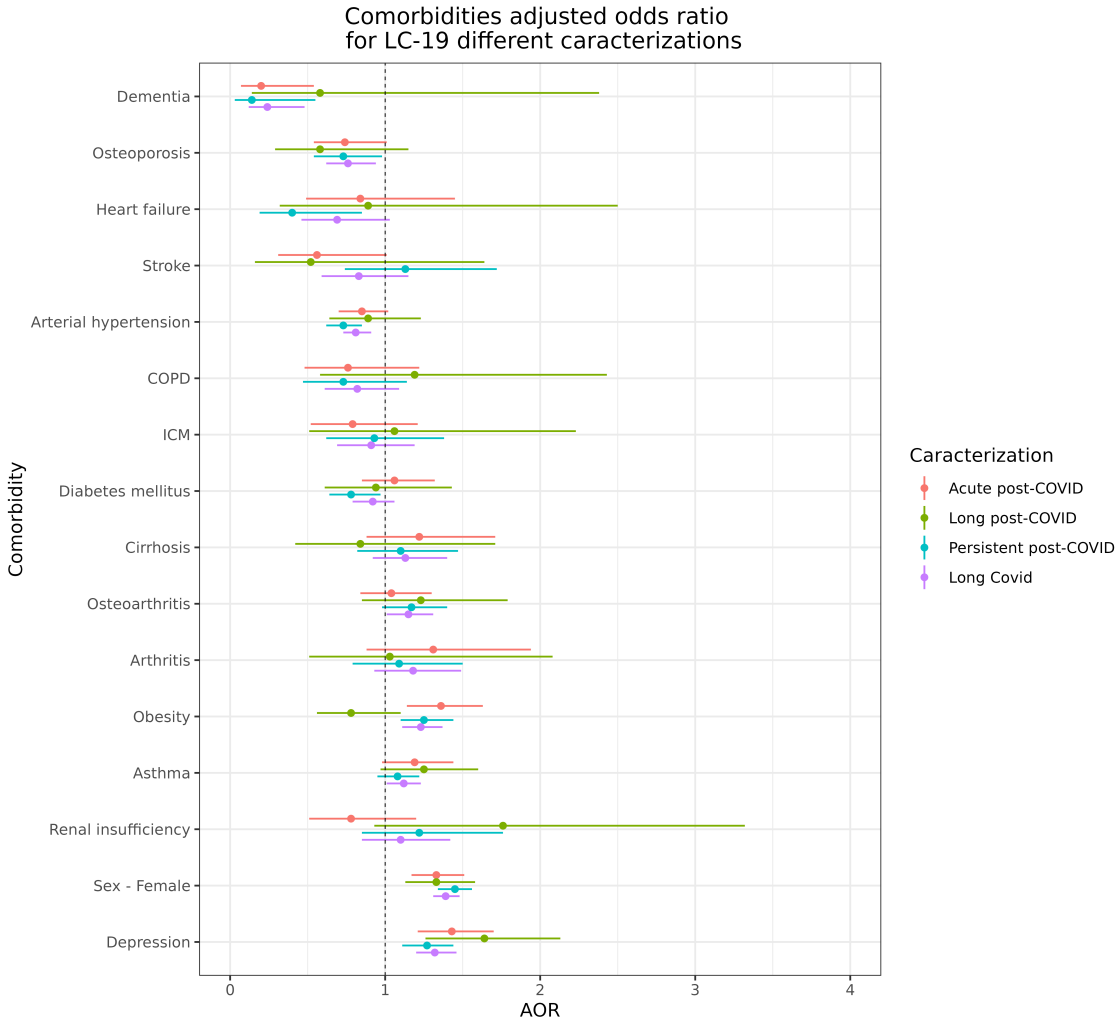


Figure 4.2: *Different types of Long COVID-19 are coded by colours. Odds ratios are adjusted for age, sex, and comorbidities.*

Concerning differences in symptomatology, we noted a higher percentage of initial symptoms in LC-19 patients compared to non-Long COVID-19 patients (Table 4.2 and Figure 4.4). The five symptoms most significantly increasing the likelihood of LC-19 development were chest pain (AOR [95% CI] = 1.67 [1.41, 1.99] for males and AOR [95% CI] = 1.58 [1.40, 1.79] for females), hyposmia (AOR [95% CI] = 1.54 [1.27, 1.87] for males and AOR [95% CI] = 1.57 [1.36, 1.82] for females), and stomach pain (AOR [95% CI] = 1.54 [1.28, 1.86] for males and AOR [95% CI] = 1.52 [1.34, 1.72] for females), followed by rash (AOR [95% CI] = 1.46 [1, 2.12] for males and AOR [95% CI] = 1.73 [1.38, 2.18]) for females), and dyspnea (AOR [95% CI] = 1.72 [1.36, 2.17] for males and AOR [95%

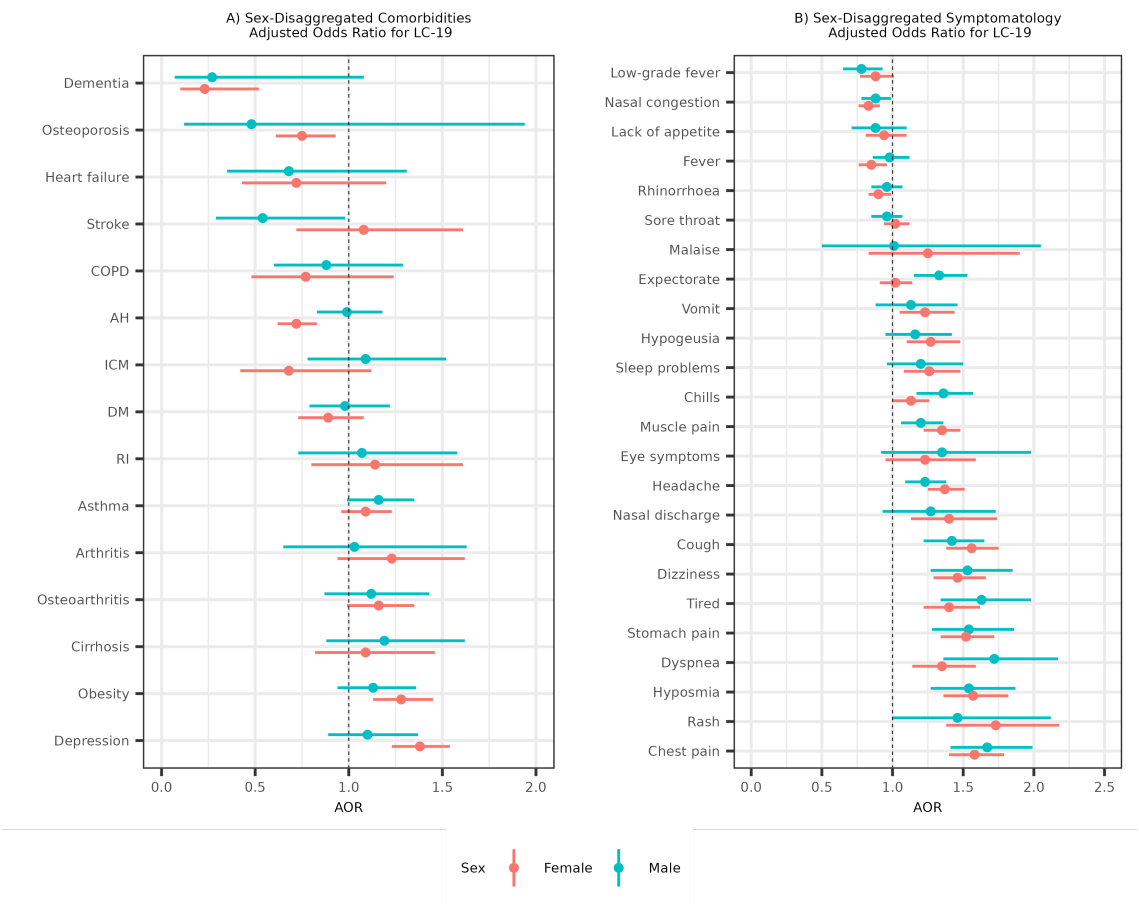


Figure 4.3: *A) Adjusted odds ratios for developing LC-19 based on different comorbidities. B) Adjusted odds ratios for developing LC-19 based on acute-phase COVID-19 symptoms. The female sex is coloured in red and the male sex in blue. Odds ratios are adjusted by age, the symptoms and the comorbidities. COPD is chronic obstructive pulmonary disease, AH is arterial hypertension, ICM is ischemic cardiomyopathy, DM is diabetes mellitus and RI is renal insufficiency.*

CI] = 1.35 [1.14, 1.59] for females) (figure 4.3B). Regarding symptom sex differences, no significant variations were observed in the likelihood of experiencing a particular symptom during the primary COVID-19 infection and subsequently developing LC-19. We conducted the same analysis, stratified by LC-19 patient type, and found that for most symptoms, there were no significant differences among LC-19 patient subtypes (See Figure 4.5). However, hyposmia showed a markedly stronger association with persistent LC-19 patients (AOR [95% CI] = 1.92 [1.66, 2.22], adjusted p <0.005) than with acute post-COVID (AOR [95% CI] = 1.16 [0.93, 1.46], adjusted p >0.05) or long post-COVID patients (AOR [95% CI] = 1.03 [0.72, 1.45], adjusted p >0.05). Similarly, the association of chest pain was notably more pronounced in acute post-COVID patients (AOR [95% CI] = 2.08 [1.77, 2.46], adjusted p <0.001) compared to those with persistent LC-19 (AOR [95% CI] = 1.42 [1.24, 1.62], adjusted p <0.001) and long post-COVID (AOR [95% CI] = 1.28 [0.94, 1.75], adjusted p >0.05).

Table 4.2: *Symptomatology for COVID-19 patients (Non-LC-19) and different subtypes of long COVID-19 patients. Categorical data are expressed as percentages (%).*

Characteristics	Non-LC-19	LC-19	Acute post-COVID	Long post-COVID	Persistent post-COVID
No. of individuals (N)	293377	4360	1058	590	2807
Symptoms					
Chest pain (%)	12117 (4.2%)	616 (14.1%)	266 (25.1%)	55 (9.3%)	309 (11%)
Muscle pain (%)	77623 (27%)	1993 (45.7%)	611 (57.8%)	209 (35.4%)	1226 (43.7%)
Cough (%)	31338 (10.9%)	1234 (28.3%)	494 (46.7%)	128 (21.7%)	650 (23.2%)
Lack of appetite (%)	9970 (3.5%)	518 (11.9%)	251 (23.7%)	49 (8.3%)	239 (8.5%)
Sleep problems (%)	7526 (2.6%)	471 (10.8%)	242 (22.9%)	43 (7.3%)	204 (7.3%)
Nasal congestion (%)	75252 (26.2%)	1377 (31.6%)	352 (33.3%)	163 (27.6%)	896 (31.9%)
Dyspnea (%)	4630 (1.6%)	379 (8.7%)	210 (19.8%)	33 (5.6%)	147 (5.2%)
Rhinorrhoea (%)	87434 (30.5%)	1553 (35.6%)	385 (36.4%)	205 (34.7%)	1008 (35.9%)
Low grade fever (%)	23368 (8.1%)	411 (9.4%)	101 (9.5%)	57 (9.7%)	266 (9.5%)
Dizziness (%)	10290 (3.6%)	523 (12%)	211 (19.9%)	58 (9.8%)	277 (9.9%)
Hyposmia (%)	41460 (14.4%)	1427 (32.7%)	349 (33%)	127 (21.5%)	990 (35.3%)
Hypogeusia (%)	37582 (13.1%)	1320 (30.3%)	338 (31.9%)	121 (20.5%)	901 (32.1%)
Headache (%)	87205 (30.4%)	2158 (49.5%)	629 (59.5%)	250 (42.4%)	1337 (47.6%)
Tired (%)	16097 (5.6%)	892 (20.5%)	402 (38%)	92 (15.6%)	432 (15.4%)
Eye symptoms (%)	1930 (0.7%)	102 (2.3%)	40 (3.8%)	15 (2.5%)	50 (1.8%)
Expectorate (%)	27041 (9.4%)	776 (17.8%)	304 (28.7%)	89 (15.1%)	407 (14.5%)
Fever (%)	37070 (12.9%)	706 (16.2%)	199 (18.8%)	98 (16.6%)	433 (15.4%)
Sore throat (%)	72993 (25.4%)	1467 (33.6%)	409 (38.7%)	187 (31.7%)	918 (32.7%)
Stomach pain (%)	11128 (3.9%)	503 (11.5%)	181 (17.1%)	54 (9.2%)	287 (10.2%)
Vomit (%)	7360 (2.6%)	284 (6.5%)	105 (9.9%)	29 (4.9%)	160 (5.7%)
Nasal discharge (%)	4558 (1.6%)	170 (3.9%)	72 (6.8%)	16 (2.7%)	92 (3.3%)
Chills (%)	23910 (8.3%)	794 (18.2%)	290 (27.4%)	84 (14.2%)	450 (16%)
Rash (%)	2163 (0.8%)	124 (2.8%)	44 (4.2%)	12 (2%)	71 (2.5%)
Malaise (%)	859 (0.3%)	38 (0.9%)	20 (1.9%)	4 (0.7%)	16 (0.6%)

4.1.2.3 Impact of COVID-19 hospitalisation on the development of LC-19

To understand the implications of COVID-19, the need of hospitalisation serves as a crucial indicator, shedding light on the disease's progression and severity across diverse patient profiles. This analysis focuses on the AORs for developing LC-19 among hospitalised and non-hospitalised patients, paying close attention to the interplay between sex and disease severity. We particularly differentiate among various LC-19 patient classifications through their impact on health outcomes. Illustrated in figure 4.6, our analysis offers a clear depiction of the odds associated with LC-19. Notably, hospitalised patients in the LC-19 category exhibit an AOR [95% CI] = 3.57 [3.22, 3.95], with an adjusted p <0.001, adjusted by age and sex. Upon disaggregation by sex, an AOR [95% CI] = 4.49 [3.85, 5.23], with an adjusted p <0.001, was observed for males and an AOR [95% CI] = 3.05 [2.66, 3.51], with an adjusted p <0.001, was noted for females, suggesting a significantly higher likelihood of developing LC-19 overall compared to non-hospitalised counterparts. Specifically, acute post-COVID conditions present a markedly increased risk for hospitalised patients (AOR [95% CI] = 10.26 [8.83, 11.92], adjusted p <0.001), especially for males (AOR [95% CI] = 15.56 [13.08, 20.97], adjusted p <0.001) compared to females (AOR [95% CI] = 7.36 [6.02, 9.01], adjusted p <0.001). Similarly, for persistent post-COVID conditions, a statistically significant higher risk is observed for hospitalised

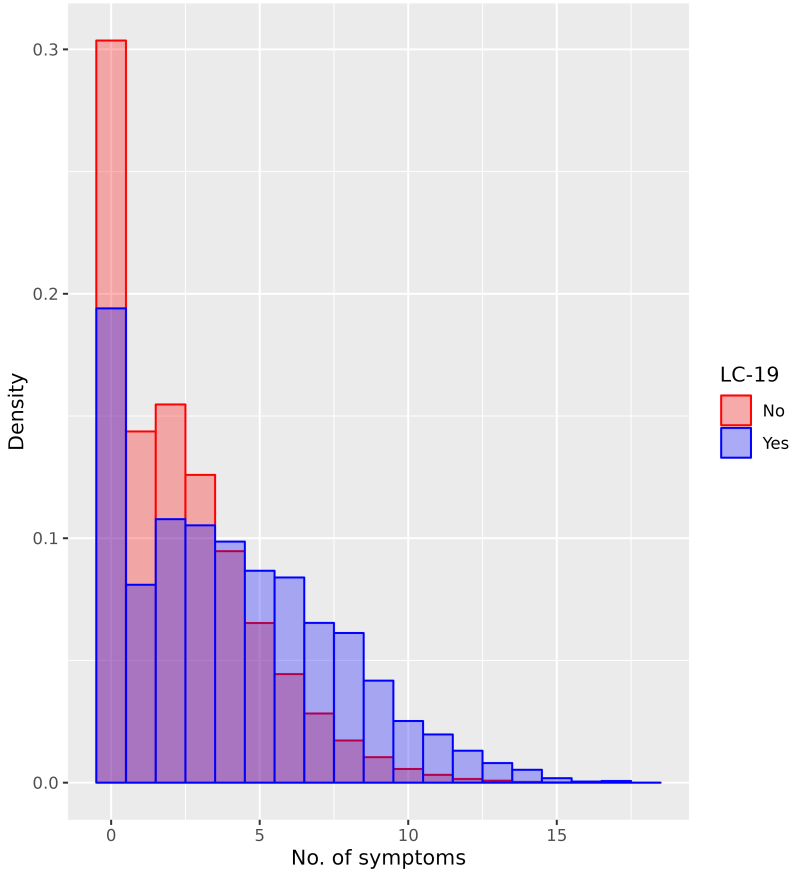


Figure 4.4: *Distribution of the number of symptoms for LC-19 (blue) vs. non-LC-19 (red) patients during primary infection.*

patients of both sex (AOR [95% CI] = 1.55 [1.29, 1.85], adjusted p <0.001). For Long post-COVID conditions, hospitalised patients face a considerable risk (AOR [95% CI] = 1.75 [1.21, 2.53], adjusted p <0.005), with this increased risk being significant solely for male patients (AOR [95% CI] = 2.44 [1.46, 4.07], adjusted p <0.001). This suggests that hospitalisation is not only a high-risk factor for developing long post-COVID and persistent post-COVID manifestations of LC-19 classifications but is particularly crucial for acute post-COVID patients. The AOR for acute post-COVID patients is significantly higher than for other LC-19 classifications.

4.1.2.4 Effects of vaccination on LC-19

We examined the effects of COVID-19 vaccination, including the impact of dosage, patterns, and brand differences on the course of LC-19. Only 15.95% of the cohort is unvaccinated. The most commonly used vaccine brands, in our data, for the first administration are Pfizer (72.39%) and Moderna (15%), followed by Astra Zeneca (8.22%) and Janssen (4.39%). The initial dominance of Pfizer decreases over time. This is shown by the transition from first to second doses. Pfizer still leads, but with a reduced majority of 67.07% of the second dose administration, and it is followed by Moderna (24.25%), Astra Zeneca

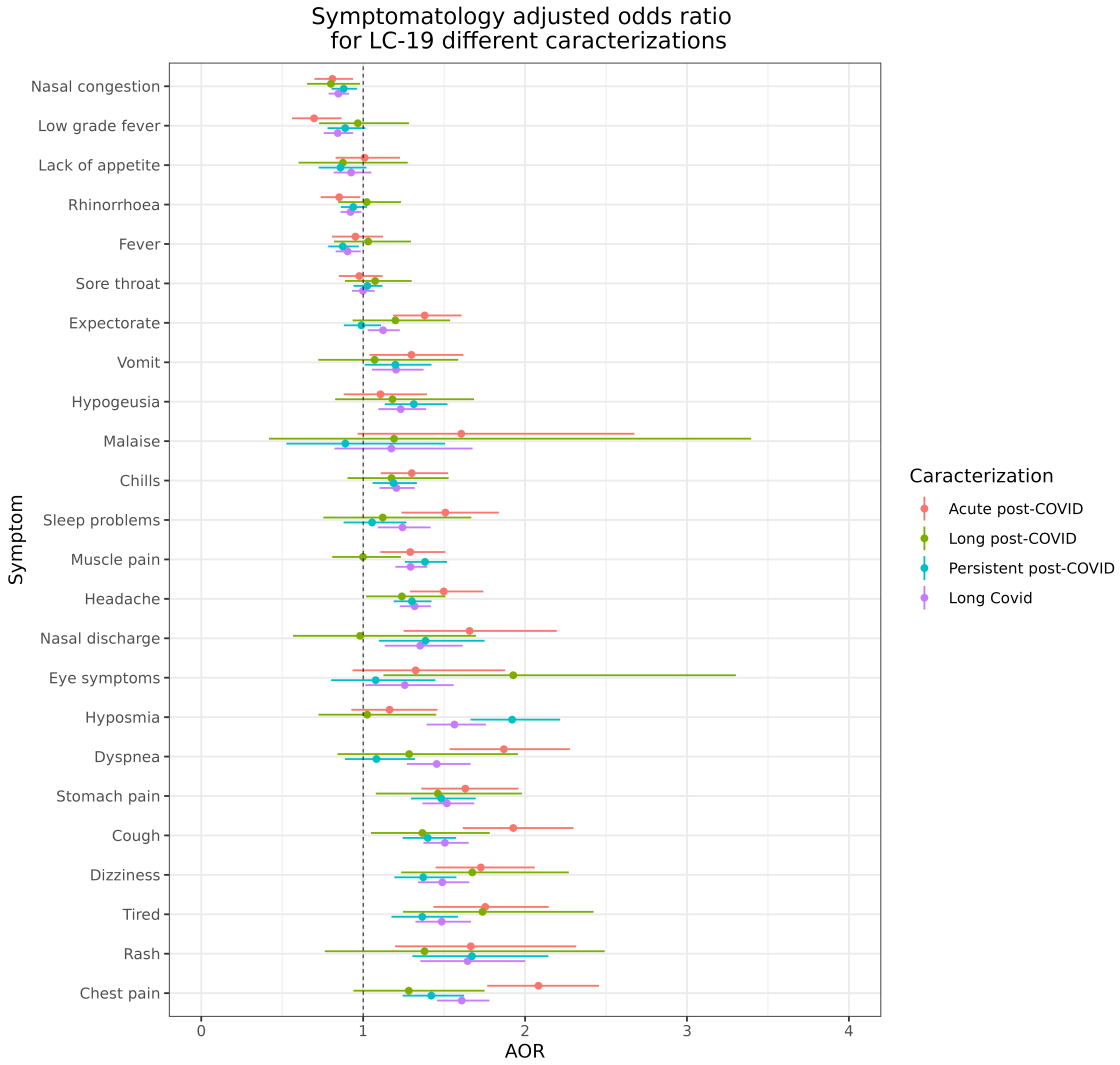


Figure 4.5: *Different types of Long COVID-19 are coded by colours. Odds ratios are adjusted for age, sex, and symptoms.*

(8.68%) and Janssen (<0.01%). Meanwhile, 33.6% of the cohort did not receive a second dose. Furthermore, 80.46% of the population has not yet received a third dose. Among the available brands, Moderna has administered the highest proportion of third doses at 80.20%, followed by Pfizer at 19.79% and Astra Zeneca (0.01%). No Janssen doses were recorded during the administration of the third dose.

When evaluating the protective effect of vaccination in relation to the health impact of COVID-19, the results are significantly influenced by vaccination status. Table 4.3 demonstrates a clear protective effect of vaccination vs no vaccine against LC-19, after adjusting by age and sex the AOR [95% CI] = 0.10 [0.09, 0.12], and adjusted p <0.001. The analysis of vaccine doses and their impact on LC-19 incidence reveals a notably more pronounced protective effect when comparing two doses against one dose (AOR [95% CI] = 0.38 [0.29, 0.50], adjusted p <0.001) and even three doses against both two doses (AOR [95% CI] = 0.39 [0.27, 0.58], adjusted p <0.001) and one dose (AOR [95% CI] =

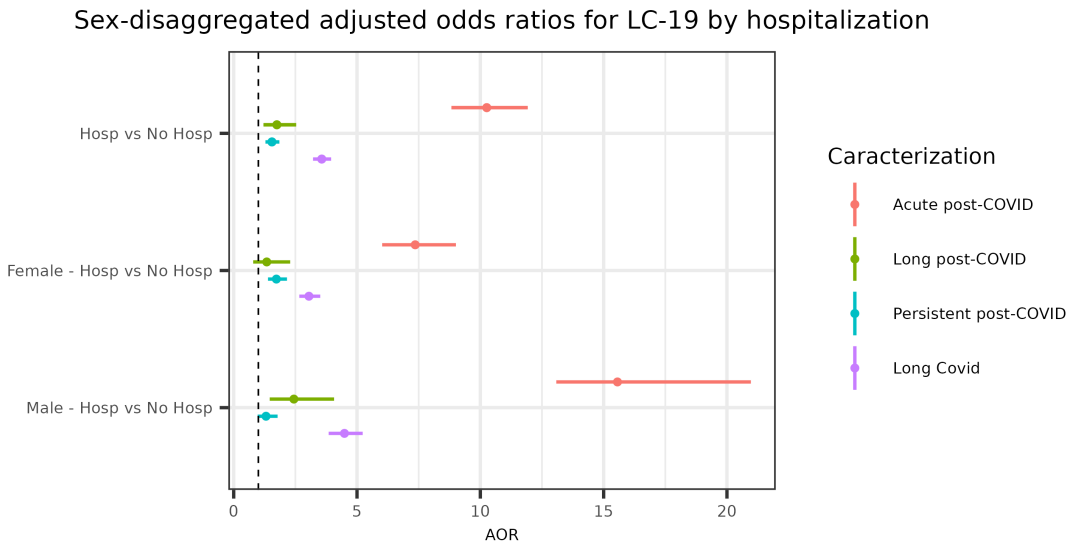


Figure 4.6: *Odds ratio for suffering each categorisation of LC-19 adjusted by age with and without sex-disaggregated.*

0.11 [0.06, 0.18], adjusted p <0.001). We also observed that mRNA vaccines provided a more protective effect against LC-19 than viral vector-based vaccines (AOR [95% CI] = 0.48 [0.37, 0.63], adjusted p <0.001). We performed the analysis of mRNA vaccines vs viral vector-based vaccines disaggregated by sex and found this effect to be statistically significant only in women (See Figure 4.7). Considering the effect of mRNA vaccines, we studied a combination of mRNA and viral vector-based vaccines, comparing them against two doses of either mRNA or viral vector-based vaccines alone. A mix of mRNA and viral vector-based vaccines in any order was more protective than two doses of viral vector-based vaccines (AOR [95% CI] = 0.3 [0.15, 0.58], adjusted p <0.001) and as protective as two doses of mRNA vaccines against LC-19 (AOR [95% CI] = 0.77 [0.41, 1.46], adjusted p = 0.43).

Furthermore, we investigated the effects of various COVID-19 vaccination schedules, segmented by vaccine brand, on LC-19 incidence (figure 4.8). AORs were utilised to determine the relative protection offered by specific vaccination schedules. Figure 4.8 summarises the protective effects of different vaccination schedules against LC-19. Initially, it was found that any vaccination schedule provides more protection against the development of LC-19 compared to being unvaccinated. The three most protective schedules were Pfizer + Pfizer + Moderna, Pfizer + Pfizer + Pfizer and Janssen + Moderna. Conversely, the least protective vaccine schedules against LC-19 were a single dose of AstraZeneca (no statistically significant protection against LC-19 comparing with no vaccine administration), a single dose of Pfizer, and two doses of AstraZeneca.

Table 4.3: *Table of vaccination differences for non LC-19 and LC-19 patients.*

	LC-19	**No LC-19**		
	N (%)		AOR [95% CI]	Adj p
No vaccinated (Ref)	3916 (2.4%)	160007 (97.6%)	0.10 [0.09, 0.12]	<0.001
Vaccinated	413 (0.3%)	155805 (99.7%)		
COVID-19 viral-based vaccine (Ref)	87 (0.71%)	12096 (99.29%)	0.48 [0.37, 0.63]	<0.001
mRNA-based vaccine	324 (0.23%)	140671 (99.77%)		
2 COVID viral-based vaccine doses (Ref)	62 (0.70%)	8799 (99.30%)	0.30 [0.15, 0.58]	<0.001
COVID viral-based dose + mRNA-based dose	10 (0.20%)	5073 (99.80%)		
2 mRNA-based vaccine doses (Ref)	242 (0.23%)	105482 (99.77%)	0.77 [0.41, 1.46]	0.43
COVID viral-based dose + mRNA-based dose	10 (0.20%)	5073 (99.80%)		
No. of doses				
0 (Ref) vs	3916 (2.4%)	160007 (97.6%)	-	
1	66 (0.6%)	11251 (99.4%)	0.25 [0.19, 0.32]	<0.001
2	314 (0.3%)	119354 (99.7%)	0.11 [0.09, 0.12]	<0.001
3	33 (0.1%)	25200 (99.9%)	0.05 [0.03, 0.07]	<0.001
1 (Ref) vs	66 (0.6%)	11251 (99.4%)	-	
2	314 (0.3%)	119354 (99.7%)	0.38 [0.29, 0.50]	<0.001
3	33 (0.1%)	25200 (99.9%)	0.11 [0.06, 0.18]	<0.001
2 (Ref) vs	314 (0.3%)	119354 (99.7%)	-	
3	33 (0.1%)	25200 (99.9%)	0.39 [0.27, 0.58]	<0.001

4.1.2.5 LC-19 calculator tool

A predictive model was developed to predict the probability of suffering from LC-19, depending on predictors like comorbidities, sex, age, the vaccination schedule up to 21 days before the positive diagnosis of COVID-19, and the symptomatology of the individual at the initial diagnosis of the COVID-19 infection and up to 28 days after the positive test for COVID-19. Only data collected prior to the LC-19 diagnosis were considered for training the model. Only patients with all the above information available were considered for this objective. The training dataset (199527 no LC-19 patients and 3031 LC-19 patients) was used to train the model and the test dataset (85511 no LC-19 patients and 1298 LC-19 patients) was used to evaluate the model.

IPIP [203] (an ensemble-based method for imbalanced learning described in Chapter 5) and SMOTE [213] were used to deal with the imbalanced proportion of LC-19 and no LC-19 patients together with logistic regressions as ML algorithms. IPIP with logistic regression obtained a balanced accuracy of 73% and an Area Under the Receiver Operating Characteristic Curve (rocAUC) of 0.80 on the test dataset and for Smote with logistic regression we obtained a balanced accuracy of 65% and a rocAUC of 0.79. Then, we used IPIP with logistic regression models to create an online LC-19 calculator that is accessible at `https://provia.inf.um.es/longcovid`.

In order to utilise the calculator to get a probability of suffering LC-19, it is necessary to input the patient's sex, age, comorbidities, and symptomatology during the acute-phase of COVID-19, as well as their vaccination schedule prior to 21 days of infection.

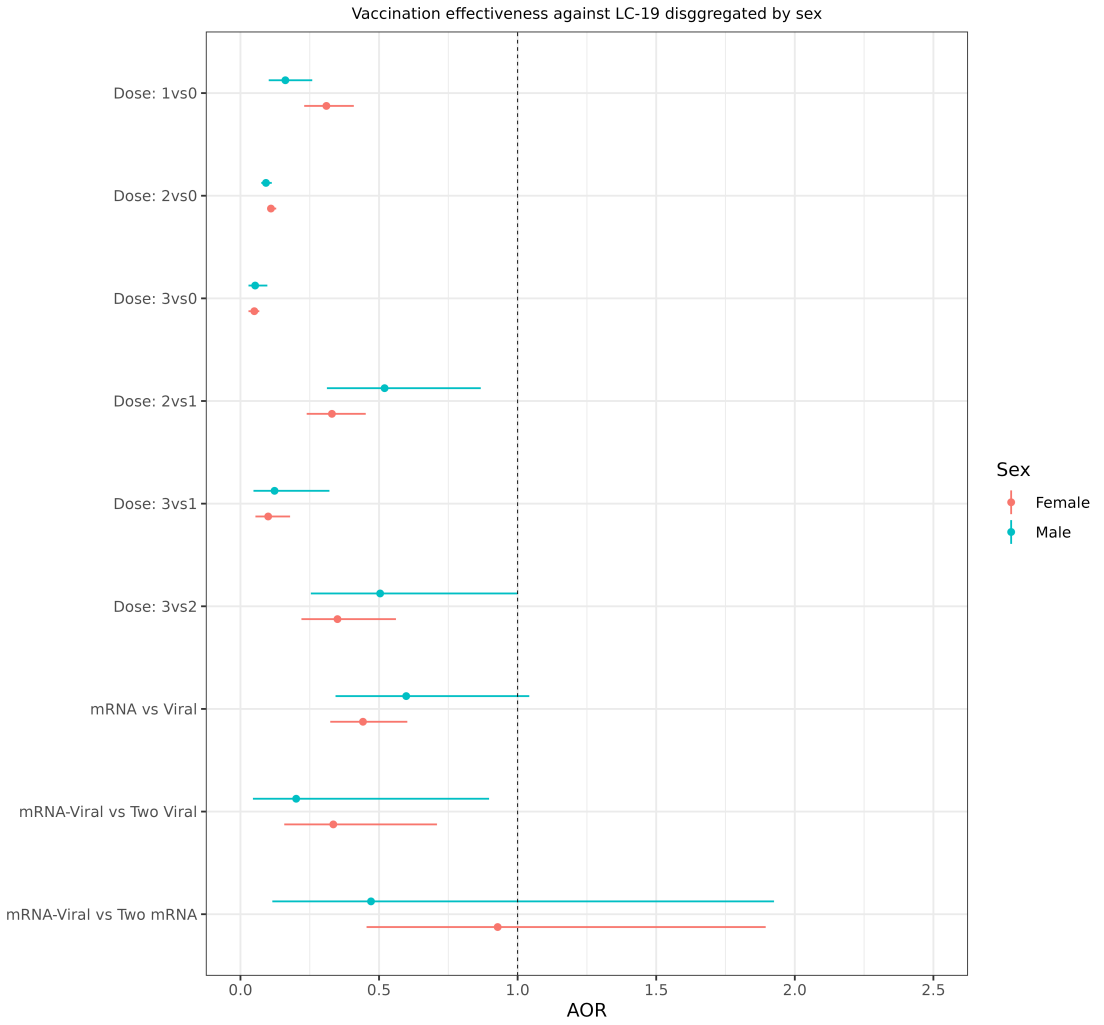


Figure 4.7: *The dose comparison is represented as the first number vs. the second number; for example, 'Dose: 1vs0' shows the odds ratio of having 1 dose of any vaccine compared to 0 doses. Female patients are represented in red, and male patients in blue.*

4.1.3 Discussion

In this study, we investigated the demographic factors, comorbidities, and symptoms of the acute phase of COVID-19 as risk factors for LC-19. Our study offers valuable insights into the factors that influence the development of long COVID. Notably, our data support that female sex is associated with a higher likelihood of developing long COVID. Another risk factor is the number of symptoms experienced during the acute phase of COVID-19; patients who exhibit a high number of symptoms are more likely to develop long COVID. Additionally, the need for hospitalisation during the acute phase significantly increases the risk of developing long COVID, emphasising the severity of the initial infection as a key factor in long-term outcomes. Our analysis also shows no strongly significant differences in symptoms and comorbidities across different subtypes of LC-19 patients, suggesting that the manifestation of long COVID is consistent among LC-19 subtypes of patients. Crucially, vaccination before acute SARS-CoV-2 infection signifi-

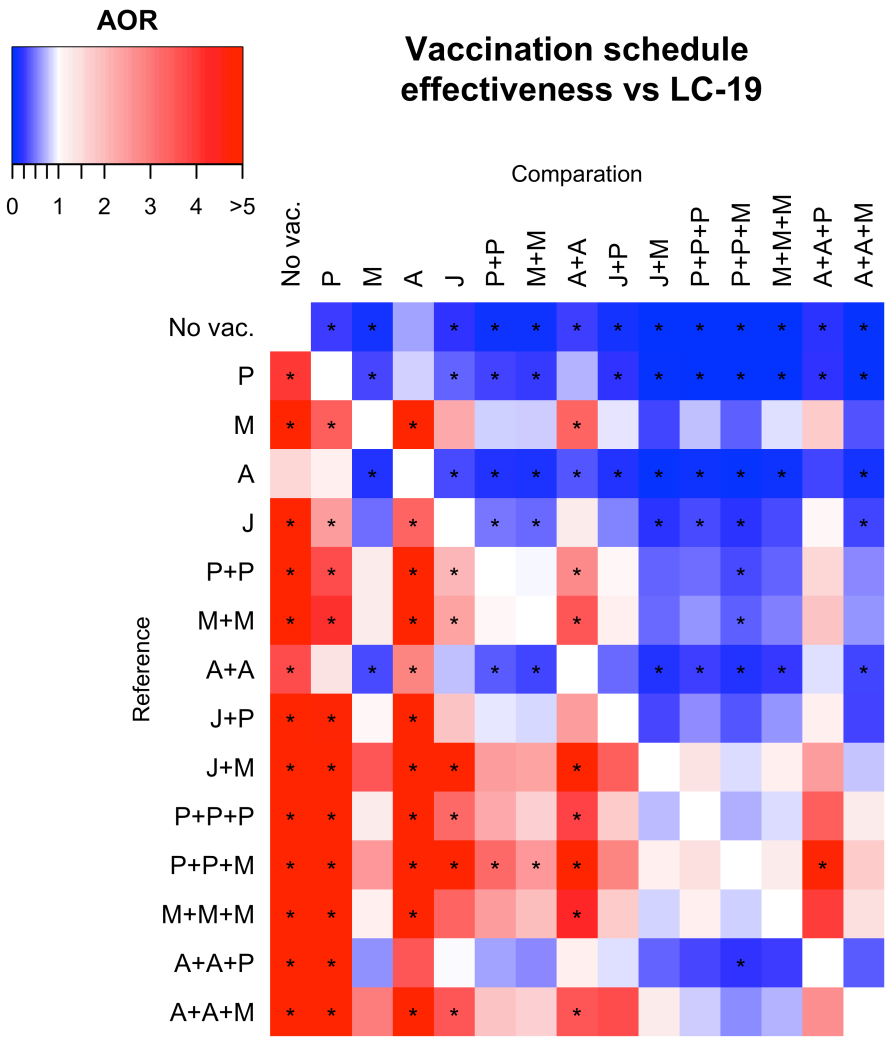


Figure 4.8: *AORs for various vaccination schedules against LC-19. Vaccination schedules are represented by combinations of letters that reflects the order in which the vaccines were administered to participants. With A for AstraZeneca, P for Pfizer, M for Moderna, and J for Janssen vaccines. Blue squares represent greater protective behaviour indicated by lower AORs, when comparing the reference schedules (rows) with the comparison ones (columns), being then the comparison one more protective. In contrast, red squares signify a higher AOR, indicating less protective effect against the outcome. If the adjusted p-value for the AOR is less than 0.05, the comparison is flagged with a star (*).*

cantly reduces the likelihood of developing long COVID. Among the vaccines, mRNA vaccines provide greater protection against long COVID compared to viral vector-based vaccines. Furthermore, a combination of mRNA and viral vector-based vaccines, in any sequence, offers more protection than two doses of viral vector-based vaccines and is as effective as two doses of mRNA vaccines.

Despite the valuable insights provided by our study, several limitations must be acknowledged. First, the number of individuals with LC-19 in our sample is relatively

small compared to non-LC-19 patients. However, the majority of our results are consistent with existing studies. Second, the diagnosis of LC-19 in our study was not clinically confirmed. Instead, we relied on a database derived from health registry data from COVID-19 patients, identifying LC-19 cases based on reported symptoms and follow-up consultations with clinicians. This method may introduce bias due to variability in symptom interpretation, as well as differences in patients' willingness to report symptoms or seek medical care. Another limitation is the lack of detailed historical data on comorbidities. While the database included pre-existing conditions, we lacked information on the duration and timing of these conditions. For example, the database registered whether a patient had experienced a heart attack prior to infection but did not capture when it occurred. Another limitation is that our study did not account for the specific variants of the SARS-CoV-2 virus. Variants may have different impacts on both the likelihood and severity of Long COVID, which was not addressed in our analysis[214–216].

According to our analysis of sociodemographic factors, in the cohort under study female sex is associated with a higher likelihood of developing long COVID. This finding is consistent with other published studies that have identified female sex as a significant risk factor for LC-19 [204–209]. Moreover, the impact of female sex appears to be uniform across all subtypes of LC-19 patients, as illustrated in Figure 4.2 Regarding the risk factors associated with LC-19, we found that comorbidities such as depression and obesity increase the risk of LC-19 in females. This finding aligns with a study predominantly involving female participants (96%) that also identified depression before COVID as a significant risk factor for developing long COVID [217]. Other studies also pointed out obesity as a risk factor for developing LC-19 [218–221]. Additionally, another study in Denmark concluded that obesity and depression are risk factors for LC-19 [222]. On the other hand, we found that female patients with dementia have a lower likelihood of developing LC-19. It is well-documented that patients with dementia have higher odds of dying from COVID-19 [203, 223, 224]. Additionally, COVID-19 appears to increase the likelihood of developing dementia after infection, suggesting a potential association between COVID-19 and dementia [225–227]. However, to the best of our knowledge, there is no evidence indicating whether patients with dementia have higher or lower odds of developing LC-19. We hypothesise that our results could be influenced by the sample size of patients with dementia and the fact that underreporting and undiagnosing are more common among patients with dementia [228–230].

Regarding acute-phase symptoms, we found that patients exhibiting several symptoms during the acute phase are at an increased likelihood of developing LC-19 after a COVID-19 infection. This finding aligns with a multicenter study that identified a greater number of symptoms at hospital admission as the most significant risk factor for developing more symptoms after COVID-19 [205]. In detail, we found that the acute-phase symptoms most strongly associated with an increased likelihood of developing LC-19 are chest pain, rash, hyposmia, dyspnea, and stomach pain. Our findings align with a Polish study, indicating that chest pain and dyspnea are risk factors for the development

of LC-19 [209]. The same symptoms that we previously identified were also reported as risk factors in a large cohort of non-hospitalised COVID-19 patients [221]. Some studies which examine cohorts of individuals from different age groups have identified chest pain, hyposmia and dyspnea as risk factors for suffering from LC-19, not only in adults but also in children [231, 232].

Another relevant question, due to ambiguous conclusions, is whether the severity of the COVID-19 infection influences the development of LC-19. Some studies have concluded that the severity of the acute disease does not significantly contribute to the risk of developing long COVID[204, 233, 234]. However, the majority of studies have found that the severity or need for hospitalisation during the acute phase is associated with a higher likelihood of developing LC-19 symptoms [235–239]. Our results align with these latter studies, suggesting a significantly higher likelihood of developing LC-19 for hospitalised patients compared to non-hospitalised counterparts. This effect is particularly strong in the acute-post COVID phase and moderates over time, as observed for persistent post-COVID in figure 4.6.

Several publications have investigated the impact of vaccination schedules prior to acute SARS-CoV-2 infection in LC-19 [240–247]. These studies support the notion of a protective effect of COVID-19 vaccination against long COVID. Our findings align with these conclusions, indicating that vaccinated individuals have a lower risk of developing LC-19 compared to unvaccinated individuals (AOR [95% CI] = 0.10 [0.09, 0.12]). Furthermore, consistent with some of these previous studies [241, 247], our research suggests that receiving two vaccine doses provides greater protection against LC-19 compared to receiving only one dose (AOR [95% CI] = 0.38 [0.29, 0.58]). In our investigation, we observed that three doses exhibited a more pronounced protective effect against LC-19 than both two doses (AOR [95% CI] = 0.39 [0.27, 0.58]) and one dose (AOR [95% CI] = 0.11 [0.06, 0.18]), respectively. Therefore, the third dose could be beneficial for preventing LC-19.

Another question arises as to whether mRNA vaccines and viral vector-based vaccines exhibit differences in the likelihood of developing LC-19. Some studies [248, 249] provide evidence suggesting a more pronounced protective effect of mRNA vaccines compared to viral vector-based ones in terms of immune response, efficacy, and effectiveness. However, there is limited investigation on this question regarding LC-19. To the best of our knowledge, we found only one publication indicating a slightly stronger preventive effect for BNT162b2 (Pfizer/mRNA) compared to ChAdOx1 (AstraZeneca/viral vector-based) [244]. Consistently, our findings in the cohort corroborate and expand this observation; while both types of vaccines confer protection against LC-19, mRNA vaccines offer a higher degree of protection than viral vector-based vaccines (AOR [95% CI] = 0.48 [0.37, 0.63]). However, according to our results, this effect is only statistically significant in women (Figure 4.7). Additionally, we concluded that, regarding two doses, a mix of mRNA and viral vector-based vaccines in any order was more protective than two doses of viral vector-based vaccines and as protective as two doses of mRNA

vaccines against LC-19. These findings underscore the critical role of vaccination, especially mRNA vaccines, in preventing long COVID and support the use of mixed vaccine regimens in protection. Our study highlights the necessity for ongoing public health efforts to promote vaccination and manage acute COVID-19 symptoms to lessen the burden of long COVID. We also presented a detailed analysis of the effects of various vaccination schedules on LC-19 in Figure 4.8. Compared to not being vaccinated, receiving any vaccination schedule, except a single dose of AstraZeneca, significantly reduces the risk of developing LC-19. Among individuals vaccinated with one dose, those who received Moderna had a lower probability of developing LC-19 than the others. For those with two doses, the Pfizer+Pfizer, Moderna+Moderna, and Janssen+Moderna schedules were the best ones in terms of reducing risk. In individuals with three doses, the Pfizer+Pfizer+Moderna schedule showed the lowest probability of developing LC-19. As illustrated in the figure and previous analysis, individuals with more doses, particularly mRNA vaccines, demonstrated a reduced risk of developing LC-19. These findings are consistent with other studies and provide additional evidence supporting the most protective vaccination combinations[241, 244, 247–249]. However, regarding the three doses effects on LC-19 there are fewer studies [250, 251]. Our results align with those studies on the most protective effect of three doses on LC-19 development and detail the schedules and brand combinations.

As described in Section 4.1.2.5, we also developed a model to predict the likelihood of developing LC-19, made publicly accessible at `https://provia.inf.um.es/longcovid`, achieving a balanced accuracy of 73% and a rocAUC of 0.80. Other studies have also developed predictive models using machine learning algorithms and obtained similar results [252, 253]. For instance, the model developed by Antony et al. achieved a rocAUC of 0.76, although they did not include vaccination records [252]. Kessler et al. developed a more complex predictive model using light gradient boosting machine algorithm, which achieved a rocAUC of 0.84, although they included information about SARS-CoV-2 variants [253]. We aimed to develop a simple and practical model using information collected at the time of COVID-19 diagnosis, allowing clinicians to use it as an effective triage tool. Given the model's accuracy and the study's limitations, it is clear that clinical decisions should not rely solely on this tool. However, we believe it can contribute to closer monitoring of higher-risk individuals, especially in the Murcia Region.

The external validity of our study is supported by the alignment of many findings with previous research. However, several factors could limit its generalisability. Our study is based on a retrospective cohort analysis using health registry data from a specific period. While the large sample size strengthens our conclusions compared to smaller cohort studies, the quality and granularity of our data are not directly comparable to prospective follow-up studies. This study covers the period from January 4, 2020, to May 3, 2022, and the patterns observed may have evolved due to the emergence of new SARS-CoV-2 variants and changes in vaccination strategies. Additionally, our sample is drawn from

a single region, which may limit the external validity of our findings. Differences in demographics, healthcare systems, and pandemic response measures between regions may affect the applicability of our results to other populations. To enhance generalisability, future research should include larger, more diverse populations and examine the impact of SARS-CoV-2 variants and vaccines on Long COVID risk factors across broader contexts.

In summary, the comprehensive clinical characterization of the SMS registry provided in this section serves as more than just a population study. The significant impact of vaccination schedules on risk profiles and the imbalance of COVID-19 and LC-19 cases demonstrate that this is a domain where data is both non-stationary and skewed. Therefore, it is clear that there is a need for machine learning techniques capable of addressing both concept drift and class imbalance to provide reliable clinical decision support. This justification leads to the following sections, where we describe the benchmark and stream-based datasets used to evaluate the predictive framework developed in the subsequent chapters under these specific conditions.

4.2 Benchmark datasets used throughout the thesis

In addition to the SMS clinical registry described in Section 4.1, the experimental evaluation relies on a collection of publicly available benchmark datasets. These datasets are used to (i) study learning and evaluation under controlled class imbalance in stationary settings, (ii) assess robustness and adaptation under non-stationary data streams affected by concept drift, and (iii) evaluate visual monitoring techniques for feature importance in evolving environments. To improve readability and avoid repetition in subsequent chapters, we summarise the benchmarks here and report their main characteristics through compact tables at the end of each subsection.

4.2.1 Datasets from stationary problems

These datasets represent settings where the underlying data-generating process is assumed to be stationary. In these scenarios, the main difficulty arises exclusively from the imbalance between classes, without the added complexity of temporal changes. These benchmarks support two complementary evaluations in this thesis. First, they are used to analyse how common metrics behave under varying imbalance ratios and to assess IPIP and BAIC under controlled imbalance conditions (Chapter 5). Second, a subset of these stationary problem datasets is used to study the stability and bias of SHAP-based explanations under imbalance (Chapter 7).

Seven datasets are employed in the stationary imbalance experiments, presenting a diverse range of class imbalance ratios (IR), sample sizes, and feature dimensions. This

heterogeneity enhances the robustness assessment of the proposed evaluations. Table 4.4 shows a summary of each dataset ordered by IR, defined as the fraction of the sample size of the majority class over the sample size of the minority class.

With respect to the SMS dataset, a stationary cohort comprising two demographic features and 19 comorbidities was derived from 86,867 electronic medical records of COVID-19 patients. It is structured for binary classification (survivor vs. non-survivor) and contains 1,141 positive cases. Regarding the multiclass sets, the Forest dataset [254] originally had seven classes; only the two largest classes were maintained to convert it into a binary problem. The Satimage dataset originally contained six classes; the minority class was preserved as the target, while a synthetic majority class was created by merging all other samples.

Datasets	Instances	IR	Attributes	Categorical atts.	Min labels in cat. att.	Max labels in cat. att.	Mean labels in cat. att.
SMS	86867	75.34	22	19	2	2	2
Mammography	11183	42.10	7	0	0	0	0
Forest	38501	13.03	28	17	2	2	2
Satimage	6430	9.29	37	0	0	0	0
Adult	48842	3.18	15	8	2	42	12.75
Phoneme	5404	2.41	6	0	0	0	0
Diabetes	768	1.87	9	1	17	17	17

Table 4.4: *Datasets from stationary systems used throughout the thesis, including the number of instances, IR, attributes, and detail on categorical labels.*

4.2.2 Concept Drift Benchmark Datasets

To evaluate the experimental contributions of this thesis under non-stationary conditions, we use both real-world and synthetic data streams commonly employed in the concept drift literature. These streams allow us to assess (i) performance degradation over time, (ii) robustness to evolving class distributions, and (iii) the effectiveness of passive adaptation strategies in imbalanced drifting environments. In all stream experiments, data are processed sequentially in chunks, following the methodologies described in Chapter 6.

Eight datasets widely established as benchmarks in the drift literature are considered, encompassing both real-world and synthetic streams with diverse dimensionalities and imbalance ratios. Support for heterogeneous streams fortifies the robustness assessment. Table 4.5 summarises each dataset. Specifically, four real-world datasets are utilised (Airlines, Electricity (Elec), Weather, and an SMS subset), alongside four synthetic streams (SEA [255] and three Hyperplane streams [256] where drift is introduced across 2, 5, and 8 attributes). The public benchmarks are sourced from the Massive Online Analysis (MOA) [257] framework. For the dynamic scenarios, the SMS stream consists of two demographic features, 18 comorbidities, and 24 symptoms extracted from 278,956 electronic medical records, framed as binary classification (hospitalised vs. non-hospitalised) with 5,018 positive cases.

To ensure consistency and comparability in the stream evaluations, a uniform chunking and imbalance control procedure is applied. Each stream is horizontally partitioned into fixed-size chunks of 1,000 instances. For controlled evaluations, multiple stream variants are generated by explicitly enforcing a predefined minority-class proportion $p_{\min} \in \{0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ within each chunk. Chunks are populated by sequentially sampling instances from both classes under the mandated $p_{\min}$ constraint, terminating when the data generator can no longer satisfy the required class proportions for an entire chunk.

Datasets	Instances	Attributes	Categorical atts.	Chunks	Drift Type
Airlines	539383	8	5	306	Unknown
Elec	45312	8	1	26	Unknown
Weather	18159	9	1	12	Unknown
SMS	278956	45	42	16	Unknown
SEA	50000	4	1	64	Sudden
Hyperplane_2	10000	10	1	15	Incremental
Hyperplane_5	10000	10	1	15	Incremental
Hyperplane_8	10000	10	1	15	Incremental

Table 4.5: *Streaming datasets and generators used in the drift evaluations, detailing dimensions and chunk counts.*

4.2.3 Stream-based Datasets

One of the final experimental blocks of this thesis (Chapter 8) introduces specific datasets for the visual monitoring of feature importance under concept drift, apart from a subset of the SMS registry, two additional datasets and environments are employed. To support these evaluations, we rely on the MOA framework [258]. MOA is a widely recognised open-source software environment designed specifically for data stream mining. It provides several machine learning algorithms, evaluation metrics, and data generators for non-stationary environments, allowing to simulate different types of concept drift (e.g., sudden, gradual, or incremental) under controlled conditions.

Using this framework, the following datasets were employed:

- **Synthetic data stream (AGRAWAL):** Generated using the AGRAWAL stream generator [259] within the MOA framework. This generator simulates a realistic loan approval process, producing a dataset with nine features—six numerical (such as salary, commission, and age) and three categorical (such as educational level and car maker). The target variable indicates whether a loan should be approved based on underlying financial rules. The power of the AGRAWAL generator lies in its ability to inject concept drift by switching the underlying classification rules (among ten predefined functions) for the loan approvals. For our experiments, two variants of the stream are generated: one containing abrupt concept drift and another

with gradual drift, evaluating model reaction times and the stability of feature explanations during transitions.

- **Cold rolling mill simulator (TCM):** A real-world dataset from a 5-stand Tandem Cold Metal simulator, representing heavy steel manufacturing processes. The dataset consists of 49 continuous variables characterising various stages of processing (such as tension, thickness, and rolling speed) and simulates drifts caused by sudden changes in production capacity, mechanical wear, or shifts in product types. This dataset was utilised during a research stay at the Jagiellonian University, allowing us to validate a visual monitoring technique, that will be introduced in Chapter 8, within a realistic industrial setting where concept drift directly impacts physical manufacturing outcomes. A comprehensive description of the dataset's attributes can also be found in [260]

4.3 Data Availability and Reproducibility

In order to ensure that the ML contributions presented in this thesis are experimentally reproducible, this section details the availability and distribution for all datasets utilised throughout the experiments.

As previously discussed in Section 4.1, the SMS clinical registry provided by the Regional Health System of Murcia contains sensitive patient information. Due to strict medical confidentiality protocols and compliance with the GDPR, this dataset remains strictly private. Consequently, these data cannot be distributed or made publicly accessible.

The external benchmark datasets employed for the evaluation of learning under imbalanced conditions (detailed in Section 4.2.1) are entirely public and can be obtained from standard ML repositories. Specifically, the Adult [261], Forest (Covertype) [254], Satimage (Statlog Landsat) [262], and Diabetes datasets are publicly available and can be downloaded directly from the UCI Machine Learning Repository [263]. Furthermore, the Mammography dataset [264], and the Phoneme dataset, derived from the ROARS project detailed by Alinat [265], are widely recognised benchmarks and are accessible through platforms such as OpenML [266].

Regarding the real-world streaming datasets employed in the concept drift evaluations (Airlines, Electricity, and Weather), they were sourced from established public repositories. The Airlines dataset and the normalised version of the Elec dataset are publicly available through the MOA datasets repository [257]. The Weather dataset is the benchmark originally utilised in the study by Elwell and Polikar [267] regarding incremental learning in non-stationary environments, and it has been made publicly available by the authors.

About the synthetic datasets utilised for evaluating non-stationary environments, the SEA dataset corresponds to the version provided and utilised by Elwell and Polikar

[267] in their work on incremental learning. Similarly, the Hyperplane data streams were acquired from the benchmark published by Tsymbal et al. [268] for their study on handling concept drift, specifically utilizing a configuration with a drift magnitude transition parameter, that measures the change speed in the decision boundary, of 0.1. The AGRAWAL data stream was generated using the MOA framework. To simulate gradual and abrupt concept drifts, the generation protocol was configured to explicitly change the underlying classification rules over time. Specifically, the generation process of 34 chunks of data applied the decision boundary function *determineClass2* for chunks 1 through 12, transitioned to *determineClass3* for the subsequent chunks (13 through 23), and finally switched to *determineClass5* from chunk 24 until the end of the stream.

Finally, the TCM simulator dataset, utilised for the visual monitoring evaluations presented in Chapter 8, is publicly accessible. This dataset can be downloaded from the Zenodo repository [260], where the specific variant `tcm5_dataset_5` was employed for the experiments.

5

A Framework for Learning in Highly Imbalanced Problems

The datasets introduced in the previous chapter highlight a common issue in real applications: the event of interest is unusual. When the set of members of the minority class within the training data is very small, both training and evaluation become strongly affected by prevalence. Before moving to changing environments, this chapter studies that difficulty in a simpler setting where the data distribution is assumed to be stable and the main challenge is the imbalance itself.

This thesis proposes a generic methodological framework: the **Framework for Imbalanced Learning Machines** (**FILM**). It was designed to mitigate the bias introduced by skewed class distributions in binary classification problems while preserving the adaptability required in dynamic scenarios. Then, FILM is the structural backbone of this research, and it encompasses three interrelated components (See Figure 5.1):

1. **An evaluation method** (**BAIC**)**:** A bias-adjusted performance method that integrates multiple standard indicators to provide a stable evaluation robust to variations in the minority proportion.
2. **A static learning algorithm** (**IPIP**)**:** An ensemble-based method that constructs balanced subsets without resorting to artificial data generation, training robust models.
3. **A dynamic streaming algorithm** (**IPIP-CD**)**:** An extension of IPIP, developed later in Chapter 6, that continuously updates the ensemble to deal with concept drift in non-stationary streams.

This chapter focuses on formalising the static components of the FILM framework: the BAIC method and the IPIP algorithm. These two parts of FILM provide a baseline for the streaming extensions developed in the next chapter of the thesis.

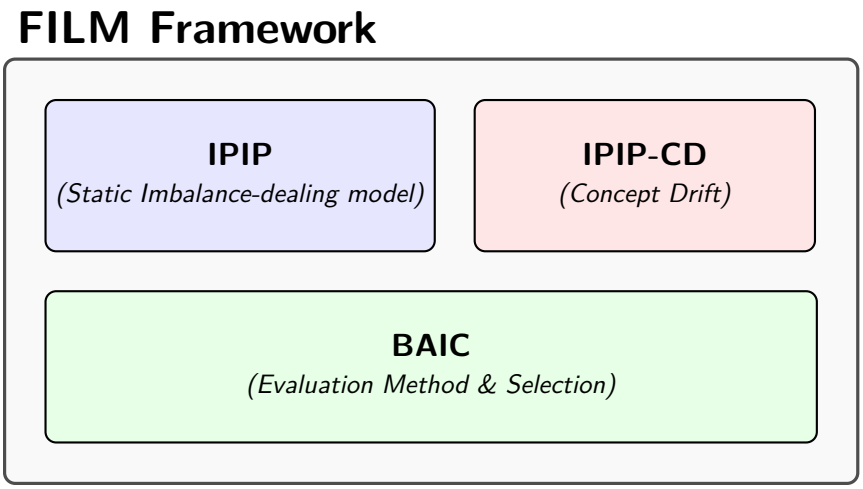


Figure 5.1: *Main modules of the FILM framework. The methodology integrates IPIP for static imbalanced learning, its dynamic extension IPIP-CD for data streams with concept drift, and the BAIC method being the model selection criterion.*

5.1 Motivation of the Framework

Machine learning is the combination of statistics, computer science, and algorithms that enable the learning of patterns from data. As introduced in Section 1.1, imbalanced data greatly influences the effectiveness of ML models. In these problems, some of the concepts we aim to learn from are underrepresented in comparison with the rest. In such cases, the resulting machine learning models tend to focus on the majority class, simultaneously forgetting the minority class.

Imbalanced problems are frequent in real-life machine learning problems. This is especially true for binary classification problems [269–274]. In such a case, the distribution of instances between the two classes is skewed, resulting in a disproportionate number of instances in one of the classes. These issues typically lead to biased results and poor performance. There are also imbalanced regression problems. In regression tasks [275, 276], the issue of imbalanced datasets arises when the distribution of the target variable is significantly skewed. And again, the resulting machine learning model will exhibit disproportionately better performance on examples with a better representation of the dependent variable. In short, it is essential to address this issue from the early stages of the data analysis tasks in any machine learning project.

One of the domains most critically affected by data imbalances is the medical domain. Examples of these include case-control cohorts of subjects with a particular disease, an intervention, or a monitoring activity. This is usually due to the fact that the most important class is usually the minority class. In medical datasets, the consequences of biased results and poor model performance are significantly more severe [277, 278]. For instance, consider the mammogram dataset example from [279], which has a distribution of 99% negative cases and 1% positive cases. A standard machine learning algorithm

applied to this dataset could potentially yield an impressive accuracy of 99%, merely by predicting all cases as negative. This approach, however, would fall short of the main goal: correctly identifying the much rarer, but crucial, positive cases while maintaining a good performance on the negative cases. These situations require special approaches to be taken. On the one hand, it is necessary to measure performance appropriately with class-imbalance-aware metrics [280, 281]. On the other hand, apart from a correct estimation of performance, we also need class-imbalance-aware data preparation techniques and machine learning algorithms [282].

One of the consequences of not using class-imbalance aware metrics is that those that are used become biased. Widely used performance metrics such as accuracy, precision, recall, area under the receiver operating characteristics (rocAUC) or F1-score can be heavily biased towards the majority class in imbalanced datasets [283]. We will demonstrate it through the chapter but we can easily see this in Fig. 5.2, in which we represent three of the most common performance measures (accuracy, F1 and rocAUC) of a conventional random forest [284] with automated grid-based hyperparameter optimisation, used on one of the datasets we will use for experimentation (SMS, see Table 4.4 and Sec. 5.4.1) in which we artificially vary the imbalance in both the training and test datasets. We observe a statistically significant association between the proportion of the minority class (from now on, p_{min}, on the x-axis) and performance values, as measured by accuracy, F1, and rocAUC, on the y-axis. The rocAUC metric even shows the opposite direction to the other two. This simple experiment highlights two distinct issues. Firstly, there is a clear association between the level of imbalance and the inflation (or depletion) of performance estimates. Secondly, apart from the bias, we observe a certain lack of agreement across metrics, as previously signalled in previous studies [285, 286]. All this eventually leads to incorrect conclusions. Taken together, these observations reveal a practical gap: machine learning engineers lack a single evaluation criterion that remains stable across variations in the proportion of the minority class while capturing the complementary information offered by conventional metrics.

In highly imbalanced settings, classic performance metrics reflect specific biases and tend to show variable inflation on the estimates due to the prevalence of one of the classes, for example F1 and Accuracy tend to inflate as the minority proportion decreases. Under such conditions, model selection becomes inconsistent across metrics and unstable under shifts in group prevalence [287]. Therefore, following recommendations from recent work [288, 289], we propose using a set of complementary performance metrics through a new approach called BAIC (Bias-Adjusted Integration Coefficient), whose aim is to provide a single, p_{min} aware technique, which integrates the basic metrics through a weighted sum. To demonstrate how effective BAIC is in reducing the bias to p_{min}, in subsequent sections, we will empirically test the possible association of the BAIC measure to p_{min}.

Examples of widely used imbalance-aware algorithms are SMOTE, ROSE, Upsample and

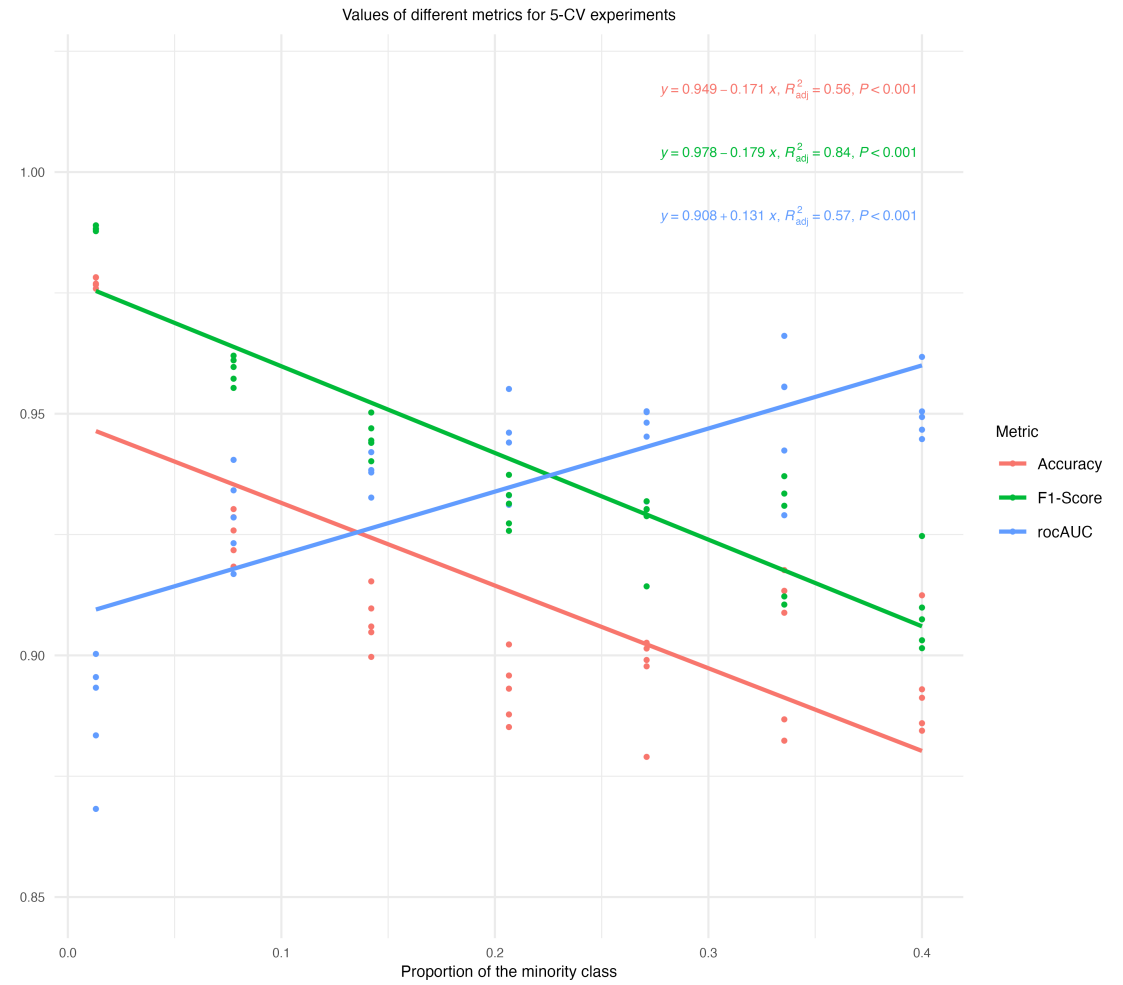


Figure 5.2: *Values of Accuracy, F1-score, and rocAUC when training a Random Forest model with upsampling on the SMS dataset, across both train and test partitions. The x-axis represents the proportion of the minority class (p_{min}), and the y-axis represents the corresponding metric value. Solid lines show the fitted linear regression for each metric, with the equation, adjusted R^2, and p-value reported in matching colour.*

Downsample. Please see Sec. 5.2 for a detailed description. Undersampling the data may discard informative majority-class instances, reducing the amount of useful signal and potentially harming model generalisation. In contrast, upsampling tends to produce overfitting and unnecessary data overlaps. To contribute to the arena of available algorithms, we include in this work a new algorithm based on resampling the original dataset [290], which addresses the aforementioned drawbacks. It is called IPIP (Identical Partitions for Imbalance Problems), and it creates ensembles of models trained on balanced subsets of the original. It is a new CPU-intensive approach but convenient in certain types of tasks (see Sec. 5.3.3).

The rest of this chapter is organised as follows. Sec. 5.2 surveys the diverse methodologies and methods currently being employed to address imbalance problems. Section 5.3 presents the datasets chosen for the study, the metrics used, along with the BAIC algorithm, our approach using the IPIP method, the design of experiments, and the evalua-

tion of the models' performance. Sec. 5.5 presents experimental results comparing our IPIP approach to other imbalance-aware approaches using all the proposed evaluation metrics and BAIC. The specific contributions of this chapter, and their relationship to the thesis objectives, are summarised together with those of the remaining chapters in Chapter 9.

5.2 Previous work

Numerous strategies have been developed to address the inherent challenges posed by imbalanced ML problems [291]. These strategies can be broadly categorised into resampling methods, cost-sensitive techniques, active learning and ensemble methods. Resampling artificially modifies the class distribution of labelled training data to increase balance by producing an artificial dataset, with the hope that machine learning algorithms will improve performance. Undersampling techniques aim to balance the class distribution by reducing the size of the majority class. This can be as simple as randomly deleting the majority class examples. While it is easy to implement, it can discard potentially valuable data. Oversampling techniques, on the other hand, increase the size of the minority class by adding more examples to it. The most straightforward approach is to randomly replicate instances of the minority class, which can lead to overfitting as it simply replicates the same data [10, 283].

A slightly more complex resampling method is SMOTE (Synthetic Minority Over-sampling Technique) [292]. It works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and creating a new sample at a point along that line. This way, the minority class is over-sampled by introducing synthetic examples along the line segments joining any or all k minority class nearest neighbours. While SMOTE is very effective, it has limitations, including the introduction of artificial noise and the oversampling of borderline examples. A SMOTE extension called Adaptive Synthetic Sampling (ADASYN) [293] is a technique that generates minority data samples based on their distributions. It focuses on generating synthetic data for complex examples, specifically those that are minority class examples, which are similar to majority class examples. This approach helps shift the decision boundary of the classifier towards challenging examples and improves the learning algorithm's performance on imbalanced datasets. ADASYN shares the same idea as SMOTE, but it focuses on generating a higher number of synthetic examples around minority class examples closer to the majority class. These examples contain, intuitively, more interesting and non-redundant information. A number of extensions of SMOTE have been proposed to overcome its limitations, including Incremental SMOTE [294], SMOTE-RkNN [295], and RSMOTE [296], each introducing different strategies to better handle class overlap, noisy instances, or sample density in the feature space. A common alternative to SMOTE is ROSE (Random Over-Sampling Examples) [297], which operates by generating arti-

ficially balanced samples using a smoothed bootstrap technique (i.e., random sampling with replacement of cases in the minority class).

Although the algorithms mentioned above are the most widely used in the literature and are the ones we compared with the new IPIP model, there are other methods for addressing class imbalance that help us provide context for IPIP and understand the type of model it is. Cost-sensitive learning employs an integral approach, where misclassification costs are embedded in the learning process, with a particular focus on accurately classifying the minority class. These costs may be calculated statistically or given by a domain expert in the form of a cost matrix. MetaCost [298] incorporates the cost matrix into the learning process itself, treating all classifiers as black boxes. Therefore, it applies virtually to any ML algorithm. The Cost-sensitive Weighted approach [299] is a method that assigns each instance a weight proportional to its misclassification cost. This is usually manually defined in a cost matrix. The weights bias the classifier towards correctly classifying instances with higher misclassification costs by combining upsampling of high-cost instances and downsampling of low-cost ones. MVCSKL[300] uses an asymmetric linear-exponential loss function in a cost-sensitive kernel learning framework to assign diverse weights to distinct samples for different classes in a dataset with incomplete information.

Active learning is a learning paradigm that enables the model to interactively acquire labels for uncertain data points from an oracle, such as a human annotator or a side selection algorithm. In imbalanced problems, active learning aims to acquire labels for the most informative instances, typically those located near the decision boundary between classes. For example, margin-based active learning with Support Vector Machines (SVM) [301] aims to identify and learn from instances which are closest to the hyperplane of separation, as their classification is more uncertain for this algorithm. Fair Active Learning using Fair Clustering, Uncertainty, and Representativeness (FAL-CUR) [302], which addresses fairness concerns in Active Learning while maintaining model accuracy by combining uncertainty and representative sampling with fair clustering. Online learning algorithms (OLA) continuously process new data, instance-by-instance and integrate it without retraining on all previous data. OLA approaches can be applied to imbalanced problems [303] through Instance Selection Strategies (ISS). ISSs are used to determine the new points to be added to the model during the active learning process. CSRDA [304] is a cost-sensitive and online algorithm that achieves a balance between low classification cost and high sparsity, and has been applied to three online anomaly detection tasks.

Ensemble techniques [305] are based on integrating basic machine learning models into meta-models. Ensembles are particularly effective in dealing with imbalance through bagging and boosting styles of ensemble construction. Bagging [306] constructs the basic models from resamples of the original training data, where repeated samples and samples left out are included. Specifically, UnderBagging combines the principles of

bagging and undersampling. It creates several balanced subsets, or bags, by undersampling the majority class. From each bag, an ML model is created. Final predictions are made by combining the individual predictions from the components of the meta-model in a specific manner. Variants include Roughly Balanced Bagging [307], based on creating resamples based on the inverse binomial distribution, and OverBagging, i.e., bagging with oversampling. SMOTEBagging [308] combines bagging resampling with SMOTE to guarantee their balance. While in bagging, the basic models are built in parallel; in boosting, they are created sequentially, as the current model focuses on modelling the errors from the previous model. This process continues until there is practically no error to model. For example, SMOTEBoost [309] combines boosting with SMOTE. We can also find bagging-boosting hybrids, such as EasyEnsemble [310]. It is a bagging algorithm that utilises Adaboost (a boosting approach) as the basic learner, implying that each boosting process can be performed in parallel. Another hybrid algorithm is BalanceCascade [310] and requires sequential model creation. This time, within each iteration, instances of the majority class that are classified with high confidence are purged from the dataset. This leads to a progressive reduction in the size of the majority class, helping to prevent overfitting and sharpening the model's focus on instances that are more challenging to classify correctly. HICD [311] is another hybrid algorithm that utilises data density to perform data sampling, generates subproblems for different classes of instances to construct ensemble models, and selects appropriate classification models based on the instance distribution.

A complete taxonomy of the imbalance learning techniques mentioned in this section can be seen in Fig. 5.3. The algorithm we propose in this work, IPIP, is a hybrid ensemble-based algorithm.

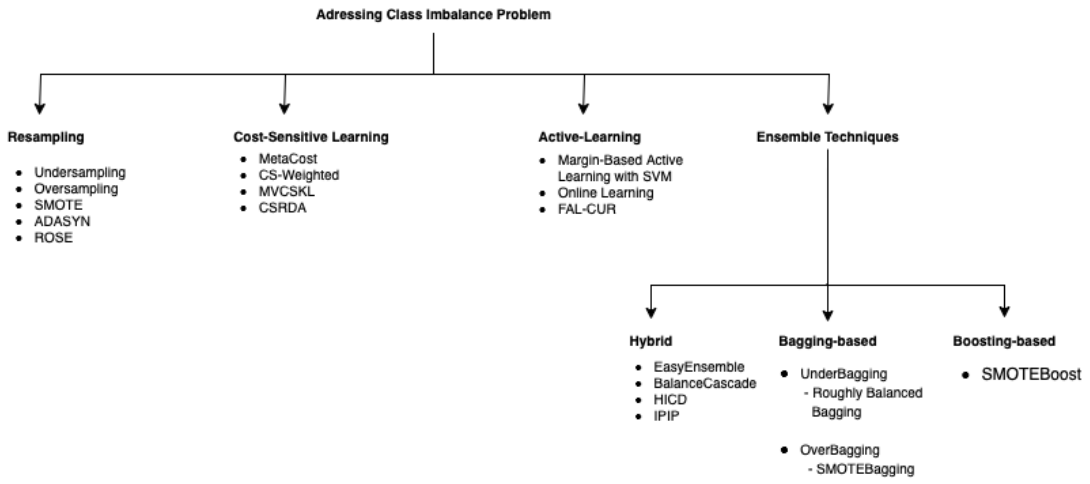


Figure 5.3: *Taxonomy of the imbalance learning techniques studied in this section.*

Based on this review, IPIP is designed as a hybrid ensemble-based approach that utilises balanced resampling and consensus learning. By opting for this specific configuration, we aim to overcome the limitations identified in prior methods, such as the introduction

of synthetic noise or the loss of significant majority class instances information, while ensuring a statistically guaranteed coverage of all available minority instances.

5.3 Materials and methods

5.3.1 Performance metrics

Henceforth, positive instances refer to the labels of the minority class samples of a binary classification problem, while negative instances shall refer to the majority class samples. We have multiple ways to inspect the performance of binary classifiers, but all are based on four elements, depending on whether the outcome from the classifier is correct or incorrect. True positives (TP) and true negatives (TN) refer to instances that are correctly classified as positive and negative, respectively. On the other hand, false positives (FP) and false negatives (FN) refer to instances that are incorrectly classified. We can compound a confusion matrix with these four values to make a fine-grained judgment on classifier performance. Moreover, we can combine those to obtain a variety of performance metrics (see Table 5.1). Accuracy is defined as the proportion of instances that are correctly classified as either positive or negative. Precision is the fraction of all relevant instances divided by the instances obtained. Recall is the fraction of relevant instances obtained over the total number of relevant instances. The F1-score is obtained through the harmonic mean of two complementary measures: precision and recall. Cohen's kappa[312] is a measure for evaluating the accuracy of a classifier that, by correcting for what you would expect by chance, takes into account the imbalance between positive and negative examples. The Matthew's correlation coefficient [313] measures the difference between the predicted values and actual values and is equivalent to a χ^2-square statistic for a 2×2 confusion matrix. The sensitivity is the proportion of correctly classified positive instances, and the specificity is the proportion of correctly classified negative instances.

Balanced Accuracy is calculated as the average of sensitivity and specificity. The Geometric mean is calculated as the geometric mean of the sensitivity and specificity. Two well-known metrics are the Area Under the Receiver Operating Characteristic Curve (rocAUC) and the Area Under the Precision-Recall Curve (prAUC), where the Receiver Operating Characteristic Curve represents the true positive rate (sensitivity) as a function of the false positive rate for different probability values. The precision-recall curve illustrates the relationship between precision and recall. Although the Imbalance Ratio (IR) is the most primarily used metric for measuring dataset imbalance, we will mainly use the proportion of the minority class, p_{min}, going forward. This is because when we work with it, we have the restriction that IR is not unbounded, while $p_{min} \in (0, 0.5)$, which makes it easier to interpret. The relationship between IR and p_{min} in a dataset can be expressed as follows:

$$p_{min} = \frac{1}{IR + 1}. \tag{5.1}$$

Of all the metrics mentioned above, we will work with Accuracy, Cohen's Kappa, Balanced Accuracy, F1-score, rocAUC, prAUC, Matthew's Correlation Coefficient and Geometric Mean. We will observe that all these metrics exhibit variable biases towards p_{min}.

Metric	Equation
Accuracy (Acc)	$\frac{TP+TN}{TP+FP+TN+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision (precision)	$\frac{TP}{TP+FP}$
Recall ($recall$)	$\frac{TP}{TP+FN}$
False positive rate (FPR)	$\frac{FP}{FP+TN}$
F1-score (F_1)	$2 \cdot \frac{precision \cdot recall}{precision+recall}$
Cohen's Kappa (Kappa)	$\frac{2 \cdot (TP \cdot TN - FN \cdot FP)}{(TP+FP) \cdot (FP+TN) + (TP+FN) \cdot (FN+TN)}$
Matthews correlation coefficient (MCC)	$\frac{RN \cdot TP - FN \cdot FP}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$
Balanced Accuracy (Bal_Acc)	$\frac{sensitivity+specificity}{2}$
Geometric mean (GMean)	$\sqrt{sensitivity \cdot specificity}$

Table 5.1: *Equations of evaluation metrics for binary classification.*

5.3.2 The Bias-Adjusted Integration Coefficient (BAIC) method

In most data analysis problems that involve ML, there is a step in the analysis pipeline devoted to the algorithm's hyperparameter optimisation and model selection. On the one hand, we need to determine the optimal values for the algorithm hyperparameters based on the target dataset. For this purpose, we typically create a set of candidate models by exploring the hyperparameter space [314, 315] to select the hyperparameter values used in the best model created during that exploration. On the other hand, we have to trust the final model we will use in exploitation mode. In both of these processes, it is crucial to use an appropriate statistic to measure performance. This issue is even more critical when classes are imbalanced because, as we have seen, the possible statistics may be biased towards the proportion of the minority class, p_{min}. Our main thesis is that we can create a new measure, as an integration of all the biased ones, in an additive fashion in which all basic measures appear corrected by their respective correlation with p_{min}, with the intent that the new statistic will be less biased to p_{min}. To do this, we need an estimate of this correlation. To obtain this estimate, we proceed as follows. Consider a

set of k performance metrics indexed by i $(1 \leq i \leq k)$. For instance, the metrics listed in Table 5.1 may be used. We want to create a joint metric as uncorrelated as possible with p_{min}. Let us denote the original minority class proportion in a specific dataset d as p_d. We then create n datasets, where n is an even number and $n \geq 6$, with artificial class proportions by randomly sampling the entire original dataset. To ensure that a broad range of minority class proportions is represented in the n datasets, we distinguish between two cases based on the original p_d. If $p_d > 0.05$, then $n/2$ of those datasets have variable, and equally distributed, class proportions within the interval $[0.05, p_d]$ and the other $n/2$ datasets have variable and equally distributed class proportions within the interval $[p_d, 0.4]$. Otherwise, if $p_d \leq 0.05$, n datasets are created in the same way within the interval $[p_d, 0.4]$, p_d values greater than 0.4 are not considered proper for imbalanced datasets and, therefore, it would make no sense to use BAIC with them.

Given an original dataset d and its n artificially sampled versions, we split each of them into training and test sets and then train ML models on all of them, including d, obtaining $n+1$ models. We evaluate each model on its respective test set using the k different metrics. Let us denote the vector of the i-th metric for a given dataset d and its n sampling subsets with $\mathbf{m_{d,i}} = (m_{d,i}, m_{1,i}, m_{2,i}, ..., m_{n,i})$ and the vector of of minority class proportions with $\mathbf{p_{min}} = (p_d, p_1, p_2, ..., p_n)$. To obtain an estimate of the bias of the k-th metric in a particular dataset d, we can calculate Pearson's Correlation, $r_{d,i}$, between $\mathbf{m_{d,i}}$ and $\mathbf{p_{min}}$.

Thus, each metric is assigned a weight according to its correlation with $\mathbf{p_{min}}$, so that the closer $r_{d,i}$ is to 0, the higher its weight should be. Therefore, to assign a weight to a given metric, we need a function that, when $r_{d,i}$ is close to 0 returns a weight close to 1, and, otherwise, when $r_{d,i}$ is close to ± 1, returns a weight close to 0. Such a function can be modelled through a Gaussian, i.e., an exponential that takes values from a concave quadratic function with parameters (a, b, c) (Eq. 5.2), being a the curve max height, b its centre where the function reaches its maximum value, and c its width:

$$G(x) = a \cdot \exp\left(-\frac{(x-b)^2}{2c^2}\right) \tag{5.2}$$

In Fig. 5.4, three different Gaussian functions are shown with different values for c, and with a maximum value $a = 1$ and centred at 0, $c = 0$. Pearson correlation values of x close to 0 lead to metric weights close to 1. Extreme Pearson values for x, i.e., ± 1, lead to weights close to 0. Note that values of c close to 0 lead to a more substantial penalty for p_{min} bias near absolute values of 1. For example, in the figure, the curve for $c = 0.15$ generates weight values closer to 0 than those we get with $c = 0.35$ (see results). In Table 5.4, several values of BAIC are 0, which means the metric is strongly biased towards p_{min}.

Therefore, given an imbalanced dataset d and a set of correlations, $\mathbf{r_d} = \{r_{d,1}, \ldots, r_{d,k}\}$, between the vector measures $\mathbf{m_{d,i}}$ for all k metrics and the vector of minority class proportions $\mathbf{p_{min}}$, and a set of weights for each measure obtained by applying the function

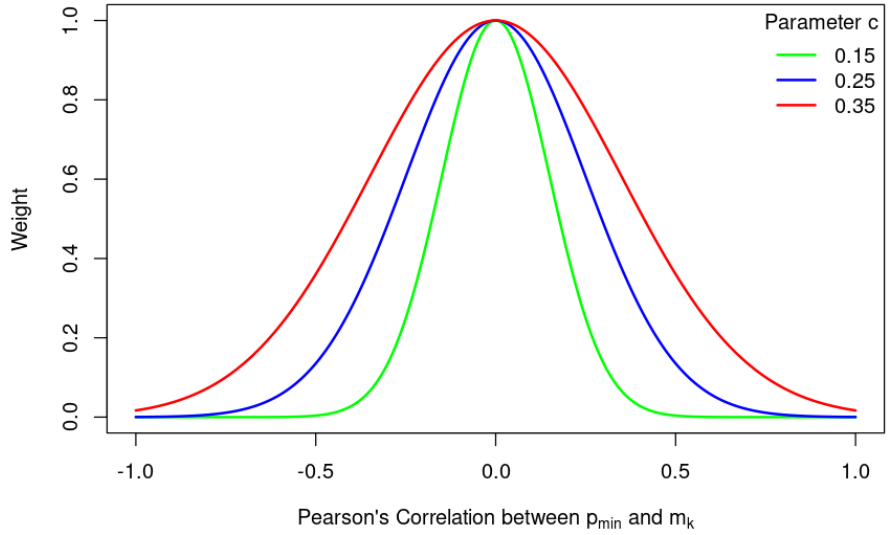


Figure 5.4: *Three different Gaussian functions. Red curve with parameters* $(1, 0, 0.35)$*, blue curve with parameters* $(1, 0, 0.25)$ *and green curve with parameters* $(1, 0, 0.15)$.

$G(x)$ to the set of correlations, that is, $\mathbf{w_d} = \{w_1 = G(r_{d,1}), \ldots, w_k = G(r_{d,k})\}$, the BAIC metric can be defined as the weighted sum of each metric (Eq. 5.3).

$$BAIC = \sum_{j=1}^{k} w_j \cdot m_{d,j}. \tag{5.3}$$

As BAIC can be obtained for any ML algorithm, it is useful for finding the best technique for a given problem. It aggregates metrics through a weighted sum. And weights are data-driven, i.e., they are derived from the correlation of the metric and p_{min}. Basic metrics that are more sensitive to prevalence receive smaller weights, yielding an estimate that is less biased to imbalance while retaining each metric's complementary information. BAIC can be used as well for model selection because it is a convex combination of normalised metrics with bias-adjusted weights. Note that it does not have an upper bound and, therefore, it is not intended for cross-dataset comparison. Its workflow is illustrated in Fig. 5.5 and can be easily applied with the FILM (Framework for Imbalanced Learning Machines) R Package. The corresponding implementation is available in `https://github.com/antoniogt/FILM`.

5.3.3 The IPIP algorithm

As reviewed in Section 5.2, the strategies to address class imbalance range from data-level resampling to algorithmic adjustments such as cost-sensitive learning. Among these, ensemble methods offer a valuable solution for complex binary classification problems by enabling multiple models to be trained from resampled training data. When constructing an ensemble for tackling imbalanced problems, it is advisable to (1) avoid generating

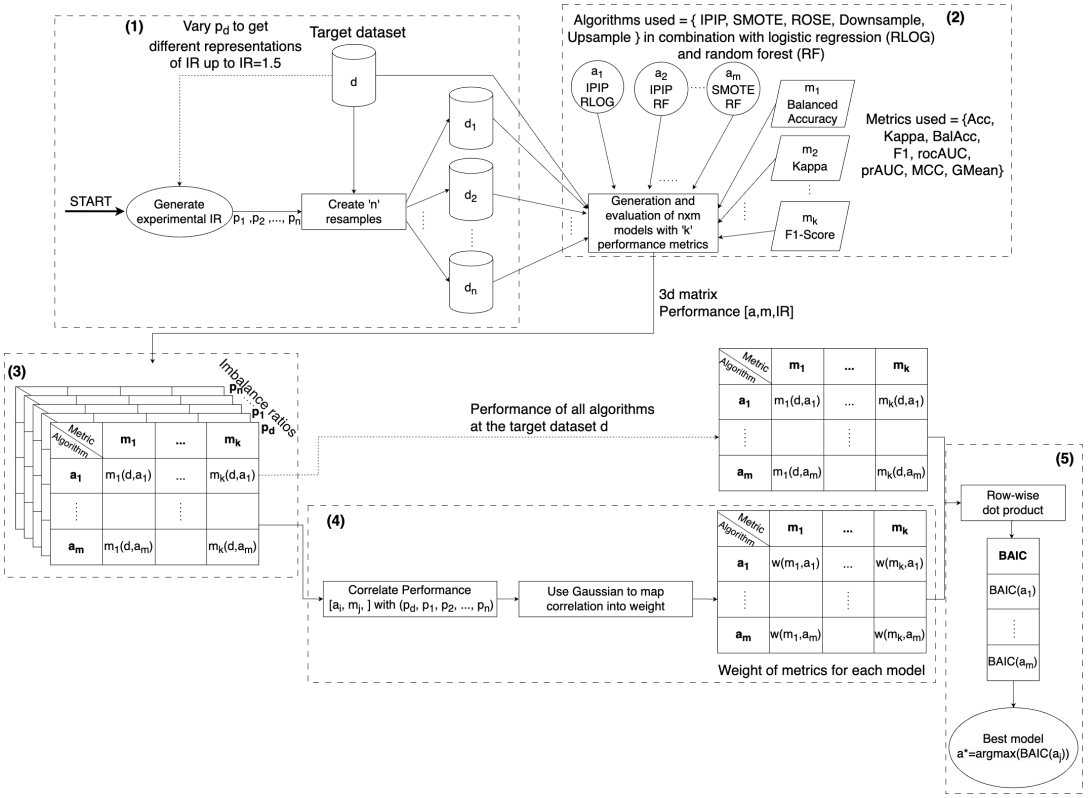


Figure 5.5: *FILM methodology for obtaining BAIC metric and choose the best model according to BAIC. In* (1)*, n datasets are obtained from the original dataset d through sampling with several imbalance ratios. In* (2)*, a set of algorithms is then trained on d and its n resamples, resulting in k metrics for each algorithm on each dataset. In* (3)*, a three-dimensional matrix is formed to retain these values. In* (4)*, we aggregate the metric results for each algorithm across all n+1 datasets, retaining the k metric values of all algorithms in d. This provides weights for each metric and algorithm. In* (5)*, the weights and values of the metrics are finally aggregated over the original dataset to obtain the BAIC metric. Then it determines which algorithm has obtained the maximum BAIC value.*

synthetic samples and (2) employ base learners as black boxes, thereby facilitating the use of any base learner on balanced resamples. To address the issue of imbalanced data, we present a novel technique termed IPIP (Identical Partitions for Imbalance Problems). IPIP generates multiple resamples by downsampling the majority class and sampling the minority class with replacement, resulting in a balanced dataset. Subsequently, each balanced resample has the potential to generate a new model, which is integrated into an ensemble through majority voting. The algorithm ensures that all examples from the minority class are selected at least once. IPIP's task is to balance the dataset and build an ensemble classifier.

The algorithm for IPIP is outlined in Algorithm 1. Firstly, the initial dataset is divided into training, validation and test sets, each containing the same class proportion as the original dataset, being the training, validation and test subsets composed of the p_{train}, p_{val} and p_{test} proportions of the total number of instances respectively (Algorithm 1 line 2). IPIP divides the training data into b_s balanced subsets by resampling, ensuring representation of all samples from the minority class throughout the resamples with

$p_{min} \approx 0.45$ of each subset (see Algorithm 1 line 3). Each of these subsets contains non-repeated and non-synthetic samples from the original dataset and is used to develop one or more models. These models are then combined into a majority voting ensemble for classification.

IPIP ensures, with confidence level $(1 - \alpha) \times 100\%$, that every single sample in the minority class is present in at least one of the b_s balanced subsets during data preparation. To achieve this, the value for b_s ought to be calculated in the following manner. Let n_{min} represent the minimum number of samples from the minority class required for the b_s resamples. We set the number of samples in the majority class, n_{max}, so that $n_{max} = n_{min} + n'$. Here, n' denotes the quantity of instances exceeding n_{min} that we need from the majority class to attain a value of $p_{min} \approx 0.45$ in every resample. Bear in mind that as n_{min} in each resample approaches n_{min} in the original dataset, b_s decreases. During our experiments, we established n_{min} to reflect 75% of the minority class samples in the original dataset. More specifically, suppose that n denotes the number of samples in the minority class of the original dataset. To obtain a set of n_{min} samples, where $n_{min} < n$, from the original minority class, we randomly select samples with replacement uniformly so that, in each iteration, each sample has a probability of $1/n$ of being chosen. In a replacement sampling, iterations are independent, therefore the random variable X_k indicating the number of times any given sample is selected for the k-th resample, with $k \in \{1, ..., b_s\}$, follows a binomial distribution with parameters $(n_{min}, 1/n)$ (Eq. 5.4)

$$P[X_k = r] = \binom{n_{min}}{r} n^r (1 - 1/n)^{n_{min} - r} \tag{5.4}$$

Since we want each sample in the minority class to be selected at least once, we have:

$$P[X_k > 0] = 1 - (1 - 1/n)^{n_{min}} = q \tag{5.5}$$

We can define another random variable, Y_k, as follows:

$$Y_k = \begin{cases} 1 & \text{if the sample is chosen for the k-th subset} \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

In this scenario, Y_k follows a Bernoulli distribution with parameter q. As each Y_k is independent, the total number of subsets in which a specific sample appears, Y, can be computed as $Y = Y_1 + ... + Y_{b_s}$. This follows a binomial distribution of parameters (b_s, q). Hence, the probability of an instance being chosen to be included in one of the b_s subsets is:

$$P[Y > 0] = 1 - (1 - q)^{b_s} \tag{5.7}$$

We want this probability to be greater than α, therefore b_s must satisfy:

$$b_s > \frac{log(1-\alpha)}{n_{min} \cdot log(1-1/n)} \tag{5.8}$$

Now, let us define a function $f_\alpha^n : N \rightarrow R$ such that

$$f_\alpha^n(n_{min}) = \frac{log(1-\alpha)}{n_{min} \cdot log(1-1/n)} \tag{5.9}$$

This function is lower bounded by the number of balanced subsets b_s and determines the number of balanced subsets required to satisfy that, with probability greater than α, every sample of the minority class appears, at least once, in one of the b_s subsamples.

For each balanced subset (algorithm 1 line 4), the algorithm builds an ensemble that progressively accumulates up to b_E ML models (algorithm 1 line 8). To achieve this, we have defined the maximum possible number of attempts to expand the ensemble. This number is given by the function $m_t(|T|)$, which depends on the number of base learners T already ensembled for a given resampling and must be a monotonically decreasing function (algorithm 1 line 8). This is because the objective is to make fewer attempts to augment the ensemble as its size increases. The rationale behind this is that it should then be more difficult to add a model that enhances the ensemble than when the ensemble is small. In our experiments, we used the function described in Eq. 5.10, but any discrete decreasing function of $|T|$ could be used.

$$m_t(|T|) = \left\lceil \frac{b_E - |T|}{3} \right\rceil \tag{5.10}$$

IPIP creates diverse resamples of each balanced subset to enable models to target different subproblems (Algorithm 1, line 9). To this end, IPIP subsamples with a $p_{min} \approx 0.5$ in each resample and n_{min} instances from the minority class. And as previously demonstrated, we only need a maximum of b_E models. Similar to the process of obtaining the lower bound for b_s (as shown in Eq. 5.8), b_E must also fulfil Eq. 5.11.

$$b_E > \frac{log(1-\alpha)}{n_{min} \cdot log(1-1/n_{min})}. \tag{5.11}$$

New base learners (line 13 of Algorithm 1) are incorporated into the ensemble only if they improve the overall performance on the validation set (line 12 of Algorithm 1), which has a class distribution equal to that of the original dataset. The ensemble generated with each balanced subset assigns an instance to the majority class if at least 75% of the models in the ensemble classify it as such. Otherwise, it is classified as belonging to the minority class (algorithm 1 line 11). The final model is formed by combining the

ensembles created from each balanced subset (line 20 of Algorithm 1).

Finally, new instances are classified by individually evaluating each component of the ensemble built for each balanced subset. The final model classifies a sample as negative if 50% of the ensembles classify it as negative (majority class). IPIP builds balanced, non-synthetic resamples that, overall and with high probability, include all minority instances at least once. These resamples are used to greedily grow ensembles with only those basic models that improve validation on the original class distribution. The IPIP algorithm is implemented in the FILM R package[1].

In terms of computational complexity, IPIP preserves the complexity of the underlying learner while introducing an overhead that depends only on the balancing parameters. In the worst case, all the $m_t(r) = \lceil (b_E - r)/3 \rceil$ attempts are used before the r-th model is added to the ensemble, the number of models tentatively trained for each balanced subset can be obtained as follows:

$$\sum_{r=0}^{b_E-1} m_t(r) = \sum_{s=1}^{b_E} \left\lceil \frac{s}{3} \right\rceil = \frac{b_E(b_E+3)}{6}, \tag{5.12}$$

and then it can be interpreted with a computational order of $\Theta(b_E^2)$. Consequently, the overall complexity of the model is bounded by

$$\Theta_{IPIP}^{worst} = \Theta(b_s b_E^2 C_{train}(S)), \tag{5.13}$$

where b_s is the number of balanced subsets, b_E the maximum ensemble size per subset, S the subset size, and $C_{train}(S)$ the cost of training a single base learner on S samples. In our experiments, the parameters b_s and b_E are always set to 7 and 5, respectively.

[1] The corresponding implementation is available in `https://github.com/antoniogt/FILM`.

Algorithm 1 IPIP($X, p_{min}, b_s, b_E, ML, tries_{max}, p_{train}, p_{val}, p_{test}$)

Input: Dataset X; p_{min}; b_s (number of subsamples, min given by Eq. 5.8); b_E (max ensemble size); ML (ML technique); $tries_{max}(\cdot)$ (max attempts function); split proportions $p_{train}, p_{val}, p_{test}$ with $p_{train} + p_{val} + p_{test} = 1$

Output: M: ensemble of ML models

$M \leftarrow \varnothing$
$(D^{train}, D^{val}, D^{test}) \leftarrow$ stratified_split$(X, [p_{train}, p_{val}, p_{test}])$
$\{D_1^{train}, D_2^{train}, \ldots, D_{b_s}^{train}\} \leftarrow$ balanced_subsampling(D^{train}, p_{min})
for all $D \in \{D_1^{train}, D_2^{train}, \ldots, D_{b_s}^{train}\}$ **do**
  $M_{bs} \leftarrow \varnothing$
  $tries \leftarrow 0$
  $best_performance \leftarrow 0$
  while $|M_{bs}| < b_E$ **and** $tries < tries_{max}(|M_{bs}|)$ **do**
    $D' \leftarrow$ balance_subsampling$(D, 0.5)$
    $model \leftarrow$ train(D', ML)
    $performance \leftarrow$ evaluate$(M_{bs} \cup \{model\}, D^{val})$
    if $performance > best_performance$ **then**
      $M_{bs} \leftarrow M_{bs} \cup \{model\}$
      $best_performance \leftarrow performance$
      $tries \leftarrow 0$
    else
      $tries \leftarrow tries + 1$
    end if
  end while
  $M \leftarrow M \cup \{M_{bs}\}$
end for
return M

5.4 Experimental setup

This section details the experimental framework designed to evaluate the proposed BAIC metric and the IPIP algorithm. First, we briefly reference the datasets selected for the evaluation. Next, we outline the implementation details, including the data partitioning strategy, the base learners applied, and the established imbalance-aware approaches used as baselines for comparison. Finally, we introduce the methodological tools created to analyse the results, specifically the novel concordance plots for assessing metric agreement and the correlation analysis designed to quantify the bias of standard metrics towards the minority class proportion.

5.4.1 Datasets

The experimental evaluation is carried out using seven datasets, covering a wide range of class imbalance ratios, sample sizes, and feature dimensions. A detailed description of these static datasets, including their preprocessing, origins, and structural characteristics, has been extensively presented in Chapter 4 (Section 4.2.1).

5.4.2 Implementation details

Each dataset is then divided into two subsets: 80% is allocated for training the model, and 20% for testing it. The training dataset is then utilised to train models using a 5-fold cross-validation method. The simplest non-overfitting logistic regression and random forests were selected as base learners in all experiments implemented. The model-building process has been implemented using the R package *caret* [316] (version 6.0-94), and *glm* for logistic regression (R version 4.3.1) and *ranger* for Random Forest (version 0.15.1). Two distinctly different modelling approaches were tested. To evaluate the effectiveness of IPIP, we compared it with well-known Imbalance-Aware Approaches (IAA), such as SMOTE, Rose, Upsample, and Downsample. Additionally, we incorporated two ensemble-based hybrid techniques, SmoteBoost and UnderBagging, which share similarities with IPIP's approach. All experimentation utilised a five-fold cross-validation strategy. For the Random Forest model, a hyperparameter search strategy has been conducted for each dataset. During the hyperparameter search, Cohen's Kappa was used as a performance metric as it presents a better trade-off between simplicity and p_{min}. The hyperparameters used for each dataset are described in Table 5.2.

	mtry	**min.node.size**	**splitrule**
SMS	{1, 4, 7, 10, 13, 16, 19}	{1, 11, 21}	gini
Forest	{1, 2, 3, 4, 5}	{3, 13, 23}	gini
Phoneme	{1, 4, 7, 10, 13, 16, 19}	{1, 11, 21}	gini
Mammography	{1, 3, 5, 7}	{1, 11, 21}	gini
Satimage	{1, 3, 5, 7, 9, 11, 13}	{1, 11, 21}	gini
Diabetes	{1, 2, 3, 4, 5}	{1, 11, 21}	gini
Adult	{1, 3, 5, 7}	{1, 11, 21}	gini

Table 5.2: *Values given to create the grid of hyperparameters for random forest models in all experiments for each dataset.*

5.4.2.1 Concordance plots between metrics: Concordance plots

The first question that needs to be addressed is whether all the metrics evaluated in our experiments are consistent in identifying a particular IAA technique as the one that achieves the highest value in all metrics for a given dataset. This was accomplished by applying all IAA algorithms to all datasets, building logistic regression and random forest models and collecting all metrics using a 5-fold cross-validation procedure, varying

the proportion of the minority class (see Sec. 5.3.2). To ensure results can be analysed in a visually interpretable manner, we propose a new type of plot called concordance plot, which represents the proportion of agreement between metrics for a given IAA technique and for two models on a given dataset.

Given a dataset, d, we apply a set of IAA techniques $A = \{a_1, a_2, ..., a_n\}$ and evaluate them with a set of performance metrics $M = \{m_1, m_2, ..., m_m\}$ obtained through a set of experiments $E = \{e_1, e_2, ..., e_t\}$. Each experiment consists of building a machine learning model with hyper-parameter optimisation based on grid search for each of the different datasets sampled from d, with different minority class proportions, $\mathbf{p_{min}}$, using 5-fold cross-validation resampling. In this context, we can establish that there is an agreement or concordance between the metrics m_i, m_j, $i \neq j$, $1 \leq i, j \leq m$ for a specific experiment $e \in E$ if both metrics give the highest value to the same technique $a_k \in A$ on e. Otherwise, we consider that there is a disagreement or discordance. To construct a concordance plot, we need to compute the agreement ratios for each technique a_k using a particular metric m_i across all experiments $e_t \in E$. This can be achieved by calculating the number of times a_k technique scores better for m_i metric in each experiment e and dividing it by the total number of experiments, $|E| = |\mathbf{p_{min}}| \times 5$. Agreement ratios can be visually represented via pie charts for each metric pair m_i, m_j, where each slice corresponds to technique $a_k \in A$. A fully coloured pie denotes agreement between m_i and m_j, whereas an empty or nearly empty pie indicates strong disagreement. The larger slice indicates the best-performing algorithm on the dataset d for cases where m_i and m_j agree. To produce a concordance plot, we collect all the slices in an upper triangular matrix, where each pie plot of the main diagonal represents a complete agreement, as it compares a metric with itself, and it can be interpreted as the proportion in which a model shows the highest value for that metric across all experiments. Each non-diagonal pie plot, which corresponds to the (i, j) index with $1 \leq i, j \leq m$, shows the agreement/disagreement when choosing the best model between the metrics m_i and m_j across all experiments.

To provide a clear conceptualisation of this graphical tool, Fig. 5.6 illustrates two extreme scenarios, evaluating three hypothetical algorithms (A1, A2, and A3) across eight metrics. In plot A), two distinct clusters emerge. The first cluster reveals a perfect, unanimous agreement among Accuracy, Kappa, Balanced Accuracy, and F1-score, with all respective pie charts entirely filled by the color corresponding to algorithm A1. This establishes that A1 achieved the highest performance across all experiments according to these specific metrics. Conversely, the second cluster demonstrates perfect agreement among rocAUC, prAUC, MCC, and GMean, which exclusively designate algorithm A2 as the optimal approach in every instance. Moreover, the intersection between these two metric clusters is entirely blank, representing a 100% discordance rate. This visually shows that an algorithm may appear perfect if evaluated under metrics like Accuracy, yet simultaneously fail to obtain a high value in metrics like MCC or G-Mean. Finally, algorithm A3 is never identified as the highest-performing model by any metric, resulting in its complete absence from the visualisation. To contrast this extreme, Panel B depicts

a scenario of total uncertainty where no clear decision can be made. In this case, performance is completely random across all the results, resulting in colour-divided pie charts in the main diagonal, showing that nor an algorithm shows the best performance for all experiments. There is also a dominant proportion of "No Match" (empty sections) in the off-diagonal pie plots, showing that there is no concordance between metrics when choosing the best algorithm. This visualisation for analysing the concordance between metrics is included in the FILM package.

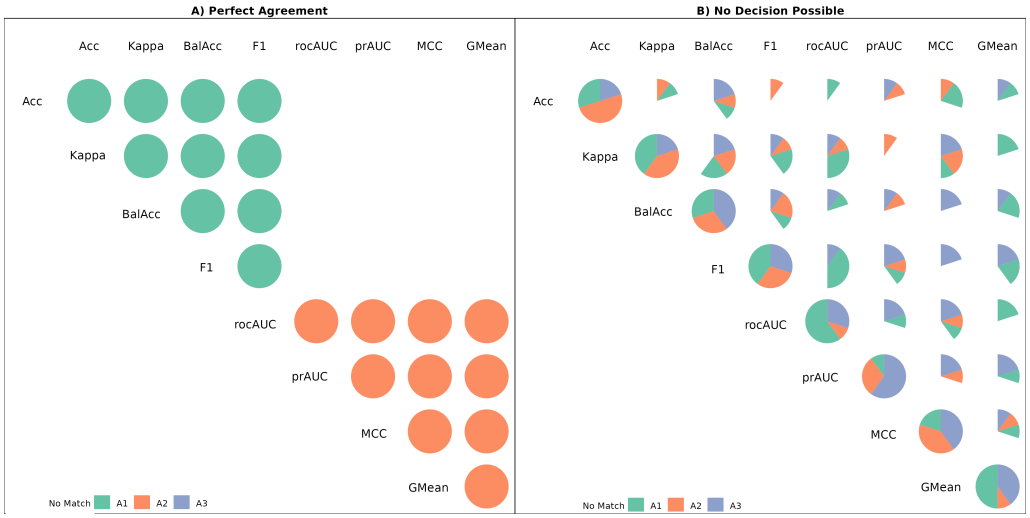


Figure 5.6: *Synthetic examples of concordance plots illustrating two extreme evaluation scenarios. A) A scenario of perfect agreement forming two distinct metric clusters that show algorithms A1 and A2 as best models, respectively, with absolute discordance between them. B) A scenario of total lack of consensus, characterised by a high proportion of empty sections in the off-diagonal pie plots.*

5.4.2.2 Correlation between metrics and proportion of the minority class $p_{\mathbf{min}}$

A key question in this chapter is whether the most commonly used metrics for evaluating binary classification models are biased by the proportion of the minority class p_{min} in a given dataset d and ML model. To shed light on this fundamental question, we investigate the correlation between the set of metrics $\{\mathbf{m_{d,i}}\}_{\mathbf{i=1,\ldots,k}}$ acquired by evaluating the machine learning model utilising a specific IAA method and the set of variations in minority class proportions, $\mathbf{p_{min}}$, employed to produce diverse versions of the initial dataset d. This procedure is explained in more detail in Sec. 5.3.2. The BAIC metric is also added to this analysis in order to be able to compare it with the other metrics.

5.4.2.3 Comparison of metrics biases towards the proportion of minority class $p_{\mathbf{min}}$

In this chapter, we introduce a new performance metric called BAIC (see Sec. 5.3), which exhibits less bias towards minority class proportions compared to other metrics employed. To empirically verify this, we incorporate the correlation between BAIC and $\mathbf{p_{min}}$ in a correlation vector $\mathbf{r_{d,BAIC}}$, for each dataset d. If the resulting vector $\mathbf{r_{d,BAIC}}$ equals

the zero vector, that will be referred to as the unbiased correlation vector; it implies that the BAIC metric is unbiased towards p_{min}, at least for the dataset d. We repeat this with all k metrics. Therefore, a measure of non-bias can be determined by calculating the Euclidean distance of vector $\mathbf{r_{d,i}} =< r_{d,i}, r_{d_1,i}, \ldots, r_{d_n,i} >$ from the unbiased correlation vector. To compare the bias of BAIC with other metrics, we visually evaluate the distribution of values for the Pearson's correlation coefficient between metrics and p_{min} across various datasets using boxplots. Additionally, pairwise statistically significant differences are assessed between the correlation values of BAIC and the other metrics using a non-parametric pairwise Wilcoxon Hypothesis Test, with the Bonferroni correction applied.

5.5 Results

5.5.1 No agreement among metrics in deciding on the best model

We have produced concordance plots for the two basic learning methods, namely logistic regression and random forest, for all seven datasets examined (see Sec. 5.4.1). For instance, Fig. 5.7 displays the pairwise agreements (see Sec. 5.4.2.1) for the Diabetes and Mammography datasets (Figures 5.8-5.13 contain plots of the remaining datasets). Pies on the main diagonal of Fig. 5.7-A correspond to a given performance measure while displaying the proportion of wins for each IAA technique under that metric. Pie charts in cells (i, j), where $i \neq j$ can be used to compare metrics i and j. The filled portion of the chart indicates agreement, while the empty section indicates disagreement.

For instance, when we examine the main diagonal of Fig. 5.7-A, for the Diabetes dataset, it is evident that, in terms of Accuracy, Kappa, and F1-score, both SMOTE and Smoteboost exhibit the highest performances based on the respective pie on the main diagonal in similar proportions. However, Balanced Accuracy indicates that IPIP, SMOTE, and Upsample are the best models, with no model proving to be predominant. Across the remaining metrics, equivalent proportions are attained by all models. Consequently, it is challenging to determine which model is the best one for these metrics. When examining the non-diagonal pie plots, there is virtually no agreement between F1 and several metrics, including Balanced Accuracy, rocAUC, prAUC, MCC, and Geometric Mean. This is because the pie plots comparing F1 with these metrics are mostly empty, as each pie plot has more than half of its area unfilled. At the same time, the Accuracy and F1 measures seem to agree in most cases, with SMOTE and SMOTEboost being the models that obtained the highest metric values. In Fig. 5.7-B, for the Mammography dataset, we observe a similar situation. By examining the main diagonal, it is evident that IPIP outperforms the other methods in terms of Accuracy, Kappa, and MCC.

Meanwhile, we can observe that Downsample and Underbagging obtain the highest proportion of the pie chart for the prAUC metric. However, when examining the non-diagonal pie plots, particularly those showing the agreement between the prAUC and the other metrics, a notable lack of agreement is evident when selecting the best model, with most pie plots being mostly empty. This observation also applies to the rocAUC metric, which agrees with no other metric in more than 50% of the cases when determining the best-performing model.

From the concordance plot, we can define a measure of discordance between two measures as the empty portion of a pie. In our experiments, we obtain a global discordance ratio across all measures of 0.48, 99% CI $[0.37, 0.59]$ in panel A, and 0.52, 99% CI $[0.36, 0.68]$ in Fig. 5.7-B. Both these numbers and the plots in Fig. 5.7 clearly reflect that we cannot rely on these metrics to select the best imbalance-aware ML algorithm for a given dataset.

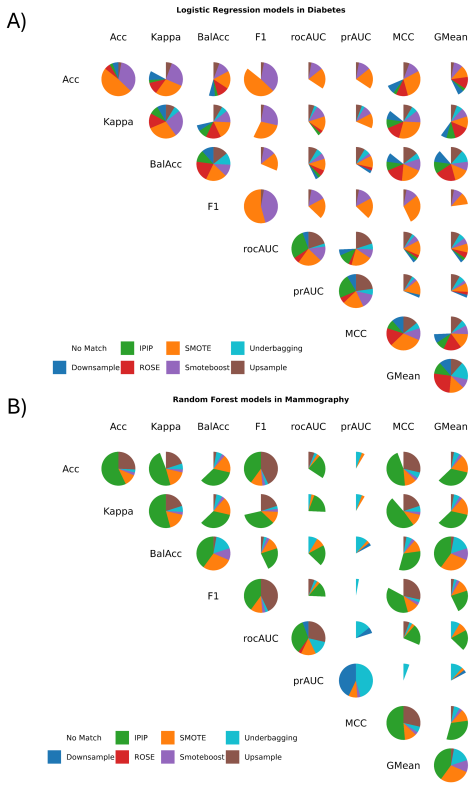


Figure 5.7: *A) Concordance plot of the Diabetes dataset using Logistic Regression. B) Concordance plot of the Mammography dataset using Random Forest models.*

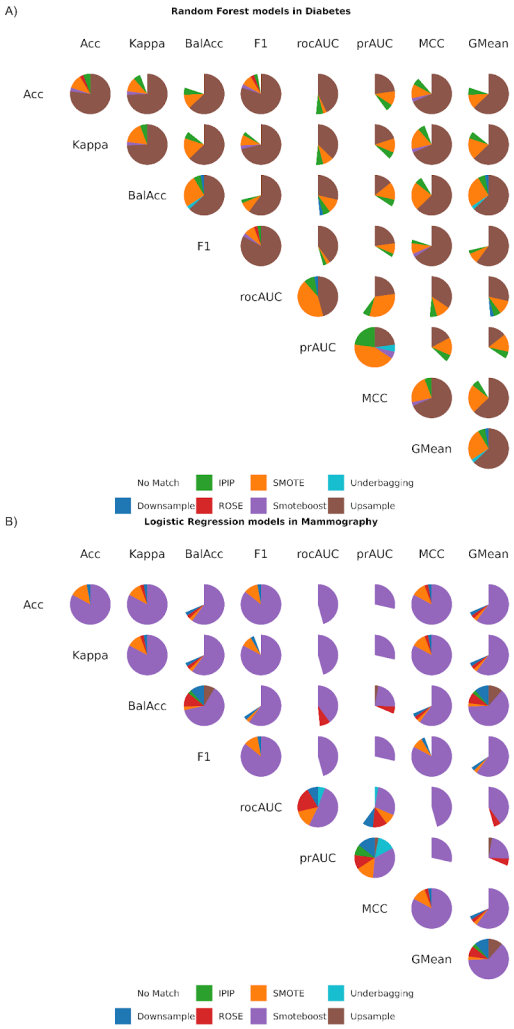


Figure 5.8: *A) Concordance plot of the Diabetes dataset using Random Forest. B) Concordance plot of the Mammography dataset using Logistic Regression models.*

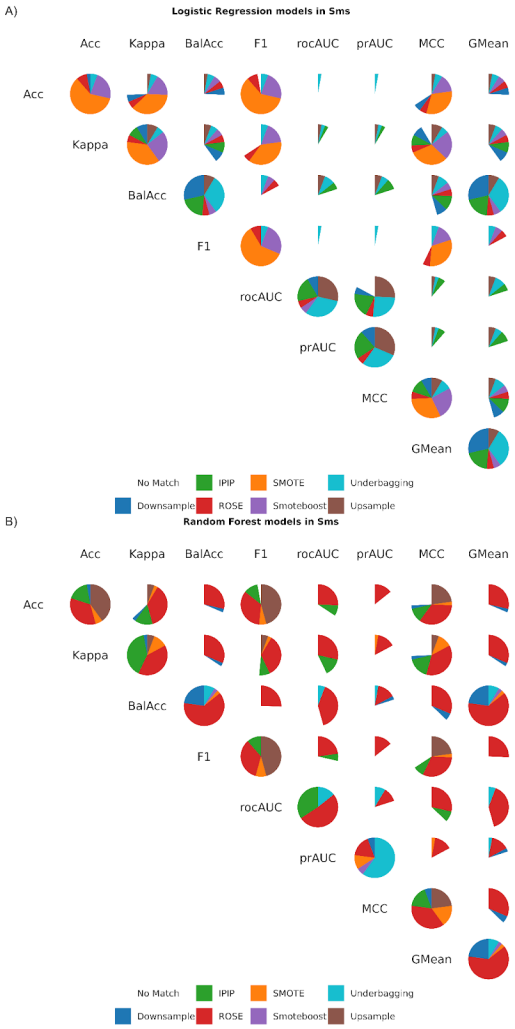


Figure 5.9: *A) Concordance plot of the SMS dataset using Logistic Regression. B) Concordance plot of the SMS dataset using Random Forest models.*

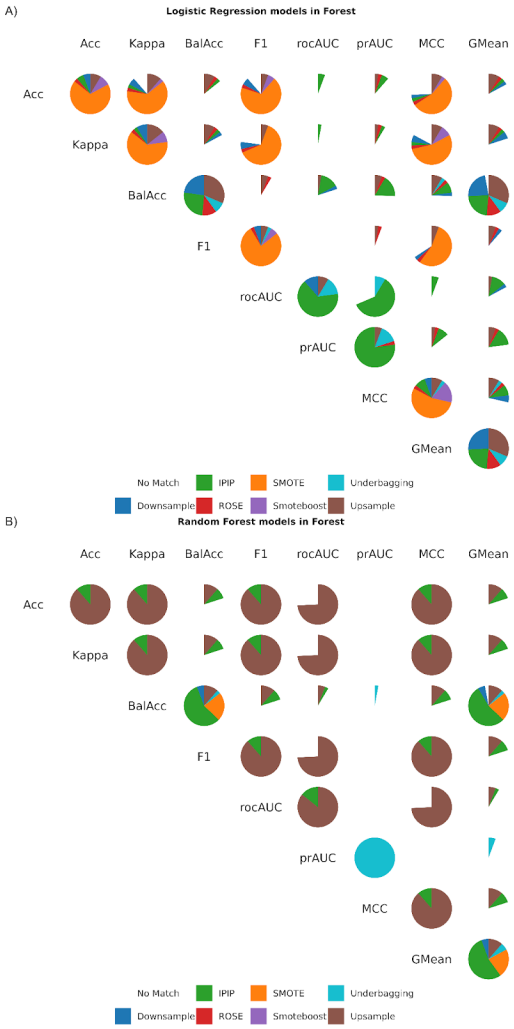


Figure 5.10: *A) Concordance plot of the Forest dataset using Logistic Regression. B) Concordance plot of the Forest dataset using Random Forest models.*

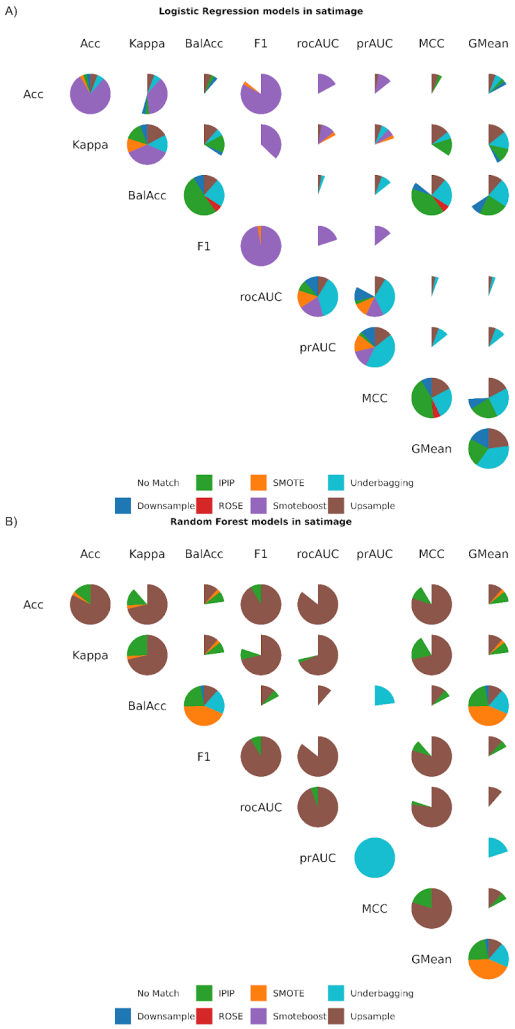


Figure 5.11: *A) Concordance plot of the Satimage dataset using Logistic Regression. B) Concordance plot of the Satimage dataset using Random Forest models.*

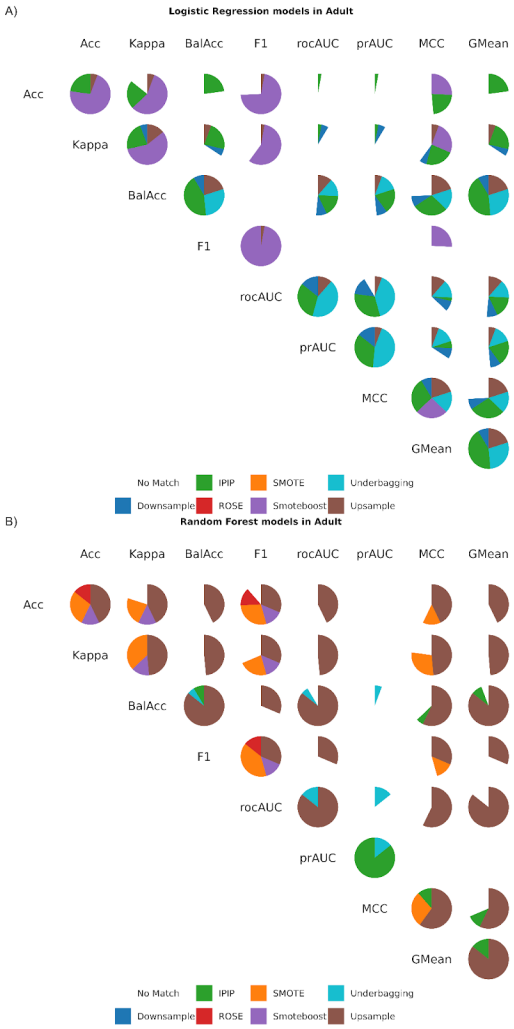


Figure 5.12: *A) Concordance plot of the Adult dataset using Logistic Regression. B) Concordance plot of the Adult dataset using Random Forest models.*

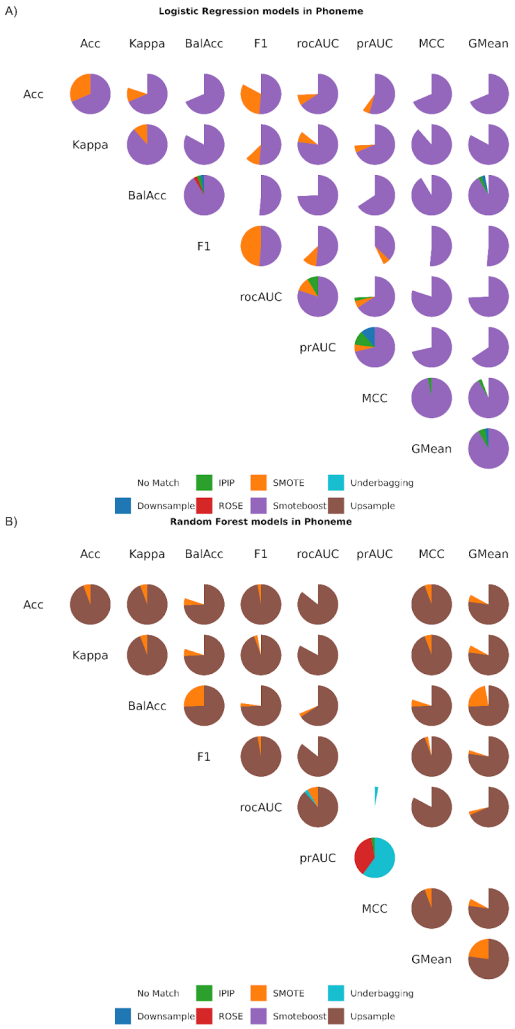


Figure 5.13: *A) Concordance plot of the Phoneme dataset using Logistic Regression. B) Concordance plot of the Phoneme dataset using Random Forest models.*

5.5.2 The proportion of the minority class biases all binary classification metrics

Once we have seen that there is no consensus among metrics, it would make sense to ask whether the metrics are distorted to the point of showing such a strong disagreement. This may be important if we want to develop an unbiased method for selecting the best imbalance-aware algorithm for a given problem. If we can estimate the level of distortion, we might use such estimates in our favour. Our goal is to consolidate all existing metrics into a single, minimally biased metric that enables accurate measurement of the performance of IAAs [283].

To uncover the bias of standard performance metrics due to the proportion of minority classes, we conducted a correlation analysis for each metric and $\mathbf{p_{min}}$. Fig. 5.14-A and Fig. 5.14-B display the correlations between techniques (rows) and metrics (columns) in the form of a heatmap using a colour gradient from white (no correlation) to red (high absolute correlation, either positive or negative) for Diabetes and Mammography datasets (Figures 5.15-5.19 contain plots of the remaining datasets). To the right of the heatmap, we included two new columns. One column shows the average of all the metrics, and the new metric, BAIC, is presented in the other column.

Additionally, Table 5.3 displays the Euclidean distance of the absolute correlation values between each metric and the 0 vector (unbiased correlation vector) for each dataset. This is further explained in Sec. 5.4.2.2. Now, leaving the BAIC metric aside, for the Diabetes dataset, we found that the least correlated metric with p_{min} is Accuracy, with a distance of 1.05 from the unbiased correlation vector. Following this is the rocAUC metric, with a distance of 1.10. More specifically, when looking at the Fig. 5.14-A, the Accuracy shows -0.005 as the lowest correlation on the Underbagging with the Logistic Regression model and -0.714 as the highest correlation on Smoteboost with the Logistic Regression model.

In the Mammography dataset, and again leaving aside BAIC, the metric that is closest to 0, in terms of distance between the correlations and the unbiased correlation vector (shown in Table 5.3), is the rocAUC, with a distance value of 1.46. This is followed by the Accuracy, with a value of 1.84. More specifically (See Fig. 5.14-B), the rocAUC showed -0.139 as the lowest correlation when using Rose and the Logistic Regression model. In contrast, the highest correlation, 0.805, was obtained when using Upsample together with Random Forest. In general, based on Fig. 5.14, we can observe that for every metric at least one method obtains a correlation value between the metric and p_{min} greater than 0.5, so it is not possible to conclude that there exists a metric that consistently has a bias towards 0 with respect to $\mathbf{p_{min}}$. Table 5.3 consistently shows optimally low distance between BAIC's correlation vector and the unbiased correlation vector.

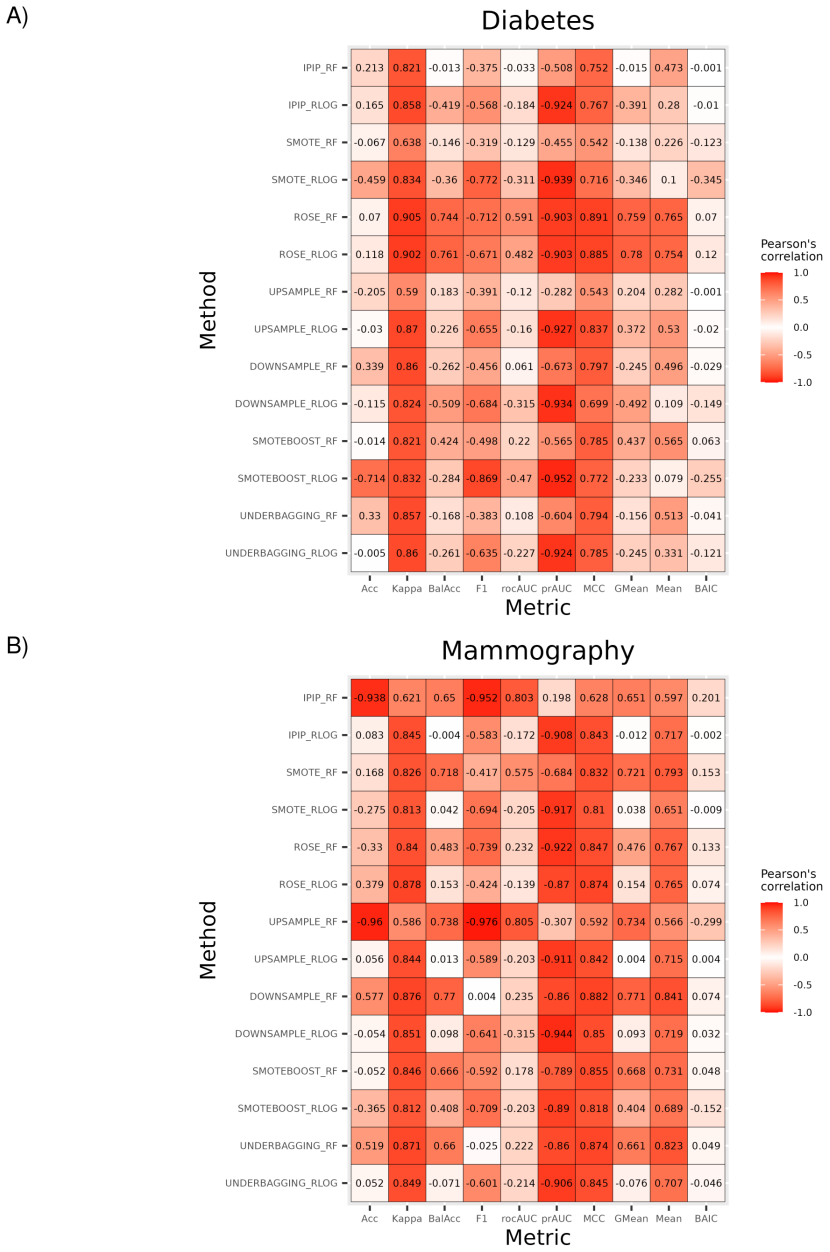


Figure 5.14: *Correlation heatmaps between metrics and p_{min} in Diabetes (A) and Mammography (B) datasets.*

	Acc	Kappa	BalAcc	F1	rocAUC	prAUC	MCC	GMean	Mean	BAIC
SMS	2.27	2.09	2.23	2.14	**2.00**	**2.00**	2.18	**2.00**	2.33	**2.00**
Forest	3.25	3.43	1.29	3.29	1.46	3.05	3.44	1.44	3.36	**1.19**
Phoneme	1.28	3.27	1.70	2.88	1.50	2.80	3.19	1.70	2.20	**0.77**
Mammography	1.71	3.05	1.84	2.36	1.46	3.05	3.06	1.84	2.71	**0.46**
Satimage	2.93	3.47	1.69	2.62	1.08	3.13	3.43	1.55	3.17	**0.69**
Diabetes	1.05	3.08	1.49	2.22	1.10	2.92	2.85	1.51	1.68	**0.51**
Adult	3.32	3.29	1.33	3.46	1.38	3.62	3.39	1.33	3.02	**1.29**

Table 5.3: *Distance matrix between correlation vectors for each metric and the unbiased correlation vector in each dataset. BAIC obtained the lowest distance value for all datasets in comparison to the remaining metrics.*

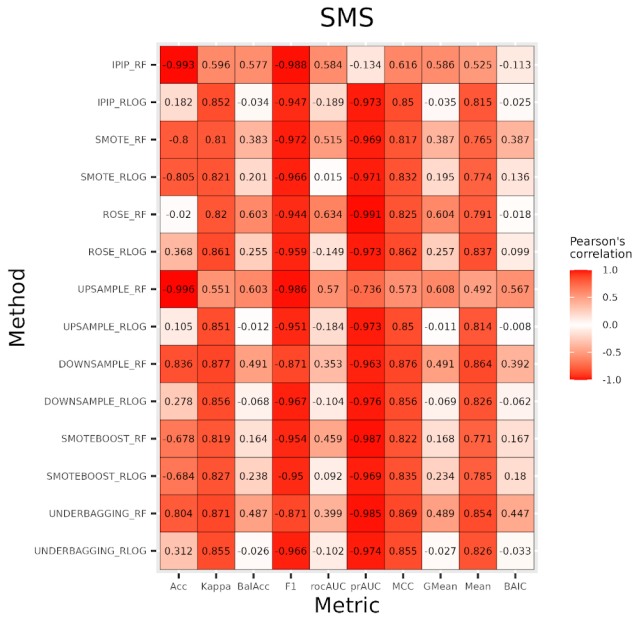


Figure 5.15: *Correlation heatmaps between metrics and pmin in SMS dataset.*

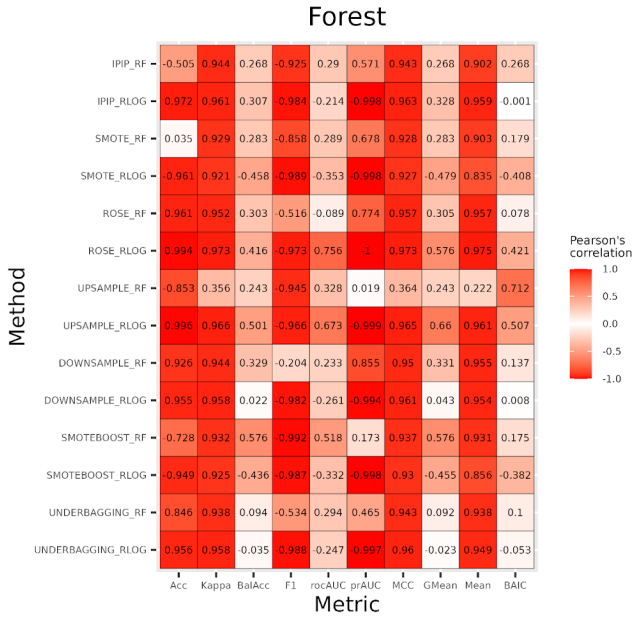


Figure 5.16: *Correlation heatmaps between metrics and pmin in Forest dataset.*

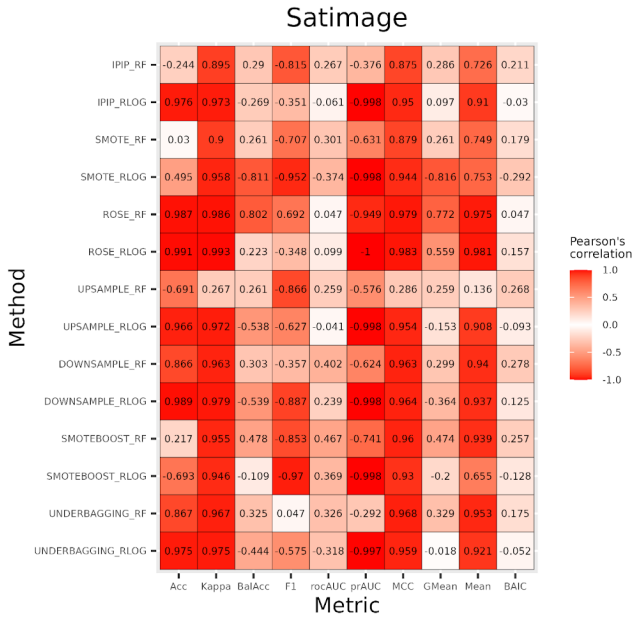


Figure 5.17: *Correlation heatmaps between metrics and pmin in Satimage dataset.*

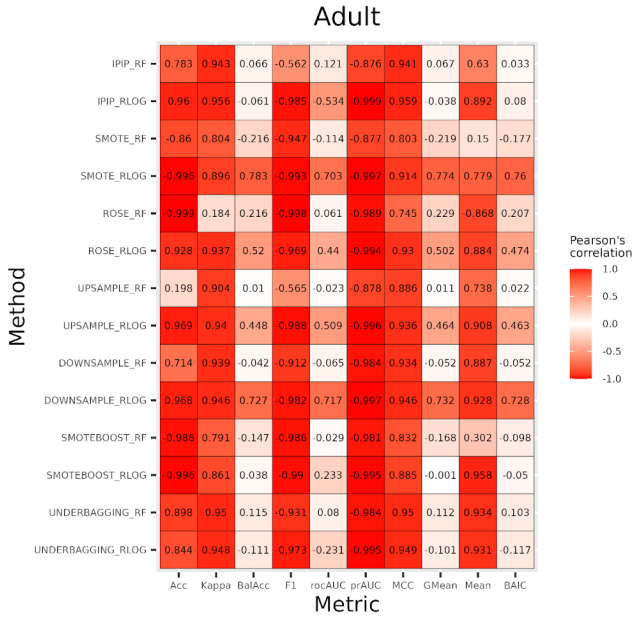


Figure 5.18: *Correlation heatmaps between metrics and pmin in Adult dataset.*

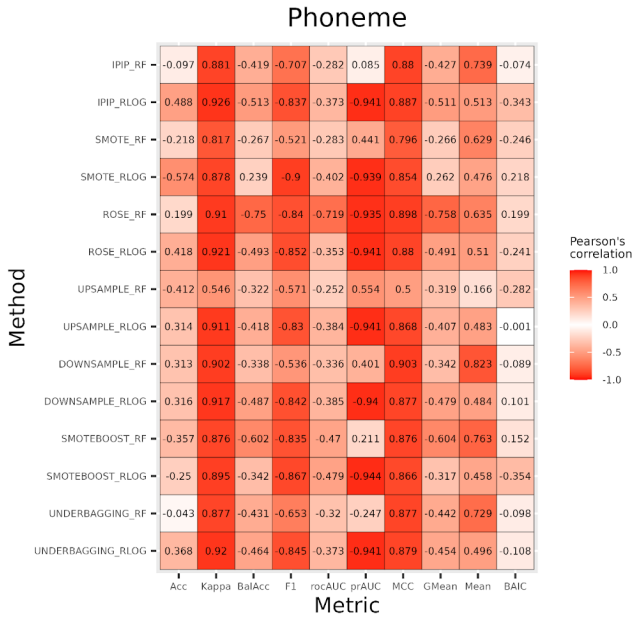


Figure 5.19: *Correlation heatmaps between metrics and pmin in Phoneme dataset.*

5.5.3 A weighted integration of the metrics reduces bias

As introduced in Sec. 5.3.2, the Pearson correlation between the performance metrics $\mathbf{m_{dk}}$ and the minority proportion vector $\mathbf{p_{min}}$ is proposed as an estimate of the bias towards minority proportions. To examine the effect regardless of direction, Fig. 5.20 depicts a boxplot illustrating the distribution of absolute correlation values for each metric, including BAIC and $\mathbf{p_{min}}$ obtained from all experiments. For every metric, we present the p-value resulting from the Wilcoxon signed-rank test performed between each metric and BAIC, with the Bonferroni correction applied. It is evident from the p-values (all $p < 0.0001$) that there exist statistically significant differences, implying that BAIC is less susceptible to bias towards $\mathbf{p_{min}}$, with a median[IQR] correlation value of 0.12[0.19].

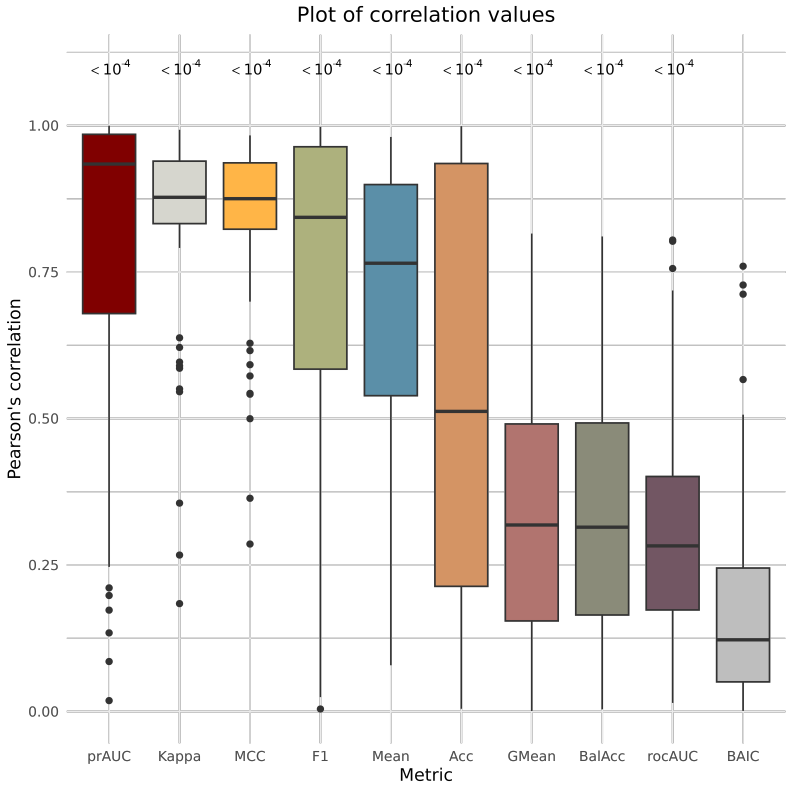


Figure 5.20: *Boxplots of correlations of the different metrics across all experiments ordered by median from highest to lowest. Above each boxplot is the p-value of the Wilcoxon's Test vs BAIC correlation values.*

5.5.4 Results of BAIC metric

After determining that our experiments imply BAIC values are less biased to p_{min}, we aimed to evaluate the performance of all IAAs, including IPIP, using BAIC, because every conventional measure (accuracy, F1, rocAUC, etc.) highlights a different trade-off between false positives and false negatives, so treating them in isolation risks favouring whichever aspect a given metric over-emphasises. Second, those same metrics are unequally distorted by class prevalence; some inflate as p_{min} decreases, others deflate, and a few even change direction. By computing the empirical sensitivity of each metric to p_{min} and then down-weighting the most sensitive ones, we can fuse their complementary information while suppressing their individual biases. BAIC aggregation behaves like a less biased integrated measure, retaining the discriminative power of the original measures while remaining stable across changes in imbalance. This criterion can be employed as a standard by which to select or compare imbalance-aware algorithms, thus ensuring that the "best" model is not an artefact of whichever single metric happens to be consulted. Table 5.4 illustrates the BAIC for all five algorithms and all seven datasets, with logistic regression and Random Forest as base learners. The table does not yield conclusive results regarding the best algorithm across all datasets. IPIP achieves the highest BAIC value in four datasets. Upsample gets the highest BAIC value in two datasets, and UnderBagging achieves the highest BAIC value in a single dataset. For the Phoneme and Diabetes datasets, which are the two datasets with the lowest number of instances (see Table 4.4), IPIP with RF achieves the highest BAIC value. However, for the Satimage dataset, which has the highest number of attributes (See Sec. 5.4.1), IPIP with Logistic Regression obtains higher BAIC values. In contrast, for the Mammography dataset, which is characterised by a severe imbalance and one of the smallest numbers of attributes, IPIP with Logistic Regression yields lower BAIC values.

Interestingly, we detected a strong negative association between the number of instances in the dataset and the BAIC values achieved by IPIP when using random forest as the base learner (correlation coefficient -0.62). This suggests that we should use random forest as IPIP's base learner on datasets with a small number of instances. There is also a positive association between the IR of the dataset and the BAIC values obtained when using IPIP together with logistic regression (correlation coefficient 0.82). Therefore, logistic regression remains a viable option as a base learner in problems with high imbalance. IPIP is faster with this algorithm.

	SMS	Mammography	Forest	Satimage	Adult	Phoneme	Diabetes
IPIP_RF	0.32	0.38	0.55	0.74	2.29	**1.69**	**2.54**
IPIP_RLOG	2.60	**3.05**	0.53	**1.49**	1.57	0.05	0.83
SMOTE_RF	0.07	0.53	1.43	1.39	1.33	0.86	2.21
SMOTE_RLOG	1.68	2.28	0.07	0.04	0.00	0.41	0.19
ROSE_RF	0.91	0.38	1.07	0.86	1.40	0.33	0.66
ROSE_RLOG	1.04	1.74	0.02	0.88	0.02	0.08	0.54
UPSAMPLE_RF	0.00	0.09	0.99	0.77	**3.11**	0.44	1.67
UPSAMPLE_RLOG	**2.93**	3.01	0.00	1.17	0.02	0.15	1.46
DOWNSAMPLE_RF	0.07	1.24	0.86	0.31	2.46	0.34	1.15
DOWNSAMPLE_RLOG	2.52	2.41	1.94	0.26	0.00	0.12	0.65
SMOTEBOOST_RF	0.97	1.37	0.36	0.33	1.81	0.39	1.04
SMOTEBOOST_RLOG	1.30	0.48	0.10	0.82	1.80	0.35	0.33
UNDERBAGGING_RF	0.04	1.27	1.74	1.27	2.09	1.21	1.53
UNDERBAGGING_RLOG	2.63	2.77	**1.98**	0.81	1.56	0.09	1.36

Table 5.4: *BAIC values of the performance of the different models on the seven datasets.*

5.5.4.1 Execution times analysis.

Table 5.5 presents the time taken by every IAA method on each dataset. It takes into account the average time spent implementing each technique, using the Logistic Regression and Random Forest ML models, in a five-fold cross-validation procedure. IPIP is, on average, 16.25 times slower than the fastest alternative, Downsample, and 1.37 times faster, on average, than the slowest alternative, Smoteboost. Note that SmoteBoost is the algorithm that requires the most time to complete on SMS, Mammography, Forest, and Satimage datasets, whereas IPIP is the longest-running algorithm on Adult, Phoneme, and Diabetes datasets. In contrast, Downsample, which trains a model on a smaller dataset, is the most straightforward method and takes the least time on all datasets. All experiments were performed on an AMD Ryzen Threadripper 3960X (24 cores) server with 252 GB RAM and using Linux kernel 5.10.0-31-amd64.

	SMS	Mammography	Forest	Satimage	Adult	Phoneme	Diabetes	Mean
DOWNSAMPLE_RLOG	1.11	0.50	3.66	1.23	71.81	0.79	0.71	**11.40**
ROSE_RLOG	32.36	1.73	24.76	3.43	141.23	1.02	0.84	**29.34**
SMOTE_RLOG	3.82	0.86	15.08	2.25	229.61	1.67	1.32	**36.37**
UPSAMPLE_RLOG	68.95	3.23	45.20	5.97	192.95	1.25	0.85	**45.49**
UNDERBAGGING_RLOG	7.26	3.63	26.70	6.47	321.70	4.63	5.39	**53.68**
SMOTEBOOST_RLOG	167.87	10.78	203.37	21.12	94.02	8.70	4.47	**72.90**
DOWNSAMPLE_RF	45.01	17.56	67.45	22.04	436.57	48.40	17.03	**93.44**
IPIP_RLOG	88.67	22.15	79.35	30.55	1312.81	17.98	16.89	**224.06**
SMOTE_RF	97.72	25.19	180.59	220.47	2152.45	93.10	24.45	**399.14**
ROSE_RF	1166.42	116.60	466.19	66.48	1089.99	88.86	19.77	**430.62**
UNDERBAGGING_RF	292.65	116.69	559.34	153.19	2367.74	313.43	113.09	**559.45**
UPSAMPLE_RF	1801.88	108.32	497.46	78.40	1787.54	90.64	20.88	**626.45**
IPIP_RF	822.10	397.58	1151.21	419.51	6434.39	794.04	338.84	**1479.68**
SMOTEBOOST_RF	5396.05	565.39	2935.45	508.42	5652.97	602.18	143.41	**2257.70**

Table 5.5: *Elapsed mean time in seconds taken to train a 5-fold cross-validation of each model ordered by column Mean, which is the mean time taken to train each model across all datasets.*

5.6 Concluding the chapter

In this chapter, a novel approach to addressing imbalanced datasets is proposed. Firstly, empirical investigations were conducted, revealing that most commonly used performance metrics exhibit inadequate concordance and a tendency to overemphasise the minority class proportion. To address the issues mentioned above, a new metric is needed that incorporates existing metrics while also considering their properties. This new metric would serve to reduce the correlation between the metrics and the proportion of the minority class. It is therefore proposed that BAIC be introduced as a new metric to measure the performance of machine learning algorithms on imbalanced data.

The BAIC measure, which encompasses various metrics analysed in this chapter, exhibits a statistically significant lower correlation ($p < 10^{-4}$) with the proportions of minority classes. BAIC is a suitable alternative strategy for selecting the optimal model on a given dataset. This innovative metric is particularly advantageous in a production environment where changes in class proportions over time can negatively impact model performance. As BAIC has a lower correlation with the imbalance ratio, it can help identify a more effective ML model in cases where changes in class proportions degrade the model's performance. Unlike prior attempts to integrate multiple basic metrics, that average scores uniformly, BAIC assigns weights estimated on the target dataset, suppressing whichever metric inflates/deflates with p_{min}.

Secondly, we have also proposed IPIP, a novel IAA ensemble-based technique. IPIP involves obtaining resamples of the original dataset that are balanced in such a way that all instances of the minority class are represented in at least one of the samples. Another significant advantage of IPIP is that it does not generate synthetic instances. The algorithm has shown promising results as a viable alternative with higher BAIC values in four out of seven datasets. It is noteworthy that IPIP, when combined with Random Forest as the base learner, has the highest BAIC for the Phoneme and Diabetes datasets, which have the fewest instances. This finding emphasises the value of retaining all observations when data are limited. IPIP with Logistic Regression has the highest BAIC on the Satimage dataset, which has the most significant number of attributes, and on the Mammography dataset, characterised by severe imbalance and a very small feature set. We found a strong negative correlation of –0.62 between dataset size and the BAIC obtained by IPIP with Random Forest, suggesting Random Forest is preferable for small datasets. In contrast, a strong positive correlation of 0.82 between the imbalance ratio and the BAIC delivered by IPIP, as determined by logistic regression, indicates that logistic regression remains a practical option for highly imbalanced problems. Regarding IPIP execution time, it is considered efficient when compared to other imbalance-based methods. IPIP's complexity (See Eq. 5.13) comes from training several balanced resamples and checking at each step whether adding a new model really improves results. Empirically, IPIP is ~16× slower than Downsample but ~1.4× faster than SMOTEBoost

on average across datasets (see Table 5.5). Such computational complexity is affordable when the imbalance is very high and covering all minority cases is essential and when oversampling would add too much overlap or noise. For large datasets, IPIP can be made faster by running resamples in parallel and by stopping the ensemble growth when improvements in adding models to it become reasonably small. To calculate the unbiased BAIC evaluation metric, a high computation time is required as experiments are performed on multiple auxiliary datasets to aggregate the basic performance metrics.

Furthermore, while BAIC and IPIP provide a robust and reliable baseline for learning from stationary imbalanced datasets, real-world applications rarely operate in strictly static environments. As data continuously arrives in streams, both the underlying patterns and the minority class proportions are able to shift over time. Consequently, the next chapter will build upon these components, extending the FILM framework to adaptively handle concept drift in non-stationary data streams.

6

Extending the Framework to Dynamic Data Streams

The methods introduced in the previous chapter address a problem commonly found in real-world applications: evaluation and learning under class imbalance, where the class of interest is rare. However, that chapter considered this problem in a simpler setting, assuming that the data distribution remained static over time and that class imbalance was the main challenge. As discussed throughout this thesis, clinical environments and production systems continuously evolve.

The next step in this thesis is therefore to extend these ideas to changing environments. In this chapter, the FILM framework is extended to dynamic data streams, where data arrive sequentially over time and the underlying concepts may change, producing the phenomenon known as concept drift.

To achieve this, the chapter focuses on extending the static components of the FILM framework to dynamic scenarios:

- Adaptive BAIC is a time-dependent version of the evaluation method that updates metric weights sequentially over time. This allows the evaluation metrics to remain less biased as the instances distribution and the minority class proportion fluctuate throughout the stream.
- IPIP-CD is an extension of the IPIP algorithm designed for continuous data streams. The method continuously evaluates and updates the ensemble using the most recent data chunks, incorporating new models trained on recent data while removing historical models whose performance has degraded, without relying on statistical drift detectors.

In this way, the methods developed in this chapter are based on the static framework introduced in the previous chapter to provide an adaptation strategy for scenarios where

class imbalance and concept drift occur simultaneously.

6.1 Introduction

Machine learning algorithms aim to learn and generalize patterns from data. However, real-world environments often undergo changes over time [8, 317], which gives rise to the previously introduced phenomenon known as concept drift (See Chapter 3). As previously defined in Chapter 1 (Section 1.2), concept drift is defined as a discrepancy between the statistical properties of the target variable in the data used for training a model and those observed in the data deployed. This is particularly challenging in the context of classification tasks, as models trained on outdated data may encounter difficulties in accurately predicting new instances, which can lead to a decline in performance. It is of paramount importance to detect and adapt to concept drift in order to maintain model reliability in dynamic environments.

The challenge is further exacerbated when concept drift is combined with class imbalance. In binary classification tasks, one class may be significantly underrepresented in comparison to the other, which results in models that are biased towards the majority class. When the distribution of classes shifts over time due to concept drift, this imbalance can lead to a further drop in model performance. The minority class, which is inherently more challenging to learn, may become even less distinguishable, resulting in suboptimal performance on the instances where accurate predictions are most crucial. It is imperative that both concept drift and class imbalance are addressed simultaneously in order to construct robust and adaptive machine learning systems. The implementation of such systems in real-world applications is the main focus of Machine Learning Operations (MLOps) [318]. MLOps is defined as the set of practices and tools aimed at automating and scaling the lifecycle of machine learning models in production, ensuring their reliability through continuous integration, deployment, and monitoring [319]. In the context of dynamic data streams, MLOps serves to manage the automated updating of models as there are new data distributions, preventing the degradation of predictive performance.

Building on the analysis presented in Chapter 5 regarding the limitations of performance assessment, one of the major challenges introduced by class imbalance in the presence of concept drift is its effect on evaluation metrics. As previously established, common metrics like accuracy can become misleading, as they may mask poor model performance on the minority class while showing high overall accuracy. However, class imbalance combined with concept drift can lead to further confusion. A decline in model performance may be mistakenly attributed solely to concept drift, when in fact, the imbalance in the data may also be a significant factor. In such cases, the underlying shift in the data distribution due to drift may go unnoticed because the imbalance obscures the true extent of the change. For instance, in a fraud detection system, where fraudulent transactions are

rare, the model's ability to detect fraud may degrade over time, not only due to changes in fraudulent behaviour but also because the imbalance persists or worsens.

In section 5.5.2, we have demonstrated that evaluation metrics can be biased by class imbalance, leading to disagreements between metrics when handling imbalanced problems. This issue becomes even more complex when concept drift is introduced. As class distribution shifts over time, the standard evaluation metrics may further misrepresent the true performance of the model, highlighting the need to address both class imbalance and concept drift simultaneously. A robust performance metric capable of handling both phenomena is essential. To this end, we propose the Adaptive BAIC (Bias-Adjusted Integration Coefficient), a novel metric designed to provide an unbiased evaluation across imbalanced datasets affected by concept drift. The Adaptive BAIC integrates fundamental metrics through a weighted sum, ensuring that minority class performance is not overshadowed by the majority class, even in the presence of drift.

In addition to evaluating adaptation through the Adaptive BAIC, this chapter extends the overarching FILM framework developed previously by introducing IPIP-CD. Relying on the foundational balanced subsampling mechanics of its static counterpart (IPIP), IPIP-CD provides a stable passive-learning approach for dynamic streaming environments. To empirically evaluate this streaming extension, we compare it against state-of-the-art ensemble-based algorithms such as the Accuracy Weighted Ensemble (AWE) [320], Uncorrelated Bagging (UB) [321], SElectively Recursive Approach (SERA) [322], and Learn++ CDS [323], alongside a basic Random Forest model as our baseline.

6.2 Background

To contextualize the performance of the proposed IPIP-CD framework, it is essential to understand the underlying mechanisms of the ensemble-based methodologies used for comparison. These baseline algorithms represent different paradigms in the streaming literature for addressing the dual challenge of non-stationary distributions and class imbalance.

The Accuracy Weighted Ensemble is a classic chunk-based ensemble method designed for data streams. It is based on weighting each base classifier in the ensemble according to its predictive accuracy on the most recent data chunk. This approach allows the system to effectively forget outdated concepts by continuously reducing the influence of models trained on obsolete data. However, in highly imbalanced scenarios, AWE exhibits limitations; its weighting mechanism inherently prioritizes overall accuracy, which frequently masks performance degradation on the minority class.

Alternatively, Learn++ CDS (Concept Drift with Stepwise Learning) is an extension of the Learn++ algorithm focused on non-stationary environments with skewed class distributions. This algorithm addresses class imbalance by incorporating synthetic data

generation, typically through SMOTE, within each training step to balance the incoming data chunks. Learn++ CDS maintains an evolving ensemble of classifiers where each new model is trained on a distribution that focuses on instances incorrectly classified by previous components. This stepwise learning strategy allows the system to adapt to new concepts while attempting to preserve minority class representation through targeted oversampling.

The SElectively Recursive Approach takes a different path by managing the historical memory of the minority class to stabilize the learning process. Unlike standard chunk-based methods, which process the continuous data stream in discrete, sequential blocks and strictly discard old chunks, SERA selectively retains minority class samples from previous time steps that are statistically similar to the current data. By combining these relevant historical samples with the newly arrived data, the algorithm constructs a balanced training set. This selective recursion strategy helps maintain a robust decision boundary for the minority class, particularly when addressing recurring concepts or sudden drift episodes where minority instances become scarce.

Finally, Uncorrelated Bagging adapts the basic bagging algorithm to dynamic environments by focusing on the diversity of the base learners. This method constructs multiple classifiers in parallel, utilizing different resamples of the current data chunk. The uncorrelated aspect refers to its active strategy of ensuring that the classification errors made by individual ensemble members are not highly correlated. This structural enhances the overall generalisation and robustness of the ensemble during concept drift episodes, providing a strong and robust baseline for passive adaptation.

In the field of non-stationary environments, adaptation strategies are categorised into active and passive approaches. Active adaptation relies on explicit drift detection mechanisms, mainly based on statistical tests, to identify when a model has become obsolete and requires retraining. In contrast, passive adaptation involves a continuous or periodic evolution of the model as new data arrive, without the need for a formal detection step.

Methods relying on synthetic data generation, such as Learn++ CDS, risk introducing artificial noise or overfitting the minority class, whereas accuracy-weighted approaches like AWE are susceptible to performance masking when the class imbalance is severe. To bridge these gaps, the IPIP-CD framework introduces a novel, data-preserving passive adaptation strategy. IPIP-CD distinguishes itself by achieving class balance strictly through strategic subsampling rather than synthetic data generation. It ensures that every minority class instance is utilised across multiple balanced partitions while simultaneously exploiting a portion of the majority class across the ensemble. This approach avoids the noise amplification of oversampling techniques while maintaining a highly diverse set of base learners, similar to the idea of Uncorrelated Bagging but for severe imbalance problems.

6.3 Methods

6.3.1 Adaptive BAIC method for dynamic data streams

In Chapter 5, BAIC was introduced as a bias-adjusted integration of several evaluation metrics. The core idea was to weight each metric according to how strongly it depends on the minority class proportion, so that model selection does not change merely because p_{min} changes.

In a data stream, this issue becomes even more critical. First, it is not enough now to measure performance with a single estimate: metrics evolve over time as new chunks arrive and the concept may drift. Second, the association between a metric and the minority proportion is not necessarily constant throughout the stream. For example, during a drift episode, some metrics may become more sensitive to p_{min} changes than they were before, while other metrics may remain stable. Using a single set of static BAIC weights would therefore mix two effects: true changes in the concept and artificial changes driven by p_{min}.

To address this, we extend BAIC to a time-dependent version where the metric weights are updated along the stream. The resulting score is called *Adaptive BAIC*. Instead of a single weight vector, the method produces a sequence of weight vectors $\{\mathbf{w}(t)\}_{t=1}^{T}$, one per time step. This makes the integration robust to time-varying p_{min} effects and provides a single performance curve over time that can be monitored together with the original metrics, as represented in Figure 6.1.

Let the stream be processed in T steps (chunks or evaluation points). At each step t, we compute a set of k metrics:

$$\mathbf{m}(t) = \big(m_1(t), \ldots, m_k(t)\big),$$

and we denote the minority proportion at that step by $p_{\min}(t)$.

In our experiments, each dataset is evaluated under several controlled p_{min} settings (several fixed values of $p_{\min}$). This produces, at each time step t, a set of paired observations $\{(\mathbf{m}^{(q)}(t), p_{\min}^{(q)})\}_{q=1}^{Q}$ across Q p_{min} configurations. This is precisely what enables estimating p_{min} bias as a function of time.

For each metric j and time step t, we estimate how much $m_j(t)$ depends on the minority proportion by computing the correlation across the p_{min} configurations:

$$r_j(t) = \operatorname{corr}\Big(\big[m_j^{(1)}(t), \ldots, m_j^{(Q)}(t)\big],\ \big[p_{\min}^{(1)}, \ldots, p_{\min}^{(Q)}\big]\Big). \tag{6.1}$$

A correlation close to zero indicates that metric j behaves similarly across different p_{min} levels at time t, while a larger magnitude indicates stronger dependence.

As in Section 5.3.2, correlations are converted into weights through a Gaussian penalty centred at zero:

$$w_j(t) = G\big(r_j(t)\big), \tag{6.2}$$

where $G(\cdot)$ is the same Gaussian-shaped function used in BAIC with parameters $a = 1$, $b = 0$, $c = 0.15$. Finally, the Adaptive BAIC at time t is computed as a weighted sum of the metrics at time t:

$$\text{Adaptive BAIC}(t) = \sum_{j=1}^{k} w_j(t-1)\, m_j(t). \tag{6.3}$$

We use a one-step lag $(t-1)$ so that the weights applied at time t do not use information from the same time step, which avoids optimistic bias in an online setting. In other words, weights act as a calibration learned up to the previous step and then applied to the current performance vector.

Adaptive BAIC is not meant to replace the original metrics, but to summarise them into a single monitoring signal that is less affected by p_{min} shifts. When drift happens, $\text{BAIC}(t)$ can drop due to genuine loss of predictive performance, while the weighting reduces spurious oscillations caused by changes in $p_{\min}$. This is especially useful when comparing adaptive learners across streams with different imbalance regimes, or when the minority proportion varies over time. An overview of the adaptive BAIC scheme can be observed in Figure 6.1

6.3.2 The IPIP passive adaptation to concept drift

As established in Chapter 5, the FILM framework decomposes the learning problem into a structured sequence of ensembles (IPIP) to mitigate the bias introduced by severe class imbalance. While the original IPIP construction guarantees robust representation in static datasets, translating this capability to streaming data requires a temporal adaptation strategy.

From a dynamic learning perspective, FILM operates as an evolving meta-ensemble, a structured pool of individual ensembles, where each ensemble is trained on a specific data chunk to capture its characteristics. Rather than relying on explicit drift detection mechanisms or aggressive retraining strategies, the IPIP-CD extension follows a passive update paradigm in which the ensemble evolves automatically as new data arrive sequentially. When a new data chunk is observed, the framework applies the same imbalance-aware partitioning process to generate additional balanced training sets, from which new ensemble members are trained. These new ensembles are appended to the existing meta-ensemble method, enabling IPIP-CD to progressively incorporate recent information without disrupting previously learned knowledge.

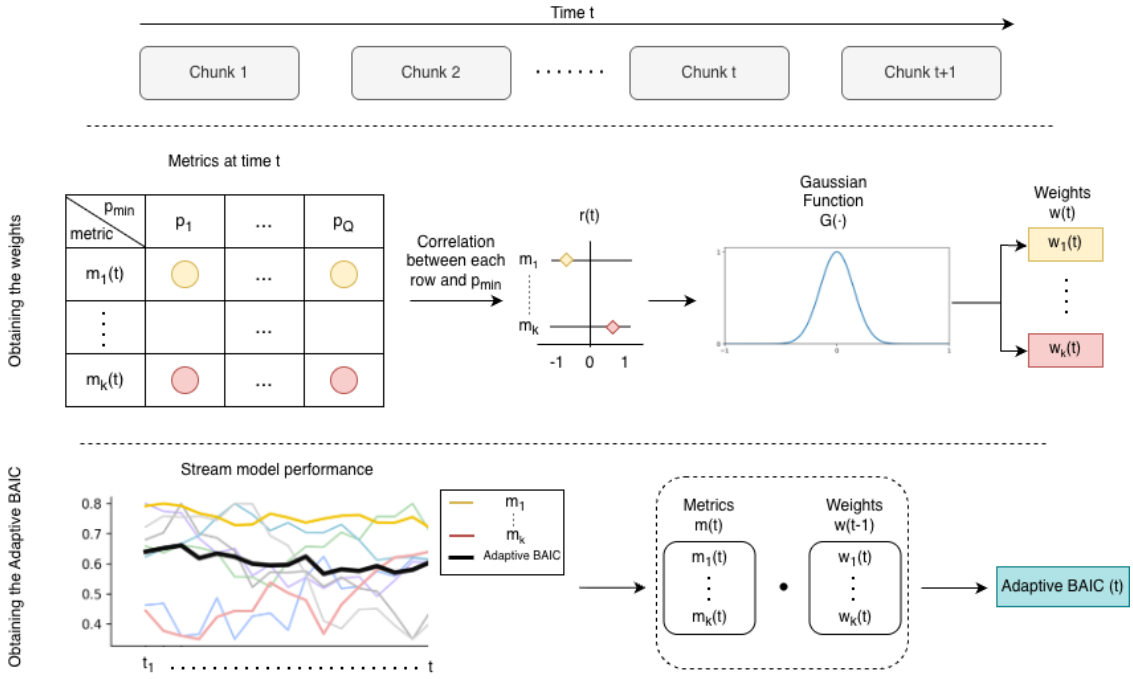


Figure 6.1: *Overview of the Adaptive BAIC computation in a chunk-based data stream. At each time step t, a vector of performance measures $\mathbf{m}(t) = (m_1(t), \ldots, m_k(t))$ is obtained (bottom-left). For the same time step, metric values are compared across a set of fixed imbalance settings $\{p_1, \ldots, p_Q\}$ to estimate the correlation $r_j(t)$ between each metric $m_j(t)$ and $p_{\min}$ (middle). These correlations are mapped through a Gaussian function to produce metric weights $w_j(t)$ (right). The integrated score is then computed online using a one-step lag, combining current metrics with the previous weights via a dot product, yielding* $\text{Adaptive BAIC}(t) = \sum_{j=1}^{k} w_j(t-1)\, m_j(t)$ *(bottom-right).*

Simultaneously, the framework continuously evaluates the predictive performance of each ensemble component on recent data using an imbalance-aware metric (Adaptive BAIC). The ensembles obtaining the lowest performance are considered obsolete and removed from the meta-ensemble. This pruning mechanism allows IPIP-CD to forget previously learnt outdated concepts while retaining relevant ensembles trained on recent data, naturally adapting to data changes over time. This would work dealing both with an abrupt drift, because all previous ensembles can be replaced by new ones, and gradual drifts, because ensembles can be added gradually over time, without explicit manual intervention on the model.

When integrating FILM into an MLOps system, a passive concept drift management strategy is adopted. In this context, passivity does not imply the absence of adaptation, but rather the absence of explicit drift detection mechanisms. Instead of actively triggering reactions based on statistical drift tests, FILM assumes there will be drift at some point and consequently adapts automatically by continuously evaluating model performance as new data become available, allowing the ensemble to evolve according to its observed predictive behaviour. To enable this process, the system first defines a chunking strategy, which determines how the continuous data stream is partitioned into discrete units. These chunks constitute the fundamental operational unit for evaluation

and adaptation within FILM. Chunking can be performed based on (1) a temporal criterion, such as monthly or weekly splits, or (2) a fixed number of instances per chunk. This design ensures that model evaluation and updates are carried out in a controlled and homogeneous manner across time.

Once a new chunk is established and analysed, FILM applies its ensemble adaptation mechanism (See Algorithm 2). All existing ensemble models are evaluated on the complete new chunk using some performance measure (e.g., balanced accuracy, F1-score,...) , while a new IPIP-based ensemble is trained exclusively on the same chunk and evaluated under identical conditions. Based on these results, FILM retains a fixed number of ensemble components corresponding to the number of balanced subsets. This evaluation and pruning phase, detailed in Algorithm 3, selects the ensembles, either old or newly trained, that achieve the highest performance score on the current data. Through this selection process, the ensemble is updated without retraining the full system, naturally replacing obsolete models with better-performing ones. In parallel, the system enables the continuous monitoring of the performance score over time, providing a clear visualization of performance evolution under class imbalance. Additionally, tracking the number of ensemble components that are replaced at each update step offers insight into the system's adaptability. A high replacement rate indicates rapid adaptation to changing data distributions, while a low rate suggests that previously learned models remain effective. Together, these indicators allow practitioners to assess both performance stability and the dynamics of ensemble evolution within the MLOps pipeline.

Algorithm 2 IPIP_CD($\mathcal{S}, p_{min}, b_s, b_E, ML, tries_{max}, p_{train}, n_{keep}, metric_{max}$)

Input: Chunked data $\mathcal{S} = \{C_1, \ldots, C_T\}$; p_{min}; b_s; b_E; ML; $tries_{max}(\cdot)$; p_{train}; n_{keep}; $metric_{max}$.

Output: M: pool of ensembles adapted over time.

$M \leftarrow \emptyset$
for $t = 1$ **to** $T - 1$ **do**
  $P \leftarrow IPIP(C_t, p_{min}, b_s, b_E, ML, tries_{max}, p_{train})$
  $M \leftarrow M \cup P$
  $M \leftarrow AdaptiveEnsembleSelection(M, C_{t+1}, metric_{max}, n_{keep})$
  $\hat{y}_{t+1} \leftarrow \text{predict}(M, C_{t+1})$
end for
return M

Algorithm 3 AdaptiveEnsembleSelection$(M, X, metric_{max}, n_{keep})$

Input: Meta-ensemble $M = \{M_1, \ldots, M_q\}$; labelled chunk $X = (x, y)$; $metric_{max}$; n_{keep}.
Output: Subset of the n_{keep} best ensembles.

1: **for** $j = 1$ **to** q **do**
2: $\quad \hat{y}_j \leftarrow \text{predict}(M_j, x)$
3: $\quad s_j \leftarrow \text{metric}_{\max}(y, \hat{y}_j)$
4: **end for**
5: $I \leftarrow \text{argsort}(s, \text{descending})[1 : n_{keep}]$
6: **return** $\{M_i : i \in I\}$

6.4 Experiments

6.4.1 Datasets

The evaluation of the proposed drift adaptation methodology is performed on eight datasets widely used as benchmarks in the concept drift literature, including both real-world and synthetic streams. A comprehensive exposition of these streaming datasets, detailing dimensions and chunk counts, can be found in Chapter 4 (Section 4.2.2, Table 4.5). It is important to note that for the datasets generated via the synthetic Hyperplane generator, a specific nomenclature is used (*Hyperplane_2*, *Hyperplane_5*, and *Hyperplane_8*), where the numerical suffix explicitly denotes the number of attributes affected by the drift in the respective dataset. By including these variations, we aim to explore how different attribute dimensions drift impact the experimental outcomes and the framework's adaptability under a varying number of features.

6.4.2 Implementation Details

In contrast to the static approach presented in the previous chapter, the dynamic data requires a different evaluation strategy. The continuous data streams were divided into sequential temporal chunks. To assess the performance of the evolving models, a chunk-based prequential evaluation strategy (test-then-train) was employed. Under this paradigm, each ensemble is first evaluated on the incoming, unseen chunk $(t + 1)$ to simulate real-world predictions, and subsequently, the real labels of this chunk become available to update, prune, or expand the ensemble.

Random Forest was selected as the primary base learner for all implemented algorithms due to its robustness against noise and its natural suitability for ensemble architectures. The model building and stream processing was implemented in R using the *caret*, and using the *ranger* package for training the Random Forests. Additionally, the *DMwR* package was utilised to handle the artificial minority class oversampling required by specific

baseline algorithms.

To evaluate the effectiveness of the proposed IPIP-CD methodology, we compared it against several concept-drift and imbalance-aware ensemble techniques previously described in Section 6.2: Accuracy Weighted Ensemble, Uncorrelated Bagging, SElectively Recursive Approach, and Learn++ CDS, alongside a standard Random Forest model as the static baseline.

For the base learners across all ensemble frameworks, an automated hyperparameter search strategy was conducted. During this process, a 10-fold cross-validation was applied internally within each training chunk. As in the static framework, Cohen's Kappa is selected as the optimisation metric for the hyperparameter tuning. The number of trees was fixed at 200 for all Random Forest models. Given the varying dimensionalities of the datasets, specific hyperparameter grids were defined for each data stream, primarily adjusting the *mtry* parameter while maintaining a consistent evaluation of node sizes. The exact hyperparameter values evaluated for each dataset are detailed in Table 6.1.

Dataset	mtry	min.node.size	splitrule
Airlines	{1, 3, 5, 7}	{1, 11, 21}	gini
Elec	{1, 3, 5, 7}	{1, 11, 21}	gini
Hyperplane_2	{1, 3, 5, 7}	{1, 11, 21}	gini
Hyperplane_5	{1, 3, 5, 7}	{1, 11, 21}	gini
Hyperplane_8	{1, 3, 5, 7}	{1, 11, 21}	gini
SEA	{1, 2, 3}	{1, 11, 21}	gini
SMS	{4, 12, 20, 28, 36, 44}	{1, 11, 21}	gini
Weather	{1, 3, 5, 7}	{1, 11, 21}	gini

Table 6.1: *Values used to create the hyperparameter grid for Random Forest base learners across all streaming datasets.*

6.4.2.1 Area under the progression of the metrics along time

In streaming environments, evaluating the performance of a machine learning model cannot be reduced to a single static measurement. As models are continuously tested on incoming sequential data chunks, any performance metric m_j evaluated at time step t produces a discrete sequence of values $\{m_j(1), m_j(2), \ldots, m_j(T)\}$. When plotted across time, these sequential evaluations define a performance progression curve that visually captures the model's behaviour in the presence of concept drift and class imbalance.

To effectively compare the overall robustness of different models throughout the entire data stream, it is necessary to summarize this temporal progression into a single global score. A standard and robust approach to achieve this is calculating the Area under the Performance Curve (AUPC) defined by the progression of the metric over time [7]. This method provides a cumulative measure of performance, rewarding models that maintain high and stable metric values across all chunks, including their recovery phases during and after concept drift episodes.

Mathematically, the total AUPC for a given metric m_j can be approximated using the trapezoidal rule. To ensure that this global score is comparable across data streams of varying lengths, the accumulated area is normalised by the total duration of the evaluated stream. Given a set of T discrete time steps, the normalised AUPC A_j for a specific performance metric $j \in \{1, ..., k\}$ is computed as:

$$A_j = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{m_j(t) + m_j(t+1)}{2} \tag{6.4}$$

where the interval between consecutive chunks is assumed to be constant ($\Delta t = 1$). This normalization bounds the resulting score to the original scale of the metric. It is important to note that while standard performance measures are bounded in the $[0, 1]$ interval, the Adaptive BAIC is defined as a weighted sum of multiple metrics (see Eq. 6.3). Since it is not upper-bounded, its values, and consequently its normalised area, can exceed 1. By calculating this normalised area for all the standard metrics, as well as for the proposed Adaptive BAIC, we obtain a comprehensive evaluation metric value. This value effectively captures the model's robustness, its degradation during drift, and its adaptability throughout the evolving data stream, allowing for a fair and unified comparison between different learning strategies.

6.4.2.2 Correlation between the area under the progression and the minority class proportion

A fundamental question when evaluating learning algorithms in dynamic data streams is whether the overall temporal performance, summarised by the AUPC, remains biased by the minority class proportion, p_{min}. To address this issue, we adapt the correlation analysis introduced in the static framework (Chapter 5) to non-stationary scenarios.

Specifically, we investigate the correlation between the global performance score, defined as the AUPC value A_j for each evaluation metric j (as formulated in Eq. 6.4), and the varying minority class proportions, $\mathbf{p_{min}}$, injected across the different experimental stream configurations.

To conduct this analysis, we obtain the AUPC generated by all evaluated ensemble strategies across the different datasets. Moreover, the AUPC values for our proposed Adaptive BAIC metric are integrated into this same experimental framework, allowing for a direct and unified comparison against standard metrics. This is designed to determine whether the dynamic integration of evaluation measures successfully mitigates the bias toward p_{min} throughout the entire streaming process.

6.5 Results

6.5.1 The AUPC value of the Adaptive BAIC progression is less biased by the minority class proportion

As established in Section 6.4.2.2, evaluating models in dynamic environments requires summarizing their performance over time. To ensure that this global evaluation is not driven by the class imbalance, it is important to analyze the correlation between the AUPC (A_j, see Eq. 6.4) and the minority class proportion ($\mathbf{p_{min}}$).

Following the same representation as Figure 5.20 in Chapter 5, Figure 6.2 shows boxplots illustrating the distribution of Pearson's correlation values between the global AUPC values of each metric and $\mathbf{p_{min}}$ across the eight experimental data streams mentioned in Section 6.4.1. It can be observed that standard evaluation metrics, particularly the F1-score, prAUC, and Cohen's Kappa, exhibit a severe dependency on the class distribution, presenting median [IQR] correlation values of 0.95 [0.18], 0.93 [0.12], and 0.90 [0.12], respectively. Even the rocAUC metric, which is frequently considered less affected by the imbalance [324], presents a high correlation distribution with p_{min}, with a median correlation of 0.79 [0.34].

In contrast, the proposed Adaptive BAIC yields the lowest median correlation among all the evaluated metrics, obtaining a median [IQR] correlation value of 0.48 [0.42]. To statistically validate this visual observation, a pairwise non-parametric Wilcoxon signed-rank test was conducted, comparing the correlation distribution of Adaptive BAIC against that of every standard metric. The resulting p-values, displayed above each corresponding boxplot in Figure 6.2, demonstrate highly significant statistical differences ($p < 10^{-4}$ in all cases). These findings suggest that Adaptive BAIC is less sensitive to changes in the proportion of the minority class instances.

6.5.2 Results on the normalised AUPC of the Adaptive BAIC metric

Having established the normalised AUPC of the Adaptive BAIC curve as a robust evaluation metric for non-stationary imbalanced streams, this section employs it to compare the proposed IPIP-CD against state-of-the-art ensemble drift adaptation methodologies.

Table 6.2 presents the AUPC achieved by the different models across the eight experimental data streams. It is important to note that these values serve as a relative comparison among models within the same stream. Because the Adaptive BAIC is not upper-bounded, values close to 0 indicate a failure to track the minority class under drift. Conversely, higher values indicate better adaptation, with the highest score in each column corresponding to the most robust model for that specific dataset. The empirical results demonstrate that the proposed IPIP-CD framework achieves the highest Adaptive BAIC

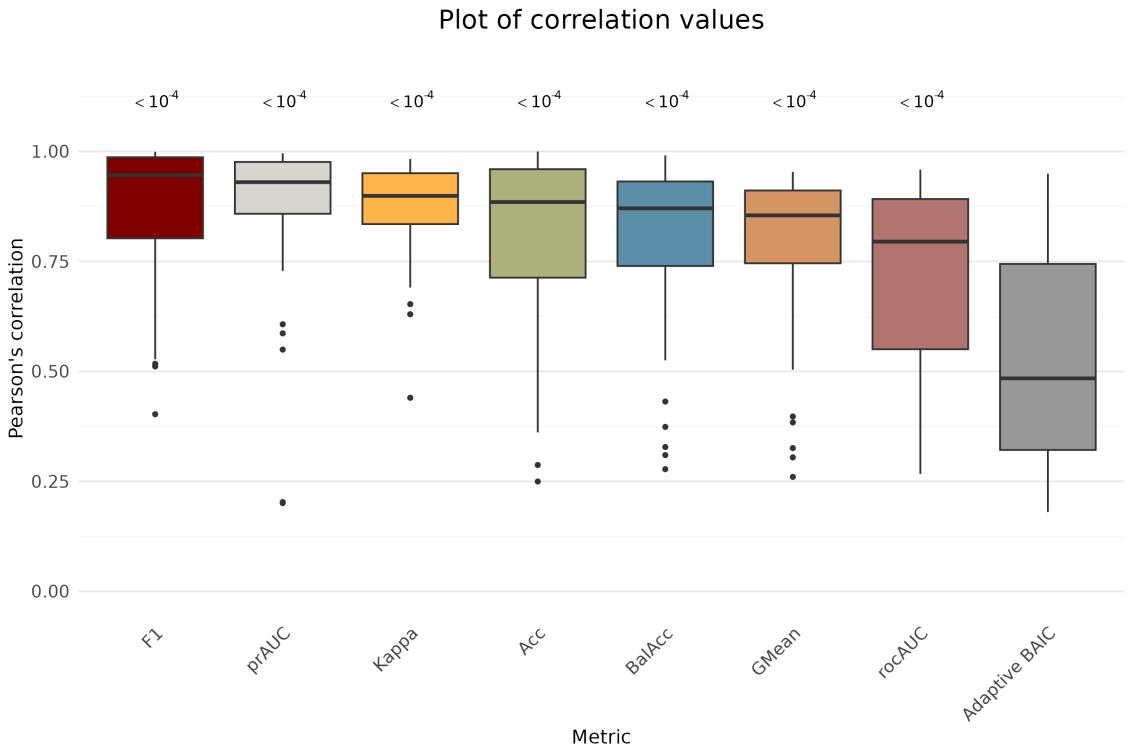


Figure 6.2: *Boxplots of Pearson's correlation values between the area under the temporal progression of different metrics and* $\mathbf{p_{min}}$ *across all streaming experiments. Metrics are ordered by median from highest to lowest. Above each boxplot is the p-value resulting from the Wilcoxon signed-rank test comparing that metric against the Adaptive BAIC correlation values.*

AUPC in six out of the eight evaluated datasets (Elec, Airlines, Hyperplane_5, Hyperplane_8, SEA, and SMS). IPIP-CD obtains more than the double of the AUPC compared to the second best model in several streams, such as Hyperplane_5 (0.406 vs. 0.146 for Learn++ CDS), Hyperplane_8 (0.316 vs. 0.063 for Learn++ CDS), and SMS (0.566 vs. 0.278 for SERA). It also obtains the highest AUPC value in Airlines (0.657 vs. 0.509 for UB) and SEA (0.582 vs. 0.330 for UB). This consistency across both real-world and synthetic streams suggests that the passive adaptation mechanism of IPIP-CD, which sequentially incorporates new balanced models and prunes obsolete ones, effectively tracks the evolving concept without being overwhelmed by the majority class.

Notably, in the Elec stream, IPIP-CD reaches a normalised Adaptive BAIC AUPC of 1.335, outperforming the UB model (normalised Adaptive BAIC AUPC of 0.728). Among the baseline ensemble methods, UB exhibits the strongest competitive performance, achieving the highest scores in the Hyperplane_2 (0.307 vs. 0.257 for IPIP-CD) and Weather (0.721 vs. 0.664 for IPIP-CD) datasets. Finally, the static Random Forest model obtains areas close to zero in several datasets, suggesting the necessity of concept drift adaptation mechanisms.

Method	Elec	Airlines	Hyperplane_2	Hyperplane_5	Hyperplane_8	SEA	SMS	Weather
AWE	0.006 (5º)	0.056 (4º)	0.000 (3º)	0.000 (4º)	0.000 (6º)	0.084 (6º)	0.000 (5º)	0.196 (3º)
UB	0.728 (2º)	0.509 (2º)	**0.307 (1º)**	0.141 (3º)	0.033 (3º)	0.330 (2º)	0.000 (6º)	**0.721 (1º)**
Learn++	0.190 (4º)	0.108 (3º)	0.000 (4º)	0.146 (2º)	0.063 (2º)	0.191 (3º)	0.077 (3º)	0.191 (4º)
SERA	0.489 (3º)	0.041 (5º)	0.000 (5º)	0.000 (5º)	0.017 (4º)	0.145 (5º)	0.278 (2º)	0.068 (5º)
Ranger	0.000 (6º)	0.002 (6º)	0.000 (6º)	0.000 (6º)	0.000 (5º)	0.190 (4º)	0.004 (4º)	0.000 (6º)
IPIP-CD	**1.335 (1º)**	**0.657 (1º)**	0.257 (2º)	**0.406 (1º)**	**0.316 (1º)**	**0.582 (1º)**	**0.566 (1º)**	0.664 (2º)

Table 6.2: *Global performance comparison using the normalised AUPC of the Adaptive BAIC progression curve. Values are reported with three decimal places followed by their relative rank within each dataset (1º to 6º). The highest performing model for each stream is highlighted in bold.*

6.6 Concluding the chapter

In this chapter, a comprehensive approach to addressing classification in dynamic data streams under severe class imbalance has been proposed [8].

Firstly, empirical investigations revealed that the combination of concept drift and class imbalance not only degrades predictive performance but also distorts traditional evaluation metrics, leading to inconsistent conclusions in model selection. To address these limitations, the Adaptive BAIC method was introduced as an evolution of the static metric, designed to provide unbiased evaluation in non-stationary environments. This metric calculates the normalised AUPC over time using a one-step lag [7]. It has been statistically demonstrated ($p < 10^{-4}$) that Adaptive BAIC successfully reduces the bias toward p_{min} (achieving a median correlation of 0.48) compared to traditional metrics like F1-score or prAUC, proving to be a robust measure for non-stationary environments.

Secondly, IPIP-CD was proposed as a natural extension of the static framework tailored for streaming environments. IPIP-CD adopts a passive update paradigm based on data chunks [325]. It does not require explicit drift detectors; instead, it evolves naturally by training new balanced models and pruning obsolete ones using the Adaptive BAIC evaluation.

Empirical results demonstrate that IPIP-CD outperformed state-of-the-art algorithms such as UB, AWE, Learn++ CDS, and SERA, achieving the highest Adaptive BAIC AUPC in six of the eight evaluated datasets (Elec, Airlines, Hyperplane_5, Hyperplane_8, SEA, and SMS). Notably, it doubled the performance of the second-best model in three datasets; SMS, Hyperplane_5 and Hyperplane_8, while remaining highly competitive by obtaining second place in the remaining two datasets led by UB. Note that both IPIP-CD and UB methods build ensembles on different data subsets to passively adapt to concept drift, and neither of them generates synthetic instances. However, they differ in how subsets are created. UB uses random resampling to uncorrelate errors across models and improve general robustness. In contrast, IPIP-CD enforces balanced partitions to guarantee that all minority class instances are used. As a result, IPIP-CD performs better when severe class imbalance is the main challenge. On the other hand, UB can achieve slightly better results when the specific type of drift favors random diversity over strict class balancing. Finally, the static Random Forest baseline completely failed to maintain

predictive performance, confirming that these adaptation strategies are necessary.

From a practical perspective, this strategy of adding and pruning submodels avoids the complete retraining of the system, making it computationally viable and ideal for continuous MLOps pipelines. Finally, the tools for implementing these methods have been integrated into the FILM-CD R package (https://github.com/antoniogt/FILMDrift), enabling model construction for all the baselines utilised in this chapter and for IPIP-CD and the Adaptive BAIC calculation for any dataset.

7

Explanation biases in Imbalanced Problems

This chapter transitions from the predictive optimization of imbalanced data streams, as addressed by the IPIP-CD framework in the preceding chapter, to the domain of Explainable Artificial Intelligence. In medical contexts, such as COVID-19 diagnosis and prognosis, achieving robust predictive performance must be accompanied by transparent decision-making to ensure clinical trust.

The research presented in this chapter was carried out during an academic stay at the Jagiellonian University in Kraków, Poland, in collaboration with the Group for Engineering of Intelligent Systems Technologies (GEIST), a research group working on topics such as explainable AI and machine learning. During this stay, an important limitation of current interpretability methods was identified: in imbalanced classification problems, class proportions can strongly affect feature importance explanations such as SHAP values. Motivated by this issue, this chapter studies how data imbalance influences SHAP explanations in stationary scenarios, providing a baseline before extending these analyses to non-stationary data streams.

7.1 Introduction

Interpretable ML has become a pivotal area of research, particularly in critical application domains such as healthcare, finance, and criminal justice [326–328]. SHAP values [195] are widely used to interpret and understand model decisions as mentioned in Chapter 3. SHAP values are fair estimates of each learning feature's contribution to predictions. The SHAP approach is widely applicable across different prediction models. But most importantly, it can be applied to study individual predictions [329]. Feature attribution refers to the process of assigning a contribution value to each feature for a specific prediction

made by an ML model. This is a critical task within SHAP-based interpretation analysis. Unfortunately, SHAP values can suffer bias due to imbalanced class distribution, feature collinearity [330] or the model-specific architecture [331] in classification problems [332, 333], potentially leading to inaccurate interpretations. Class imbalance is a significant yet frequently disregarded source of such bias, a prevalent issue in real-world datasets where one class is significantly overrepresented. This imbalance can affect the calculated feature attribution, i.e. the set of mentioned SHAP values and subsequently impact decision-making in sensitive domains [334].

As studied in Chapter 5, class imbalance is a well-known challenge in binary classification, where the majority class dominates the dataset, leading to biased model performance and misleading evaluation metrics (See Section 5.5.2 of Chapter 5). While much research has focused on mitigating this imbalance, including the IPIP algorithm included in this thesis (See Section 5.3.3), and through techniques such as resampling [335] or cost-sensitive learning [336], less attention has been given to its impact on post-hoc interpretability methods like SHAP. This method can be particularly vulnerable to imbalanced class distributions, where SHAP values tend to be artificially inflated toward the majority class. SHAP values should provide unbiased estimates of feature importance. Otherwise, critical predictors for the minority class may be ignored. Apart from this potentially harmful bias, model-specific SHAP implementations such as TreeSHAP for decision trees and their ensembles, DeepSHAP for neural networks or LinearSHAP for linear models, make specific assumptions about the underlying predictive model. Therefore, it is likely that the same dataset may lead to qualitatively different SHAP estimates depending on the chosen model family and SHAP implementation variant.

In this chapter, we examine the impact of different levels of imbalanced data on the SHAP values by observing their variability across five different datasets, when altering the target variable's class proportions in controlled experiments. We aim to detect potential biases in SHAP values across different learning problems and imbalance levels for a more accurate use of the SHAP method. Our approach is purely empirical for practical reasons. Because SHAP generation implementations are highly model-specific, as mentioned before, and a purely theoretical approach would likely remain superficial and struggle to cover the specific mathematical assumptions of all algorithmic families. An empirical approach, however, allows us to treat the underlying models as black boxes, ensuring broad, general applicability across diverse machine learning paradigms. It involves the generation of a number of resamples from each dataset under study with an IR of interest. We denote each IR with $n : m$, being n and m the number of samples in the minority and majority classes, respectively. We generate five resamples for each dataset using five IR values. Then we generate overfitted machine learning models for four model families, i.e., Random Forest (RF), XGBoost, multilayer perceptron (MLP) and a polynomial logistic regression (PLR). The models are obtained by performing an automated hyperparameter optimization guided by robust estimates of balanced accuracy across all resamples. This ensures that all models fit the training data closely to serve as a proxy

for a comprehensive explanation of the training examples. Subsequently, SHAP values are obtained from all the models to obtain estimates of variability within each IR. We then use these estimates to identify possible biases in SHAP values across IRs.

7.2 Related work on the reliability of SHAP-based post-hoc explanations

Existing literature that connects class imbalance explicitly with the reliability of post-hoc explanations is still limited, but it already suggests that SHAP outputs can be noticeably affected by skewed class distributions. In the clinical domain, [337] reported that imbalanced datasets can introduce artefacts in SHAP visualisations and can alter feature-importance patterns, and proposed balancing strategies at the explanation stage to obtain more stable outputs. Similarly, [333] analysed interpretable ML for credit scoring under multiple imbalance ratios and showed that increasing imbalance may yield less stable explanations, with larger variability and inconsistent feature rankings across resamples. These studies motivate the need for systematic experimental designs that quantify how SHAP changes across imbalance conditions. In this chapter of the thesis we are motivated by the requirements of accuracy in the medical domain. Therefore, we extend this emerging line of research with a controlled, multi-dataset and multi-model study that characterises how SHAP distributions shift across imbalance ratios.

Beyond the direct observation of explanation instability discussed above, further research explores its underlying mechanisms. Specifically, several works highlight that imbalance intrinsically modifies the learned decision function, which in turn indirectly affects any attribution method applied post-hoc. For instance, [338] studied instance-dependent and cost-sensitive classifiers, where examples are weighted differently and the effective class distribution becomes skewed. Their results indicate that strategies intended to improve predictive performance under rare-event settings may alter the stability of SHAP and LIME [339] explanations, where LIME (Local Interpretable Model-agnostic Explanations) is an alternative post-hoc interpretability method based on local surrogate models (i.e., models that are trained to approximate the predictions of the underlying black box model), unlike the game-theoretic formulation used by SHAP. They also propose quantitative stability measures to capture such effects. While these contributions do not focus on SHAP distributional behaviour across controlled imbalance ratios, they support the idea that changes in model behaviour induced by imbalance are a plausible source of explanation shifts. The contribution of this chapter is to make this link explicit by coupling SHAP distribution analysis with model-stability indicators, enabling us to relate changes in attributions to changes in model complexity and parameter stability as imbalance varies.

Finally, several works propose general tools to assess explanation stability and to clarify

when changes in feature importance are meaningful. [340] introduced a statistical framework with hypothesis tests to verify the stability and significance of top-k feature rankings, providing a rigorous way to distinguish systematic shifts from random fluctuations due to resampling. Complementary theoretical critiques also show that Shapley-based attributions can be unreliable in specific data regimes. For example, [341] demonstrated that SHAP can produce misleading attributions in settings with strong feature dependence or misspecified data-generating mechanisms, where the assigned importance may fail to reflect true causal relevance. In a related direction, [342] discussed how naive aggregation of SHAP values may distort global importance estimates and proposed refinements to obtain more faithful global summaries. Together, these works highlight that explanation robustness must be evaluated rather than assumed, and that dataset properties can be a major driver of instability. In this chapter, we contribute a practical evaluation pipeline that quantifies how imbalance affects SHAP at both the distribution and ranking levels, and we release the code and results to facilitate reproducibility and future extensions.

7.3 Background

In this section, we introduce SHAP values and how can SHAP values be obtained when training RF, XGBoost, MLP and PLR models. This serves to clarify the approach adopted in the study of SHAP biases in imbalanced problems.

7.3.1 SHapley Additive exPlanations (SHAP) values

The SHAP method is based on estimates called Shapley values. The mechanism to produce such estimates uses coalitional game theory [343], where the features of a target training dataset are treated as 'players' in a coalition. The goal of SHAP is to explain each individual prediction by producing estimates of how much each feature contributes to the output. This is achieved by comparing the model's predictions when different resamples of features are used, simulating a cooperative game in which features collaborate to make a prediction. Given a model f and an instance x, the Shapley value for the i-th feature is defined as:

$$\phi_i = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|S|!(|\mathcal{N}| - |S| - 1)!}{|\mathcal{N}|!} (f(S \cup \{i\}) - f(S))$$

where $\mathcal{N}$ is the set of all input features, $f(X)$ is the ML model created by using only the X set of variables, S means all the resamples of $\mathcal{N}$ that do not include the i-th feature, and the term $f(S \cup \{i\}) - f(S)$ captures the marginal contribution of the i-th variable. SHAP is represented as an additive feature attribution model, denoted by g, which is defined

as $g(z') = \phi_0 + \sum_{j=1}^{|\mathcal{N}|} \phi_j z'_j$ where g is the function that generates the SHAP value, and z' is the coalition vector that indicates the presence ($z'_j = 1$) or absence ($z'_j = 0$) of each feature. The Shapley value for feature j is represented by ϕ_j. This formulation enables SHAP to express feature importance as a linear model of coalitions.

Exact computation of Shapley values is usually unfeasible for real-world models, and different SHAP implementations rely on model-specific approximations [195]. From a software and practical standpoint, we rely on the Python `shap` library (version 0.46.0), which natively integrates highly optimised, model-specific explainers that make this extensive empirical evaluation computationally feasible. Specifically, TreeSHAP exploits the structure of decision trees and ensembles to compute exact feature importances. Because Shapley values satisfy the additivity property, TreeSHAP is perfectly applicable to both independent tree ensembles (like Random Forests) and sequential ones (like XGBoost). The explainer operates over the final structure of the trained trees and aggregates their individual contributions, regardless of the bagging or boosting training procedure [344]. DeepSHAP combines a deep learning important features algorithm [345] with Shapley value concepts to approximate attributions for deep neural networks. LinearSHAP provides a formula for linear models by leveraging the linearity of the prediction function. For the experiments detailed in this chapter, we use TreeSHAP for RF and XGBoost, DeepSHAP for the MLP, and LinearSHAP for the PLR. By utilising these specific, state-of-the-art implementations (TreeSHAP for RF and XGBoost, DeepSHAP for the TensorFlow MLP, and LinearSHAP for the PLR), we ensure that our analysis of structural biases relies on the exact same tools and algorithmic approximations currently deployed in real-world scenarios.

7.4 Experimental set-up

In this section we present the overall approach we have followed for empirically studying the effect of class imbalance on SHAP-based explanations in binary classification problems. As illustrated in Fig. 7.1, the target dataset is first subsampled at several predefined imbalance ratios (IRs) by means of an imbalance-aware stratified sampling procedure (see Algorithm 4), which generates multiple resamples for every ratio. Note that we assume a target dataset with an initial class labels distribution from which all resulting resamples show an increased imbalance. For each resample, all four model families considered in this study (RF, XGBoost, MLP and PLR) are trained and subsequently explained using SHAP values. The resulting SHAP values are then analysed in terms of their distributions. The procedures to perform all the analyses are described in the following subsections.

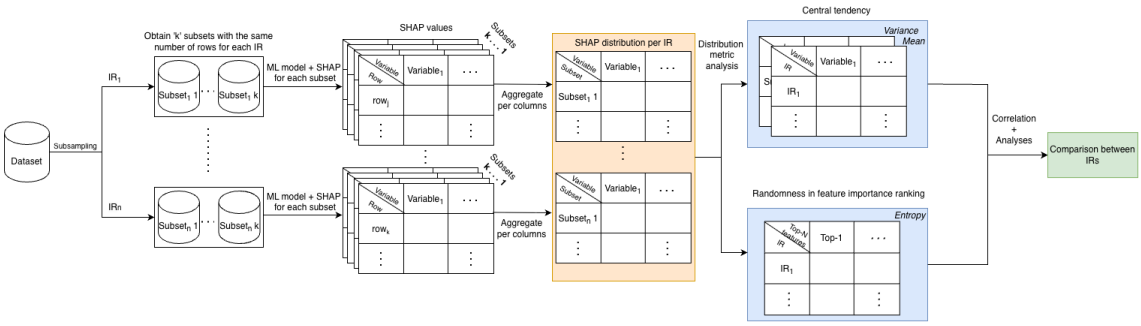


Figure 7.1: *Overview of the experimental framework that has been utilised in order to analyse the effect of class imbalance on SHAP values for each individual dataset.*

7.4.1 The four families used in the context of imbalanced problems

In all four model families used, we employ a basic treatment of imbalance, tailored for the model at hand, as follows.

RF [51] is an ensemble learning method that constructs multiple decision trees from bootstrapped resamples of training data and uses random resamples of features in each tree split decision. The final prediction is obtained by majority voting on the classification tasks or by averaging in regression tasks. RF training may lead to biased predictions favoring the majority class when dealing with imbalanced datasets. We mitigate this bias by differently weighting the available classes [346]. Rather than treating each instance equally, the loss function that guides model building uses instance-specific weights based on their class proportion to avoid prioritizing overall accuracy at the expense of the minority class.

XGBoost [54] is a gradient boosting ensemble of decision trees that builds trees sequentially, each new tree focusing on correcting the errors made by the previous ones. It has proven highly effective on tabular data and is widely used in imbalanced classification problems [347]. In our experiments we did not apply any imbalance specific regularisation techniques to XGBoost (such as class weighting, resampling strategies or modified loss functions), as even with the standard binary logistic objective and sufficiently large ensemble sizes, XGBoost already achieved high balanced accuracy across all IRs. This behaviour is expected, as gradient boosting iteratively focuses on misclassified instances, which in imbalanced problems are disproportionately drawn from the minority class, and can therefore deal with the lack of explicit class reweighting [348].

The MLP used in this chapter is a simple feed forward and fully connected neural network with two hidden layers and a non linear activation function [349]. It is trained with a standard cross entropy loss for binary classification. To mitigate the effect of class imbalance in the backpropagation algorithm that trains model parameters, we introduce class weights in the loss function so that errors on the minority class instances have a greater impact on the parameter updates than errors on the majority class instances.

Finally, the PLR model is obtained by expanding the original features with polynomial combinations up to a fixed degree and then fitting a regularised logistic regression on the resulting model [350]. This final model captures nonlinear relationships while retaining a linear parametrisation in the feature space. As in the case of RF and the MLP, we use a class-weighted logistic loss. In addition, rather than relying on the default probability classification threshold of 0.5, the decision threshold can be selected such that it maximises balanced accuracy on the corresponding data.

7.4.2 Sampling Strategy for IR Evaluation

In this section we present our Imbalance-Aware Stratified Sampling (IASS), through the pseudo-code of algorithm 4 designed to obtain data resamples with the desired IR while maintaining a consistent sample size across IR values. For each IR, a predefined proportion of instances from each class is used to randomly select examples without replacement. This process is repeated multiple times to generate approximately identically distributed resamples for each IR to reduce the variance across resamples. Additionally, we generate as many resamples as necessary to ensure that each unique instance appears in at least two resamples, enhancing the robustness of the comparison between IR conditions. It is important to note that the structure of Algorithm 4 guarantees that the exact same number of resamples is generated for every IR. The final list of resamples is used for model training and their corresponding SHAP values enable SHAP analysis in downstream tasks (See Fig. 7.1). Note that the same list of resamples is used to train all four ML methods so that comparisons across model families are based on identical data.

7.4.3 SHAP values analysis through basic statistical parameters

To characterise how feature attributions change with IRs across models and datasets we rely on SHAP values computed for every trained model and resample. For each combination of dataset, IR, IR resample and ML algorithm, we train the corresponding classifier as described in Section 7.4.1 and compute SHAP values for all dataset instances. We use the implementations provided by the `shap` library that are appropriate for each model family. Because SHAP explanations in binary classification explain the model output for one class at a time, the same instance can have different SHAP vectors depending on whether the explained output is the positive or the negative class. To keep all SHAP comparable when aggregating results across instances, we retain, for each instance, only the SHAP values computed for its true class label. In this way, SHAP magnitudes always refer to the feature importance of the correct outcome and are not mixed across opposite class outputs.

Let D denote a dataset, m a model family, r an imbalance ratio, S the list of resamples obtained with the IASS algorithm and j a feature. Let also $D_r \subseteq S$ be all resamples in S

Algorithm 4 **I**mbalance-**A**ware **S**tratified **S**ampling (IASS)

Input: Dataset $\mathcal{D}$, target variable y, imbalance ratios IRs
Output: List of resamples $\mathcal{S}$ for each IR

Split $\mathcal{D}$ into $\mathcal{D}_{\text{pos}}$ (minority class) and $\mathcal{D}_{\text{neg}}$ (majority class)
Compute instances $N_{\text{pos}} = |\mathcal{D}_{\text{pos}}|$ and $N_{\text{neg}} = |\mathcal{D}_{\text{neg}}|$
Compute set of minority proportions p_{IRs} associated to IRs
Initialise $k = 1$
$N_{\max} = \min_{p_{IR} \in p_{IRs}} \left(\frac{N_{\text{pos}}}{p_{IR}}, \frac{N_{\text{neg}}}{1-p_{IR}} \right)$
while not all instances appear at least twice across resamples for any $IR \in IRs$ **do**
  $\mathcal{S}$ = empty list
  for each $IR \in IRs$ **do**
    $\mathcal{S}_{IR}$ = empty list
    for $i = 1$ to k **do**
      Sample $N_{\text{pos}}^{(i)} = \lfloor N_{\max} \cdot p_{IR} \rfloor$
      Sample $N_{\text{neg}}^{(i)} = N_{\max} - N_{\text{pos}}^{(i)}$
      $\mathcal{S}_{IR}^{(i)} = N_{\text{pos}}^{(i)} \cup N_{\text{neg}}^{(i)}$
      $\mathcal{S}_{IR} = \mathcal{S}_{IR} \cup \mathcal{S}_{IR}^{(i)}$
    end for
    $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_{IR}$
  end for
  $k = k + 1$
end while
Return $\mathcal{S}$

as calculated for an IR of r.

For each combination (D, m, r, j) we collect all absolute SHAP values, obtaining an empirical distribution

$$S_{D,m,r,j} \;=\; \left\{ |\phi_{i,j}^{(D,m,r)}| \;:\; i \text{ indexes instances in } D_r \right\},$$

where $\phi_{i,j}^{(D,m,r)}$ denotes the SHAP value assigned to feature j for instance i by model m under imbalanced ratio r. These sets $S_{D,m,r,j}$ represent the distribution of feature importance values for a given dataset, model and imbalance ratio. We denote with μ_j the arithmetic mean of SHAP values for feature j in all combinations of datasets, IRs and models. Analogously for the variance σ_j^2. If, for example, we need to specify those statistics, for a specific dataset D, r value and method m, we denote mean and variance with $\mu_{D,m,r,j}$ and $\sigma_{D,m,r,j}^2$. If we wanted to aggregate results for the top k features as signaled by the SHAP analysis, we simply obtain arithmetic means of the corresponding $\mu_{D,m,r,j}$ and $\sigma_{D,m,r,j}^2$ for all j in the top k features and denote them with $\mu_{D,m,r}$ and $\sigma_{D,m,r}^2$, respectively.

To detect possible associations between levels of class imbalance and μ and σ^2 values for a given dataset D and ML model m, we use Spearman correlation. So we calculate $\rho_{\mu,D,m}$ as the Spearman correlation between IR values used and the corresponding $\mu_{D,m,r}$ values. Analogously for $\rho_{\sigma^2,D,m}$. In the Section 7.5.3 we report and discuss these correlations

together with the patterns observed in the violin plots.

7.4.4 Model stability indicators

Apart from those statistics, of global applicability, we consider model-specific biases towards IR values that influence model complexity as proxies for estimates of stability. An algorithm would be considered stable if the models it creates do not significantly vary when the IR indeed varies.

7.4.4.1 Predictive Complexity Score (PCS)

For tree ensembles (RF and XGBoost) we measure such stability with what we call the Predictive Complexity Score (PCS). Given a model m, and an IR value r, the PCS is the average depth of the leaves reached by examples from the minority class across all trees. The PCS reflects the effective complexity of the decision paths taken when predicting the minority class, with higher values indicating more complex decision boundaries in the feature space leading to less generalization capability and worse behaviour in predicting the minority class. Note that focusing on the minority class to provide estimates of the PCS is consistent with standard practice in imbalanced evaluation, where metrics such as recall or F_1-score explicitly prioritise performance on the minority class [3, 4].

Let M denote a tree ensemble with T trees, trained on a dataset with a minority class containing N_+ instances. In our methodological framework, PCS is computed exclusively on the minority-class training examples, so N_+ denotes the number of minority instances and all frequencies below are defined with respect to this resample. We define a tree-level complexity measure $C(t)$ for each tree $t = 1, \ldots, T$ and then average these values over all trees to obtain the PCS of the ensemble:

$$\mathrm{PCS}(M) \;=\; \frac{1}{T}\sum_{t=1}^{T} C(t). \tag{7.1}$$

For a given tree t, let $\mathcal{L}(t)$ denote the set of its leaves. For each leaf $\ell \in L(t)$, we denote by $d_{t,\ell}$ the depth of leaf ℓ (with the root at depth 0) and by $n_{t,\ell}$ the number of *minority-class* training instances that reach leaf ℓ. Let $N_t = \sum_{\ell\in\mathcal{L}(t)} n_{t,\ell}$ be the total number of minority instances used to compute these leaf frequencies (typically $N_t = N_+$ if all minority training instances are considered). We then define

$$C(t) = \sum_{\ell\in L(t)} p_{t,\ell}\, d_{t,\ell}, \qquad p_{t,\ell} = \frac{n_{t,\ell}}{N_t}, \tag{7.2}$$

so that $C(t)$ is the expected depth of the leaf reached by a randomly chosen training in-

Algorithm 5 Computation of PCS for a tree ensemble

Input: Trained ensemble M with trees $\{t_1, \ldots, t_T\}$, minority-class training data $\{x_1, \ldots, x_{N_+}\}$
Output: $\mathrm{PCS}(M)$

```
1: Initialise PCS ← 0
2: for k = 1 to T do
3:    Let t ← t_k
4:    For each leaf ℓ in L(t), set n_{t,ℓ} ← 0
5:    for i = 1 to N_+ do
6:       Determine the leaf ℓ reached by x_i in tree t
7:       Increment n_{t,ℓ} ← n_{t,ℓ} + 1
8:    end for
9:    Compute N_t ← Σ_{ℓ∈L(t)} n_{t,ℓ}
10:   Initialise C(t) ← 0
11:   for each leaf ℓ ∈ L(t) do
12:      Compute p_{t,ℓ} ← n_{t,ℓ}/N_t
13:      Let d_{t,ℓ} be the depth of leaf ℓ
14:      Update C(t) ← C(t) + p_{t,ℓ} · d_{t,ℓ}
15:   end for
16:   Update PCS ← PCS + C(t)
17: end for
18: PCS(M) ← PCS/T
19: return PCS(M)
```

stance in tree t, and $\mathrm{PCS}(M)$ is the average of these expected depths over all trees in the ensemble. Intuitively, large PCS values correspond to ensembles that rely on deeper and more complex decision paths, which we interpret as having more complex and potentially less stable decision boundaries. PCS can thus be interpreted as an average path-length measure focused on the minority class, in line with complexity methods based on expected leaf depth that are commonly used for decision trees and ensembles [351, 352].

Algorithm 5 summarises the computation of PCS in pseudocode. Lines 4-8 compute the leaf frequencies for each tree using the training data, lines 11-15 compute the tree-level complexity $C(t)$ by aggregating depths weighted by their empirical frequencies, and lines 16-18 average these values to obtain the final PCS of the ensemble.

For each dataset D, model family m and imbalance ratio r, we obtain multiple resamples generated at ratio r and train one ensemble $M_{D,m,r}^{(s)}$ per resample $s = 1, \ldots, S_{D,r}$. We then compute $\mathrm{PCS}(M_{D,m,r}^{(s)})$ for each resample and aggregate these scores by averaging over resamples:

$$\mathrm{PCS}_{D,m,r} = \frac{1}{S_{D,r}} \sum_{s=1}^{S_{D,r}} \mathrm{PCS}\big(M_{D,m,r}^{(s)}\big), \tag{7.3}$$

which yields a single PCS value per dataset, model family and imbalance ratio.

7.4.4.2 Parameter-variance indicators

For MLPs and PLR models we use as estimates of stability the empirical variance of the model parameters (VMP) across all resamples for a given imbalance ratio. Larger parameter variances would suggest more chance of overfitting to the particular resamples, therefore reducing stability.

For the MLP and the PLR model we define parameter-variance indicators that capture how strongly the learned parameters change across resamples at a given imbalance ratio. Let $M^{(s)}_{D,m,r}$ denote the model of family $m \in \{\text{MLP}, \text{Poly}\}$ trained on resample s for dataset D and imbalance ratio r. We extract the corresponding parameter vector $\theta^{(s)}_{D,m,r} \in \mathbb{R}^P$ by concatenating all weights (for the MLP) or all coefficients (for the PLR model) into a single vector of length P.

Given the collection $\{\theta^{(s)}_{D,m,r} : s = 1, \ldots, S_{D,r}\}$, we compute the empirical variance of each parameter across resamples and then average these variances to obtain a single indicator:

$$v_{D,m,r} \;=\; \frac{1}{P} \sum_{p=1}^{P} \mathrm{Var}\Big(\theta^{(1)}_{D,m,r,p}, \ldots, \theta^{(S_{D,r})}_{D,m,r,p}\Big), \tag{7.4}$$

where $\theta^{(s)}_{D,m,r,p}$ denotes the p-th component of $\theta^{(s)}_{D,m,r}$ and $\mathrm{Var}(\cdot)$ is the empirical variance across resamples. Large values of $v_{D,m,r}$ indicate that the parameters of model m must change substantially from one resample to another at imbalance ratio r, which we interpret as a sign of reduced stability.

Algorithm 6 shows the computation of the parameter-variance indicator in pseudocode. Lines 1-3 flatten the parameters of each model into a vector, and lines 5-9 compute and aggregate the variances across resamples.

The use of VMP as a stability metric is justified by the underlying mathematical structure of both PLR and MLP models. Since both model families are fundamentally defined by matrices of real-valued weights, the magnitude of these weights directly dictates the contribution of each input feature to the final prediction. In linear models, this relationship is explicit in the coefficients, while in MLPs, it is captured through the composition of weight matrices across layers.

It is worth noting that flattening the weight matrices of the MLP into a single linear vector $\theta \in \mathbb{R}^P$ is a mathematically sound operation for the purpose of calculating the VMP. Since the flattening process follows a deterministic and consistent order across all experimental runs, each index p consistently tracks the same synaptic connection across different model instances. While this global aggregation does not account for the hierarchical structure of the network (i.e., it treats weights from different layers with equal importance), it provides a robust scalar indicator of the overall parameter variance.

Algorithm 6 Computation of parameter-variance indicator

Input: Models $\{M^{(1)}, \ldots, M^{(S)}\}$ of family m trained on S resamples at a given imbalance ratio
Output: Parameter-variance indicator v

for $s = 1$ to S **do**
  Extract and flatten all parameters of $M^{(s)}$ into a vector $\theta^{(s)} \in \mathbb{R}^P$
end for
Initialise $v \leftarrow 0$
for $p = 1$ to P **do**
  Let $\theta_p^{(1)}, \ldots, \theta_p^{(S)}$ be the p-th components of $\theta^{(1)}, \ldots, \theta^{(S)}$
  Compute the empirical variance $\sigma_p^2 \leftarrow \mathrm{Var}(\theta_p^{(1)}, \ldots, \theta_p^{(S)})$
  Update $v \leftarrow v + \sigma_p^2$
end for
$v \leftarrow v/P$
return v

We later use both PCS and VMP with μ, σ^2 and their ρ_μ and ρ_{σ^2} to investigate whether changes in explanations across imbalance ratios are associated with systematic changes in model complexity or parameter stability, and whether such relationships differ between model families.

7.4.5 SHAP consistency in producing feature rankings

The focus of this chapter is quantifying possible SHAP biases towards different IRs. Therefore, we may ask ourselves whether feature ranks as produced by SHAP are stable across different IRs. Since inconsistent feature rankings for the same dataset and different resamples may compromise model interpretability, especially under class imbalance, it is essential to assess feature ranking robustness across resamples.

Given a dataset D and an ML model type M, we create feature ranks for D and M ordering the features by their mean absolute SHAP value. To analyse the stability of feature rankings for D and M across different resamples for each IR value, we focus on the top-N features. All ranks of length N, across resamples for the same IR are aggregated to form an empirical probability distribution over feature combinations. We then compute the Shannon entropy of the empirical probability estimates of the top-N feature sets. Low entropy indicates that the same features consistently appear in the top positions across all resamples, reflecting high stability in the ranking of model explanations. Conversely, high entropy suggests that the model assigns similar SHAP importance to several alternative features, leading to less stable and potentially less reliable explanations.

By construction, entropy increases monotonically with N, as more features introduce more possible combinations. To obtain robust estimates of ranking stability, we compute entropy for each IR using a non-parametric bootstrap procedure. For a given IR and a chosen value of N, resamples with replacement are used, and the entropy of the resulting

empirical distribution is recomputed. Repeating this process 10000 times yields a bootstrap distribution of entropy values, from which we derive a bootstrapped mean and a bias-corrected 95% confidence interval. This procedure provides a reliable measure of uncertainty and allows us to compare entropy levels across imbalance conditions.

Finally, entropy curves are constructed by plotting the bootstrapped mean entropy as a function of N for each IR, separately for each model. Confidence intervals are displayed as vertical error bars. This representation makes it possible to visualise not only how stable the top-N features are within each imbalance condition, but also how this stability changes as datasets become more or less balanced. These entropy-based analyses complement the SHAP magnitude and variance results, offering an additional layer of interpretability focused on the consistency of feature importance across IR changes.

7.5 Results

7.5.1 Datasets

The controlled experimentation to detect biases in SHAP values is conducted across five datasets encompassing diverse initial class imbalance ratios, sample sizes, and feature dimensions. The selected benchmarks involve four publicly accessible domains (Diabetes, Mammography, Forest, and Satimage) and an expanded cohort of the SMS COVID-19 clinical registry. A comprehensive description of these datasets, their original imbalance, and the number of resamples engineered for the experiments are meticulously detailed in Chapter 4 (Section 4.2.1). For each benchmark, multiple subsets are consistently produced employing the Imbalance-Aware Stratified Sampling (IASS) algorithm to enforce the necessary target imbalance ratios. For instance, the SMS dataset needed 713 resamples, the Diabetes dataset needed 160 subsets, the Mammography dataset generated 148 resamples, Forest needed 106 subsets, and Satimage required 102 resamples.

7.5.2 Implementation details

To create all data resamples at variable IRs we artificially generate resamples with a target minority class proportion. We use as targets 2%, 5%, 10%, 15% and 20%, which correspond to target IRs of 1:49, 1:19, 1:9, 3:17 and 1:4, respectively. Resample generation with IASS is implemented in R (version 4.2.1), while model training and explanation are carried out in Python using `scikit-learn` [353] (version 1.5.1) for all models except for the MLP which is implemented with TensorFlow.

For each resample created (see Algorithm 4 for details), we train all four model families described in Section 7.4.1. Unless otherwise stated, the grids used for hyperparameter-based model optimization are kept fixed across all target IRs and their corresponding

resamples. RFs, MLPs and polynomial models are trained with class-weighted losses to balance error based on the IRs. XGBoost is employed with its regular logistic loss without any explicit imbalance-specific regularisation. This design choice is intended to reflect a common practical scenario in which limited prior knowledge about the data distribution is available, or where the data distribution may evolve over time. In such settings, the effective imbalance ratio may be unknown, unexpected, or subject to change, and model designers may therefore rely on default training procedures without introducing additional imbalance-mitigation mechanisms. The PLR model is regularised with an L2 penalty applied to the polynomially expanded features. And its probability threshold for classification is automatically chosen to maximize balanced accuracy. SHAP values are computed for each trained model and resample following the procedure described in Section 7.4.3. Random seeds are used to ensure reproducibility of the sampling and training processes.

ML models are evaluated using balanced accuracy computed on the same resample used for training (see Table 7.1). This choice is intentional, as our goal is not to benchmark generalisation performance under imbalance, but to isolate how class distribution affects SHAP values when the underlying model has adequately captured the training signal of each resample. If a model underfits a resample, SHAP values may primarily reflect estimation noise rather than stable data structure, which would confound the feature importance analysis. Therefore, we ensure high balanced accuracy for each IR before analysing SHAP distributions, so that differences in explanations across imbalance ratios are not driven by low model performance. All models achieved a mean balanced accuracy above 0.91 across all imbalance ratios.

Dataset	Model	Balanced Accuracy by IR (BalAcc [$CI_{95\%}$])				
		1:49	1:19	1:9	3:17	1:4
SMS	MLP	0.96 [0.95, 0.96]	0.94 [0.94, 0.94]	0.93 [0.93, 0.93]	0.92 [0.92, 0.92]	0.92 [0.92, 0.92]
	PLR	0.97 [0.97, 0.97]	0.95 [0.95, 0.95]	0.94 [0.94, 0.94]	0.93 [0.93, 0.93]	0.93 [0.93, 0.93]
	RF	0.99 [0.99, 0.99]	0.98 [0.98, 0.98]	0.97 [0.97, 0.97]	0.97 [0.97, 0.97]	0.97 [0.97, 0.97]
	XGBoost	0.93 [0.93, 0.93]	0.94 [0.94, 0.94]	0.94 [0.94, 0.94]	0.95 [0.95, 0.95]	0.95 [0.95, 0.95]
Diabetes	MLP	0.99 [0.98, 0.99]	0.98 [0.98, 0.99]	0.98 [0.97, 0.98]	0.97 [0.97, 0.98]	0.96 [0.96, 0.97]
	PLR	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	RF	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	XGBoost	0.98 [0.98, 0.99]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Mammography	MLP	0.94 [0.93, 0.95]	0.92 [0.91, 0.92]	0.93 [0.92, 0.93]	0.93 [0.93, 0.93]	0.93 [0.93, 0.94]
	PLR	0.99 [0.98, 0.99]	0.98 [0.98, 0.99]	0.98 [0.98, 0.98]	0.97 [0.97, 0.97]	0.96 [0.96, 0.96]
	RF	0.99 [0.98, 0.99]	0.99 [0.98, 0.99]	0.99 [0.99, 0.99]	0.99 [0.99, 0.99]	0.99 [0.99, 0.99]
	XGBoost	0.97 [0.96, 0.97]	0.98 [0.98, 0.98]	0.98 [0.98, 0.98]	0.98 [0.98, 0.98]	0.98 [0.98, 0.98]
Forest	MLP	0.91 [0.88, 0.94]	0.91 [0.89, 0.93]	0.91 [0.89, 0.93]	0.94 [0.91, 0.97]	0.95 [0.93, 0.97]
	PLR	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	RF	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	XGBoost	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Satimage	MLP	0.96 [0.95, 0.97]	0.95 [0.94, 0.95]	0.94 [0.94, 0.95]	0.95 [0.95, 0.95]	0.95 [0.95, 0.96]
	PLR	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	RF	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	XGBoost	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]

Table 7.1: *Balanced Accuracy (BalAcc [$CI_{95\%}$]) for each dataset, model, and IR.*

7.5.3 Magnitude and variance of SHAP

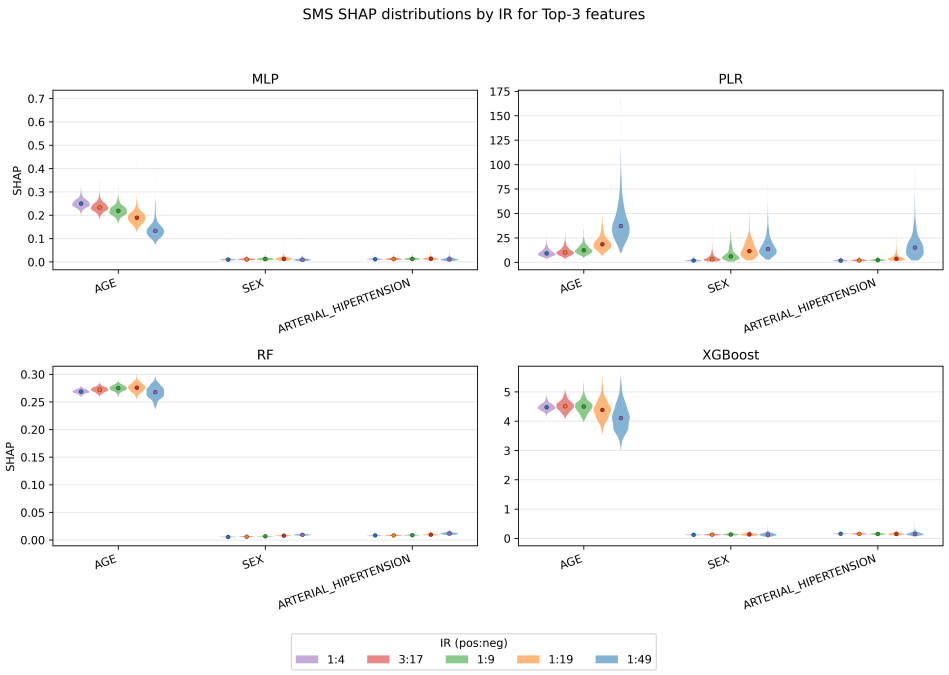
The mean and variance of absolute SHAP values play a central role in interpretability, as they reflect, respectively, how strongly and consistently each feature contributes to model predictions. The distributions shown in Figure 7.2, panels (a) and (b), illustrate how SHAP values change across IRs for the SMS and Diabetes datasets. Correlations between IR and both the mean and variance of SHAP for all datasets used are reported in Table 7.2.

Across models, some common tendencies emerge. Neural and polynomial models often exhibit pronounced shifts in the mean SHAP values as imbalance increases, whereas tree-based models maintain relatively stable means but display an increase in variance. These effects are visible even when the SHAP-based ranking of top features remains stable. For example, in the SMS dataset, all four ML algorithms consistently designate *age*, *sex* and *arterial hypertension* as the most influential variables. The influence of IR on the mean and variance of SHAP values is particularly evident for the polynomial model (see Figure 7.2(a)), and the high values of $\rho_{mean} = 0.79$, $p < 0.001$ and $\rho_{var} = 0.87$, $p < 0.001$ reported in Table 7.2. RF and XGBoost, on the other hand, generate models with nearly constant mean SHAP across IRs, yet their variance increases significantly, with ρ_{var} values near 0.51 and 0.55, respectively. The similar behaviour of these two models may suggest that SHAP variance changes are due primarily to the TreeSHAP explainer rather than to model-specific biases. The neural network algorithm is pretty stable, as neither the mean nor the variance of SHAP values show significant correlation with the IR.

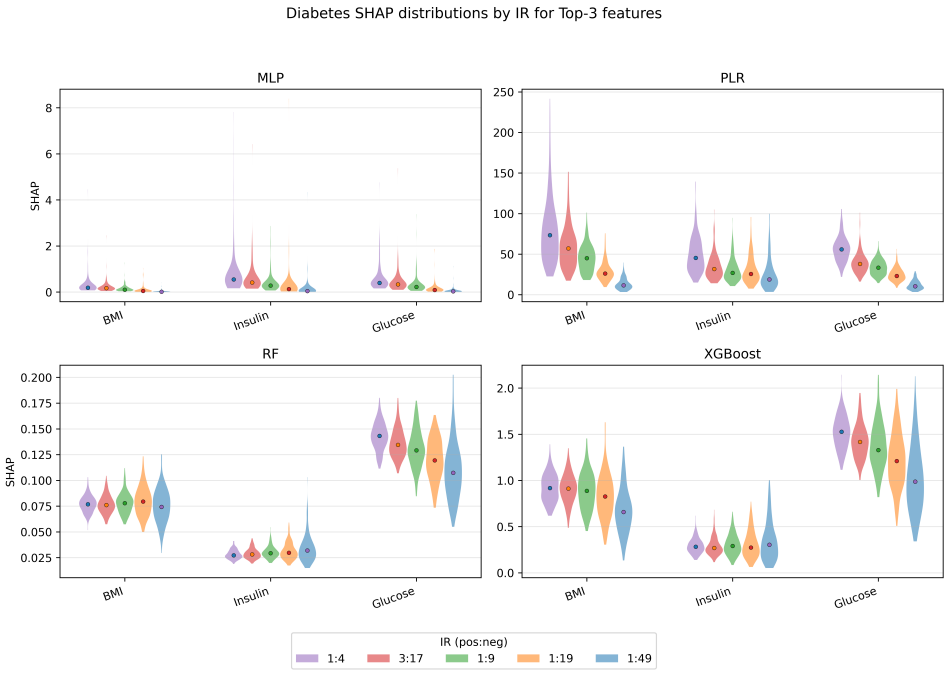
These patterns align with the stability indicators described in Section 7.4.4. For SMS, PCS decreases strongly with increasing imbalance for both RF and XGBoost (correlations of -0.96 and -0.60), pointing to simpler and less stable decision paths across trees for instances in the minority class. Neural network and PLR models show increasing VMP with larger values of IR, with correlations of 1.00 and 0.98, respectively, consistent with their larger shifts in the mean and variance of SHAP values. This would confirm that PCS and VMP are useful to account for stability in SHAP interpretation of specific ML algorithms in variable IR scenarios.

The Diabetes dataset shows a different response to imbalance. For the neural network, the mean and variance of SHAP values show a decreasing tendency as IR increases with $\rho_{mean} = -0.83$, $\rho_{var} = -0.62$, both statistically significant (see Table 7.2). Polynomial models behave similarly. Tree-based models' mean values remain stable, although their variances increase consistently ($\rho_{var} = 0.56$ for RF and $\rho_{var} = 0.67$ for XGBoost, both significant). Clearly, mean and variance work differently with variable IR as Neural networks and Polynomial models experiment changes their mean SHAP values whereas tree-based ones do it with variance.

Again, PCS and VMP stability indicators align with these observations. In Diabetes,



(a)



(b)

Figure 7.2: *SHAP distribution for each model and IR for the (a) SMS dataset and (b) diabetes dataset.*

PCS decreases almost monotonically for RF and XGBoost as the imbalance increases ($\rho = -0.99$ and -0.94), indicating that their decision boundaries become simpler under higher IR. In contrast, MLP and PLR models exhibit greater parameter variability when the data are more balanced (correlation of -1.00 for MLP), which matches the larger mean SHAP values observed at lower IRs. Thus, imbalance affects each model family in a consistent manner, which is reflected both in the mean and variance of SHAP and in model stability.

A similar diversity of behaviours is found in the Forest, Mammography and Satimage datasets (see Table 7.2). All this evidences that the sign of associations of SHAP values mean and variance with IR depends on the model family and dataset (see Figures 7.3-7.5).

Dataset	Method	ρ_{mean}	ρ_{var}
SMS	MLP	-0.05	0.42
	PLR	0.79***	0.87***
	RF	0.32	0.51*
	XGBoost	0.03	0.55*
Diabetes	MLP	-0.83***	-0.62*
	PLR	-0.92***	-0.58*
	RF	-0.01	0.56*
	XGBoost	-0.12	0.67**
Mammography	MLP	-0.83***	-0.67**
	PLR	0.73**	0.62*
	RF	0.22	0.92***
	XGBoost	-0.29	0.88***
Forest	MLP	-0.28	-0.25
	PLR	-0.46	0.01
	RF	-0.11	0.13
	XGBoost	-0.04	-0.12
Satimage	MLP	-0.92***	-0.82***
	PLR	-0.93***	-0.49
	RF	-0.11	0.68**
	XGBoost	-0.12	0.51*

Table 7.2: *Spearman correlation between IR and the mean/variance of |SHAP| for each dataset and model. Significance levels: *** for $p-value < 0.001$, ** for p-value < 0.01, and * for $p-value < 0.05$.*

7.5.4 Consistency of the most important features

We define consistency as the property of features in a dataset to maintain their SHAP-generated relevance rank across samples from the dataset. Consistency of the most relevant features is a key component in analysing SHAP values. Feature importance rankings are stable across resamples when SHAP consistently identifies the same relevant variables across them. Variability across ranks may be due partially to significant heterogeneity amongst resamples but also to the inherent variability of machine learning algorithms, especially in high IR scenarios. As an estimate of ranking variability, we use the entropy measure (see section 7.4.5).

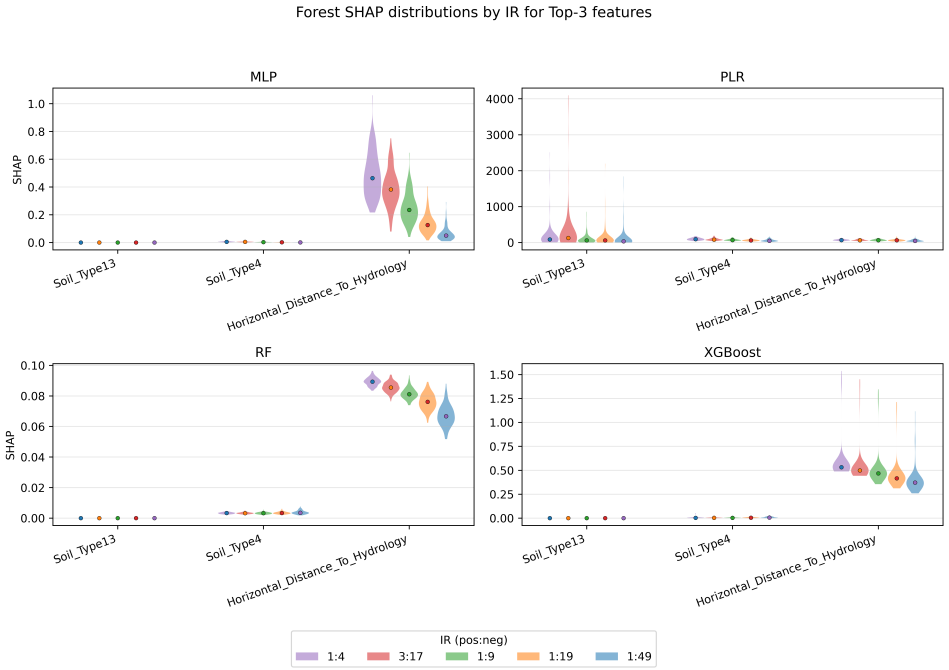


Figure 7.3: *SHAP distribution for each model and IR for the Forest dataset*

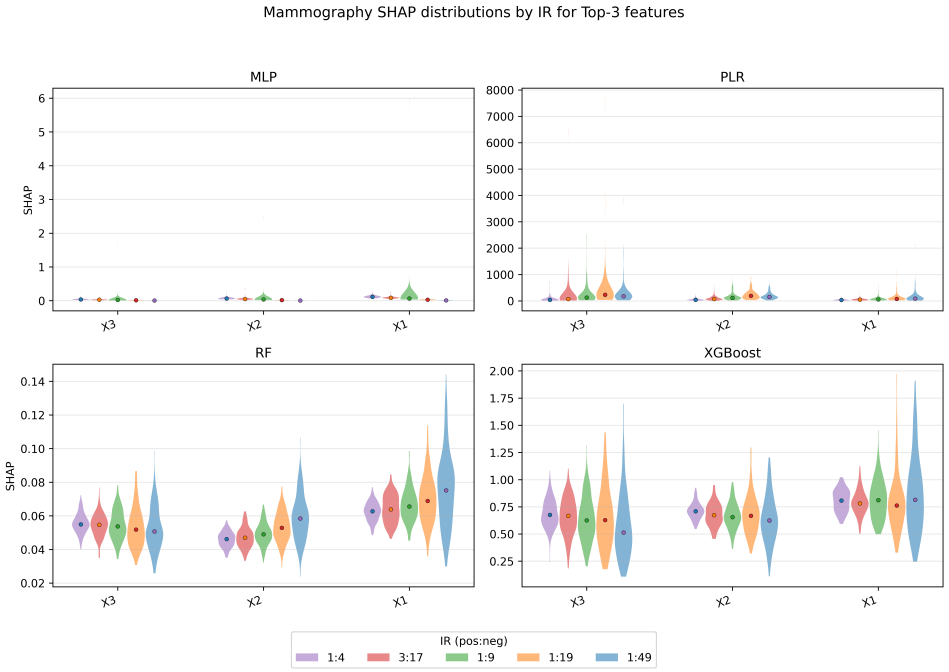


Figure 7.4: *SHAP distribution for each model and IR for the Mammography dataset*

Figure 7.6 shows rank entropy values (y axis) along number of features considered (x axis) for the IRs corresponding to SMS and Diabetes. In all curves we see a monotonically increasing trend, as expected from its definition. It is the separation between the curves that we are interested in. Particularly, for the SMS dataset, Figure 7.6(a) we consider y values across algorithms and separation amongst curves within the same panel, we can see different behaviours across ML models. In all four families, the highest IR, 1:49,

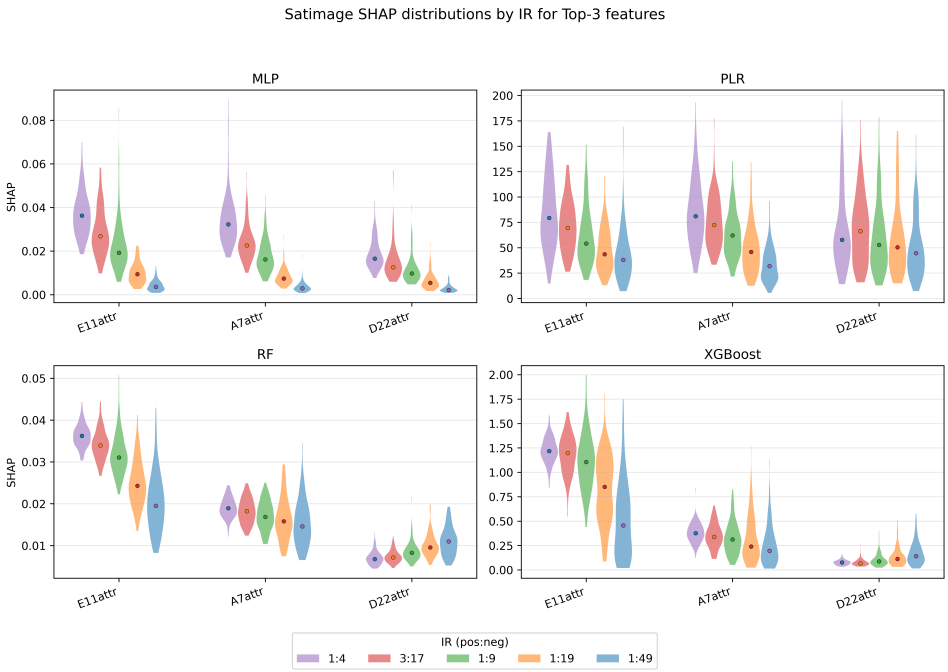
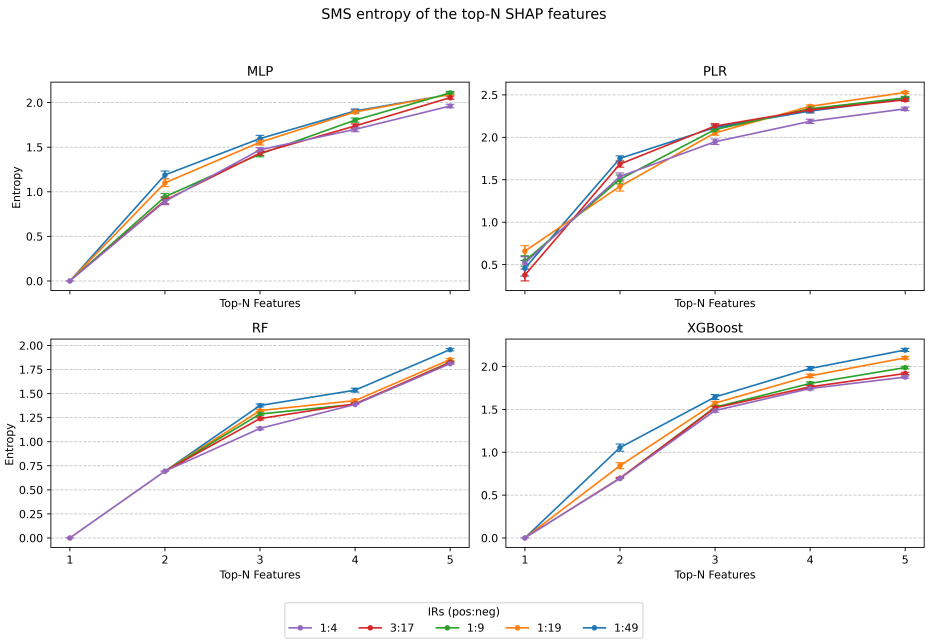


Figure 7.5: *SHAP distribution for each model and IR for the Satimage dataset*

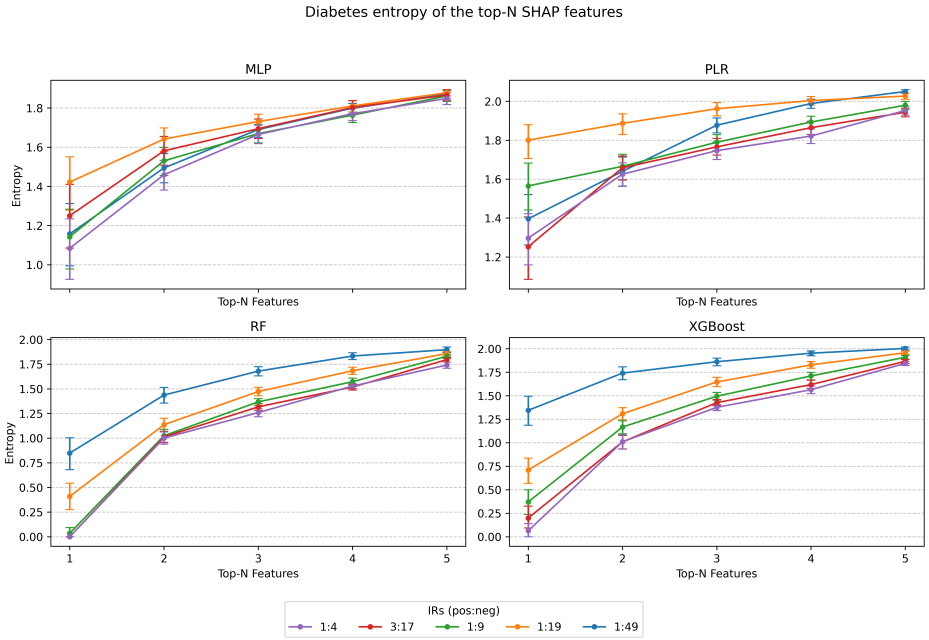
consistently produces the largest entropy values leading to less stable feature rankings and lower reliability in SHAP interpretation. RF works better up to ranks of size 3 and the polynomial model shows more entropy variability than the rest with ranks of size 1. But overall, although IR values influence the rank variability, ML models tend to preserve the identity of the dominant features in SMS, particularly for the tree-based methods.

The Diabetes dataset, shown in Figure 7.6(b), presents more separation between IR curves and variability of behaviours across model families. For the MLP and PLR models, entropy varies noticeably across IRs for the one and two top features, with fast convergence from Top-3 onwards. This suggests that, in this dataset, only the very top positions are sensitive to imbalance, while the entropy trend remains stable with enough rank size. In contrast, RF and XGBoost exhibit a clear stratification by IR. The last two highest IRs consistently obtain higher entropy, i.e., a more random distribution of SHAP importance among the leading features, while the three lower IRs are practically indistinguishable. This suggests that, in Diabetes, the IR affects not only the mean and variance of SHAP values but also the order of feature relevance for tree-based models, whose rankings become more random as imbalance grows.

A comparable heterogeneity is observed in the Forest, Mammography, and Satimage datasets, as detailed in Figures 7.7- 7.9. These confirm that the sign and magnitude of the association between SHAP value stability and IR are dependent on the model family and the specific dataset imbalance, reinforcing the patterns observed in SMS and Diabetes.



(a)



(b)

Figure 7.6: *Entropy of the most important feature sets across different IRs for the (a) SMS dataset and (b) diabetes dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR.*

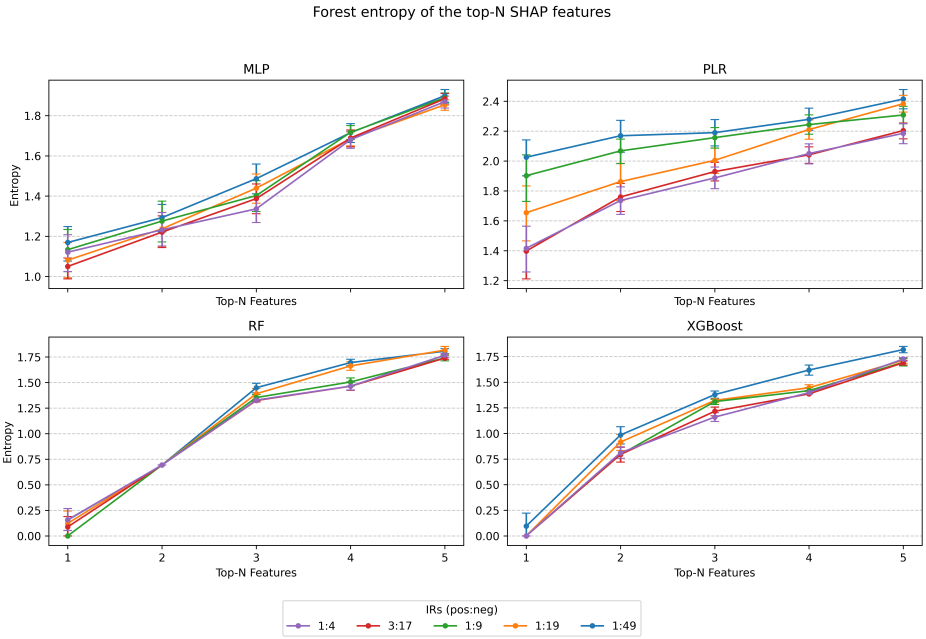


Figure 7.7: *Entropy of the most important feature sets across different IRs for the Forest dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR.*

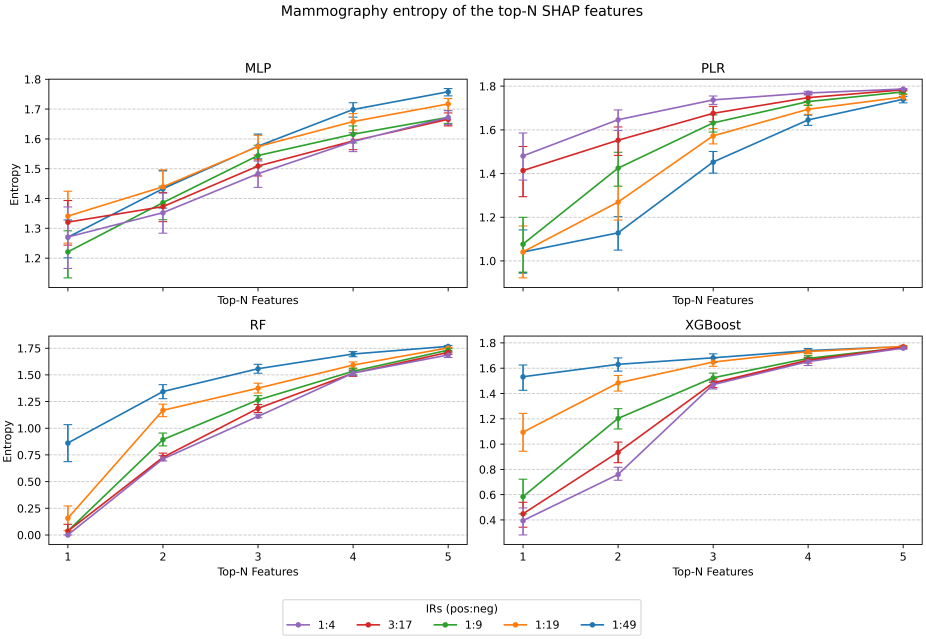


Figure 7.8: *Entropy of the most important feature sets across different IRs for the Mammography dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR.*

7.6 Concluding the chapter

This study shows that class imbalance has a clear and measurable effect on SHAP-based explanations, influencing both the mean and variance of SHAP values and the stability of

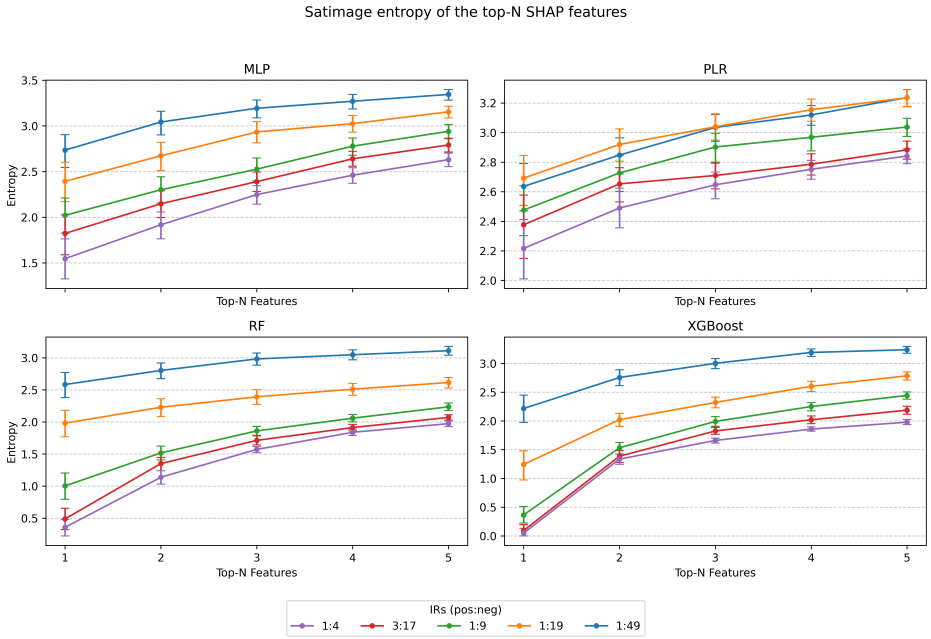


Figure 7.9: *Entropy of the most important feature sets across different IRs for the Satimage dataset. The x-axis represents the number of top k-ranked features considered in each resample, while the y-axis shows the entropy of feature rankings within each IR.*

feature rankings across resamples. By analysing five binary classification datasets under several IRs, and contrasting these results across ML algorithms of different inductive biases and prediction models, we observe that, clearly, class imbalance modifies SHAP estimates of feature importance, where specific datasets and ML models biases play a role in their variability, as pointed out by the Predictive Complexity Score and the variance of learned parameters.

We observe that the mean and variance of SHAP values do not follow a universal monotonic trend with respect to the imbalance ratio. Instead, they depend on both the dataset and the chosen model family. The correlations presented in Table 7.2 illustrate that for certain dataset-model pairs, higher IRs are associated with larger and more variable SHAP values, while for others, the opposite occurs. For example, in the SMS dataset, the PLR model shows a significant positive correlation between the IR and both the mean ($\rho_{\text{mean}} = 0.79$, $p < 0.001$) and variance ($\rho_{\text{var}} = 0.87$, $p < 0.001$) of SHAP values. Conversely, this does not hold for the MLP, suggesting that SHAP explanations for this model family are relatively robust to changes in IR on this dataset. In the Diabetes dataset, both the MLP and the PLR models exhibit a negative correlation between IR and mean SHAP values ($\rho_{\text{mean}} = -0.83$ and $\rho_{\text{mean}} = -0.92$, respectively), along with significant negative correlations for their variances. In contrast, for tree-based models (RF and XGBoost) on the same dataset, no significant association is detected regarding mean values, but there is a positive association with SHAP variances. All these observations support our central finding: SHAP explanations should not be interpreted in isolation from the underlying class distribution in scenarios with significant class imbalance. We

demonstrate that PCS for tree-based ensembles and VMP for MLPs and PLR models are useful as indicators of model stability in relation to scenarios of varying IR.

To assess how consistently models rank their most relevant features in the light of SHAP analysis, we used the notion of entropy to analyse entropy curves from a range beteween Top-1 and Top-5 feature ranks. Entropy monotonically increases with rank size. Consistently across datasets, we observe that imbalanced scenarios produce higher feature rank entropy, i.e., rankings tend to show features uniformly distributed within the rank. These same rankings stabilize at lower IR values. Besides, the strength of this effect varies across datasets and ML models. Interestingly, entropy behaviour is consistent with PCS and VMP. All three help clarify whether changes in attribution magnitude are accompanied by changes in the feature importance.

For SMS, all models show higher entropy in the most imbalanced IRs, meaning that when the data are severely imbalanced, models distribute importance more uniformly across features. As the dataset becomes more balanced, entropy decreases, reflecting more stable and concentrated feature rankings. RF stands out as the most stable model for SMS, as its entropy curves for Top-1 and Top-2 features practically overlap for all IRs, showing that its two most relevant features remain consistently important regardless of imbalance. MLP and XGBoost behave similarly, with small separations between IR curves, whereas the PLR model shows more noticeable differences between IRs, in line with the larger changes observed in its SHAP magnitudes and stability indicators. Overall, SHAP analysis on the SMS dataset exhibits robustness across models, with feature importance ranks remaining relatively stable except in the most imbalanced conditions.

In the Diabetes dataset, the entropy curves reveal that the MLP and PLR models behave in a similar way, with the largest differences between imbalance ratios appearing in the highest-ranked features, particularly for the Top-1 and Top-2 variables. As additional features are incorporated, up to Top-5, the curves progressively converge, indicating that the effect of imbalance is concentrated mainly on the most influential features. At Top-5 ranks, both models show almost identical entropy levels across IRs, reflecting a stable distribution of feature importance at broader feature depths. In contrast, RF and XGBoost exhibit a markedly different pattern, as in both models, the most imbalanced settings consistently produce higher entropy values, suggesting a more uniform spread of feature importance across the top variables. Meanwhile, lower IRs do not differentiate in entropy, indicating increased stability in feature-importance allocation when the minority class is better represented. Taken together, these results show that, in Diabetes, the impact of imbalance on the structure of feature relevance clearly differs across model families, as MLP and PLR models mainly change the magnitude of their attributions, while tree-based models are more affected in terms of how importance is distributed across the top features.

Overall, our results across datasets and models highlight that SHAP explanations are not invariant to class imbalance, as both their magnitude, variance and feature rank sta-

bility can change significantly with IR. The global correlations in Table 7.2, together with the PCS and VMP indicators, show that different model families react differently to imbalance, so feature importance changes may be due to model behaviour rather than to the data alone. Moreover, entropy provides a complementary view of explanation stability, capturing whether the same features remain highly ranked across resamples with identical IR. Balanced accuracy alone is insufficient for evaluating the robustness of explainability, as even models with stable predictive performance can show substantial changes in SHAP distributions, feature rankings and stability indicators when the class distribution changes.

From a practical perspective, if the primary objective is to prioritize stability in the magnitude of the feature importance (mean SHAP values), the use of tree-based models such as Random Forest or XGBoost is recommended, as these models tend to maintain relatively constant SHAP means under class imbalance, showing the impact of data imbalance mainly through an increase in the variance. In contrast, if the priority is to obtain a stable feature importance ranking, the choice depends on the scope of the analysis: MLP and PLR models are preferable when consistency across larger sets of variables is required (from the Top-3 onward), since their instability is usually concentrated only in the first one or two positions before rapidly stabilizing. Conversely, although tree-based models may exhibit high consistency in certain domains, under severe imbalance scenarios they can present higher entropy distribution, distributing importance more uniformly among the leading variables and reducing the reliability of the resulting feature importance ranking.

8

Explainability and Visualization in Dynamic Imbalanced Problems

Following the findings of Chapter 7, which demonstrated that SHAP values are susceptible to changes induced by severe class imbalance, this chapter extends the interpretative framework into dynamic, non-stationary data streams. While the preceding chapter studies the static biases of feature importance, real-world clinical environments continuously evolve due to concept drift. It is therefore necessary to monitor how the model behaviour shifts over time without being confounded by the underlying class imbalance distributions. To address this, the visual analytics proposed in this chapter are coupled with the IPIP-CD methodologies developed in Chapter 6, ensuring that the generated explanations remain reliable against both temporal drift and extreme minority class scenarios.

Continuing the international collaboration established at the Jagiellonian University in Kraków, this chapter applies the theoretical foundations of XAI robustness to practical streaming scenarios. Specifically, it introduces one study designed to monitor concept drift visually under highly imbalanced conditions. The foundational study, entitled "Proposing multi-perspective approach for detecting and explaining concepts drifts in evolving data", was presented at the 2nd TempXAI Workshop for Explainable AI in Time Series and Data Streams (ECML PKDD 2025).

8.1 Introduction

In dynamic environments where data streams continuously evolve, detecting and adapting to concept drift is crucial for maintaining the reliability and accuracy of machine learning models, as discussed in Chapter 3. This challenge is particularly prevalent in applications such as industry, financial forecasting, and healthcare analytics, where shifts in underlying data patterns are common [354]. Beyond drift detection, understanding

and explaining the causes behind these changes is equally important. In industries, particularly for predictive maintenance [355], identifying the root causes of drift can help detect early signs of equipment degradation or process inefficiencies. This allows for early interventions, reducing costly downtime and extending the operational lifespan of the machinery. Similarly, in healthcare, explaining concept drift is crucial during pandemics [356], where sudden shifts in patient data, treatment responses, or infection rates can signal emerging health risks. Understanding these changes enables faster responses and more effective allocation of resources to mitigate the spread and impact of disease outbreaks. This chapter presents a novel method for concept drift detection and explanation using a 3P (three-perspective) approach that evaluates changes across data, model, and explanation perspective. This nomenclature is adopted to emphasize that relying on a single viewpoint is often insufficient; only by triangulating these three dimensions can we achieve a holistic understanding of the underlying causes and dynamics of concept drift. The proposed approach not only tracks shifts in data distribution but also monitors model confidence and interprets the evolving influence of individual features on predictions. The data perspective provides insights into how the distribution shifts. The model perspective captures variations in model confidence, serving as an early indicator of potential drift. The explanation perspective clarifies how feature importance evolves, enabling a deeper understanding of the causes behind model performance fluctuations. The proposed 3P framework is presented in Figure 8.1.

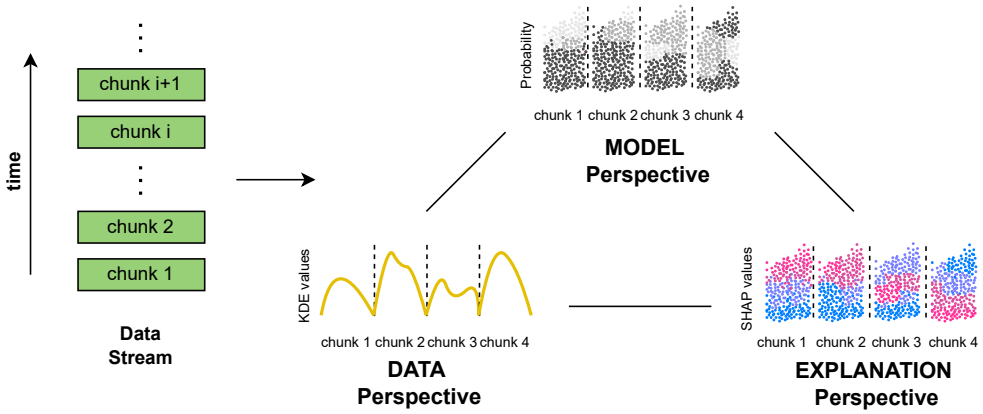


Figure 8.1: *Flowchart of 3P approach. The data stream is divided into chunks, and each chunk is analyzed from three key perspectives—data, explanation, and model perspective—such that the data perspective focuses on tracking the Kernel Density Estimation of test data across chunks, the explanation perspective involves monitoring changes in feature importance across successive models while maintaining a fixed dataset, and the model perspective captures variations in prediction probabilities across successive models while using a fixed dataset.*

As established in the systematic review of the literature (Chapter 3), existing methods for concept drift detection often lack interpretability or provide explanations only after drift has been identified. They frequently fail to pinpoint which features cause the drift and struggle to handle both abrupt and gradual changes in real time. To address these gaps, the present chapter of the thesis adds, to the tools presented in chapters from 5 to

7, an approach that natively integrates drift detection with explainability.

The rest of the chapter is structured as follows: section 8.2 offers an in-depth explanation of the proposed 3P approach. Section 8.3 presents empirical studies on the influence of SHAP values, data distribution and model performance in drift detection, using both benchmark and real–world datasets. Finally, section 8.4 concludes the chapter by summarizing the most relevant conclusions.

8.2 Integrating data, explanation, and model performance into a 3P approach

Building upon the insights regarding XAI robustness established in Chapter 7 and the stream processing methodologies detailed in Chapter 6, this section formalizes the integration of these concepts into a unified monitoring framework. In real-life scenarios, data streams are often heavily imbalanced, requiring robust methodological interventions. To address this, we employ IPIP method, which was previously formalised in Section 5.3.3 and extended to dynamic data streams using a passive adaptation approach in Chapter 6. Then, we ensure that the minority class remains adequately represented throughout the stream while continuously updating the underlying ensembles to deal with a passive approach with concept drift.

Let us consider the data stream $S = \{(\mathbf{x}^0, \mathbf{y}^0), (\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^N, \mathbf{y}^N)\}$ described by n-dimensional attribute vector $\mathbf{x}^i = [x^i_j]^n_{j=1}$ and a class label $\mathbf{y}^i \in \{0, 1\}$. We divide this data stream into m equally sampled chunks X^k_{chunk}, so that $S = \{X^k_{chunk}\}^m_{k=1}$. Each chunk is further split into training $(X^k_{train}, y^k_{train})$ and testing datasets (X^k_{test}, y^k_{test}) with split ratio 80% for training and 20% for testing. We train a classification model $f^k : \mathcal{R}^n \rightarrow \{0, 1\}$ on each training dataset X^k_{train}. These models are then utilized to compute SHAP values [75], denoted as ϕ^k_j, which quantify the contribution of j-th feature to the k-th model prediction. The computation of SHAP values requires defining a dataset on which they will be evaluated. To this end, we define an evaluation dataset X_{eval} derived from the data stream S, which can, for instance, be the first testing dataset. The specific selection of this X_{eval} does not influence the overall analysis of the results, as the primary objective of this methodology is to track relative changes and shifting decision boundaries across successive models, rather than evaluating absolute predictive performance on a specific subset. Consequently, SHAP values for the n-th feature are expressed as $\phi^k_n(f^k(X_{eval}))$. In this approach, models are updated iteratively after each data chunk, as it was introduced in Chapter 6, while SHAP values are consistently computed on a fixed evaluation dataset. This approach allows for tracking changes in the decision boundary across successive models.

The proposed 3P approach relies on temporal visualization to integrate data, model, and explanation perspectives. As illustrated in Figure 8.2, the visualization tracks the

evolution of a single selected feature over time. Time advances from left to right across the x-axis, divided into discrete data chunks. To provide a clear understanding of the drift dynamics, the visual elements are constructed and interpreted as follows:

1. **Data Perspective (Bottom KDE plots):** To monitor shifts in the raw data distribution, we extract the feature values from each successive testing dataset (X^k_{test}). This is visualized at the bottom of the figure using orange Kernel Density Estimation (KDE) plots for each chunk. For these specific plots, the x-axis represents the actual value of the feature, and the y-axis represents the data density. Noticeable variations in these distributions indicate underlying data drift.
2. **Model Perspective (Main scatter plot transparency and accuracy lines):** To capture changes in model confidence, we calculate the prediction probabilities generated by each successive model (f^k) on the fixed evaluation dataset (X_{eval}). In the main central plot, this probability is visually encoded into the transparency of the data points: highly transparent points indicate a lower probability (prediction towards class 0), while opaque points indicate a higher probability (prediction towards class 1). Additionally, the overall model performance is tracked using blue lines across the chunks (dashed for test dataset accuracy, solid for X_{eval} accuracy).
3. **Explanation Perspective (Main scatter plot colors and Top KDE plots):** Finally, to understand how the feature's importance evolves, we compute the SHAP values (ϕ^k_n) for the selected feature using the successive models evaluated on X_{eval}. In the central scatter plot, the y-axis represents the actual feature value. Each data point is colored based on its corresponding SHAP value using a colormap: pink indicates a positive influence on the prediction, blue represents a negative impact, and purple signifies no impact at all. To summarize these changes, green KDE plots are displayed along an additional axis, where the x-axis represents the magnitude of the SHAP values and the y-axis their density.

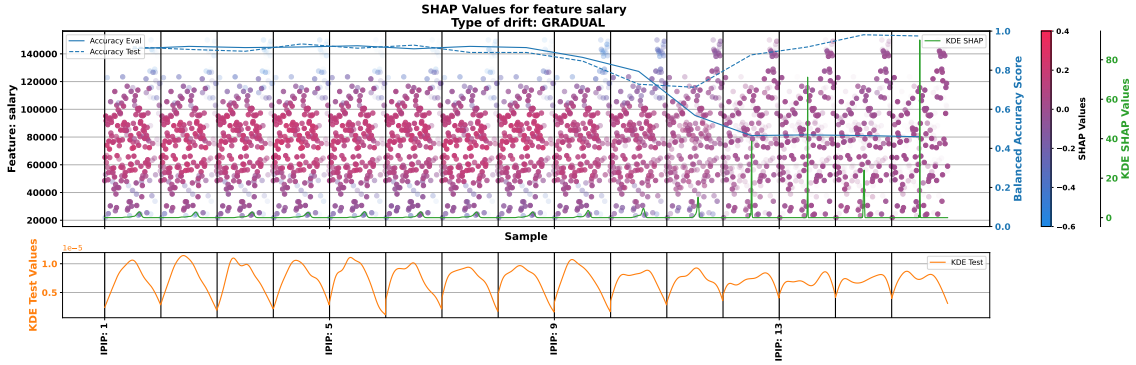


Figure 8.2: *Example of visualization of proposed 3P approach.*

8.3 Empirical studies on concept drift analysis through the lens of the 3P approach

This section illustrates the practical application of the 3P approach for detecting and analyzing concept drift, along with an empirical study of the results. The evaluation is conducted on three streams exhibiting different drift dynamics: a synthetic stream generated with the data streams simulator AGRAWAL, industrial data from a cold rolling mill simulator (TCM), and the clinical SMS stream for predicting whether or not a COVID-19 patient requires hospitalisation. Detailed characteristics of these datasets and their drift generation protocols have been described in Chapter 4 (Section 4.2.3).

Our study shows that the proposed approach enables the detection and explanation of drift patterns, along with the dynamics of model adaptation, which would otherwise remain invisible to both the system user and the model designer. To ensure reproducibility, the source code and implementation details for the synthetic and cold rolling mill datasets can be found at https://github.com/antoniogt/3PApproach, whereas the clinical SMS dataset remains private due to strict medical confidentiality protocols. The remainder of this section details these experiments, beginning with an examination of the synthetically generated drifts, followed by an evaluation on the real-world datasets.

8.3.1 Synthetic dataset

Before evaluating the proposed methodology on complex real-world data streams, it is essential to validate it in a controlled environment. The use of a synthetic dataset provides a known ground truth regarding when and how concept drift occurs, allowing for a precise and unambiguous assessment of the 3P approach's detection capabilities. For our experiment, we use the AGRAWAL data stream generator [259], implemented within the MOA framework [258]. The generator creates a data stream with nine features, including six numerical and three categorical attributes. It uses ten predefined functions to assign binary class labels, likely indicating whether a loan should be approved. These functions, which define different decision rules, can be found in the original paper [259]. We generated two datasets: one with abrupt concept drift and another with gradual concept drift. The generated data stream is first divided into 34 distinct chunks. To induce the desired class imbalance ratio, each chunk is then randomly undersampled to contain exactly 1000 instances, enforcing a 20% minority-class instances and 80% majority-class instances. The evaluation dataset for computing SHAP values is the test data from the first chunk $X_{eval} = X^1_{test}$. Concept drift occurs in chunk 12, where the decision boundary shifts from AGRAWAL's function 2 to function 3, and again in chunk 24, where it transitions from AGRAWAL's function 3 to function 5.

To generate classifiers on the data stream we employed the IPIP algorithm (See Section

8.2) to process imbalanced data chunks efficiently and adapt to changes. Figures 8.3 and 8.4 present the visualization of our approach for the AGRAWAL stream, illustrating abrupt and gradual concept drift in the loan and salary features, respectively. Red boxes, which mark the exact periods where drift occurs (around chunks 12 and 24), indicate how the 3P approach captures these dynamics. Several major differences emerge between these cases.

In the case of abrupt drift (Figure 8.3), the red boxes reveal an immediate transformation. First, if we look at the salary feature subplot around chunks 12 and 13: in the top red box (main scatter plot), the balanced accuracy drops instantly and the data point colors switch quickly, signaling a sudden change in feature importance. Directly below it, the bottom red box (orange KDE plot) confirms a simultaneous sudden shift in the raw data distribution. A similar effect is visible for the loan feature: directing attention to the red boxes around chunks 24 and 25, one can observe the same sharp drop in accuracy at the top and abrupt change in the green SHAP KDE distributions, followed with a change in the distribution of the test KDE distribution also in the bottom orange plots.

However, if we look at the gradual drift analysis in Figure 8.4, where the transitions within the red boxes are visibly smoother and occur over a wider window of chunks. For the salary feature, looking at the top red box spanning chunks 10 to 14, the decline in the balanced accuracy is progressive rather than a sudden vertical drop. Furthermore, the visibility of the data points fades slowly due to increasing transparency, reflecting a gradual loss of model confidence. Finally, examining the loan feature subplot within the red boxes between chunks 24 and 28, it is clear how the green SHAP distributions (in the top panel) and the orange data KDE plots (in the bottom red box) evolve their shape slowly over time, demonstrating the model's adaptive process during a non-instantaneous shift.

8.3.2 Cold rolling mill simulator

In this section, we will showcase the practicality of our approach using a real-world example from a steel manufacturing plant. Steel production mainly relies on continuous casting, which results in slabs, billets, or blooms. To obtain the required shape and size, the steel undergoes two rolling processes: hot rolling and cold rolling. During hot rolling, the slabs are heated in a furnace, reshaped through rolling stands, and then cooled. If additional thickness reduction is required, the steel is further processed in a cold rolling mill. Our study focuses on this second stage, where the variety of rolled products causes data drift. The cold rolling stage of steel production is responsible for reducing the strip's thickness by passing it through a pair of rolls. Initially, the steel coil is unrolled on the uncoiler, and during the rolling process, the strip moves slowly between the rolling stands. Once the strip is gripped by the uncoiler, the process speeds up to reach the desired exit thickness. Cold rolling is a continuous process where raw data samples are collected at regular time intervals, allowing the process data to be treated as

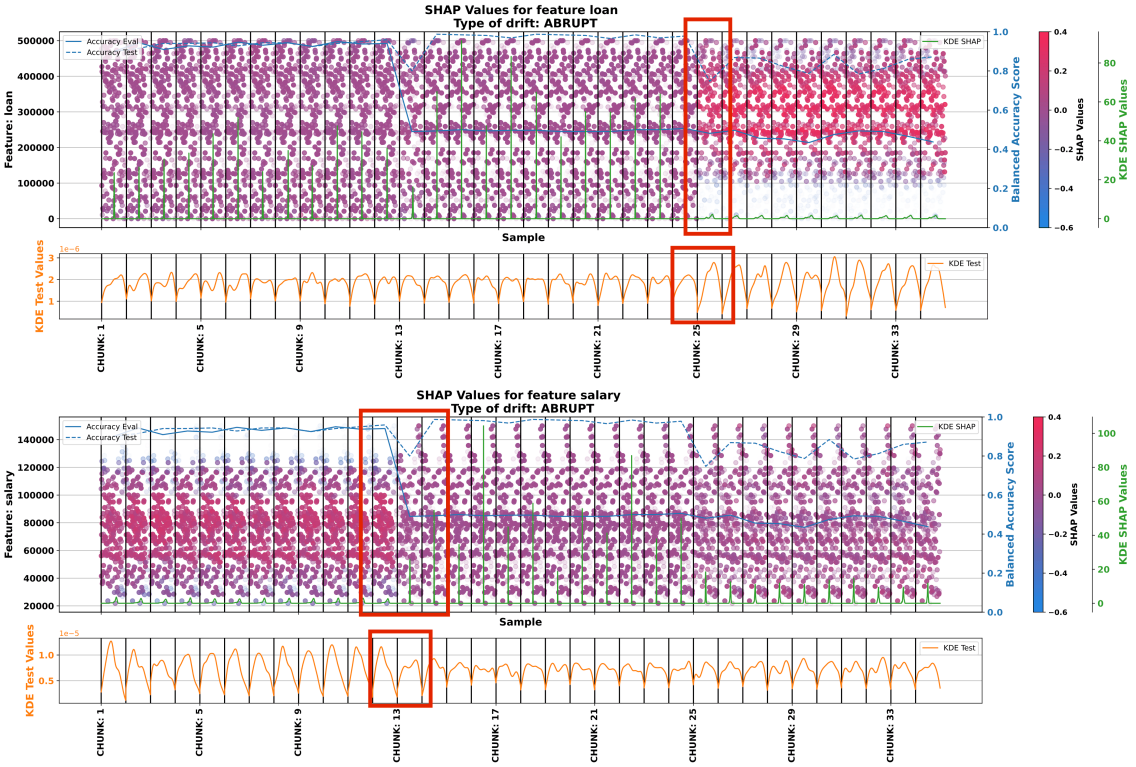


Figure 8.3: *Proposed visualization for AGRAWAL data stream. The plots correspond to abrupt concept drift.*

a data stream.

The analyzed data stream originates from a five-stand cold rolling mill simulator, which was developed using real measurements from a steel company (See Section 4.2.3 for more details). In the selected dataset, the class attribute indicates anomalies associated with physics-based deviations in the rolling process. Four types of anomalies were introduced: anomalies in reduction, work roll anomalies, bearing anomalies, and electric motor anomalies. For the purposes of this study, the problem is treated as a one-class classification task, where the class label is equal to 1 if any of the aforementioned anomaly types occur and 0 otherwise. The dataset was divided into 20 equally sized chunks. Each chunk was further split into training and testing subsets using an 80% - 20% ratio, respectively. An evaluation dataset, denoted as X_{eval}, was constructed by randomly selecting 100 samples labelled as class 1 (anomalies) and 100 samples labelled as class 0 (normal instances) from all test subsets. This approach ensures a balanced evaluation dataset, which facilitates the observation of changes in the decision boundary over time. For classification, a basic implementation of the RF (RandomForestClassifier) from the *scikit-learn* Python library was employed [357]. While the IPIP framework is the primary engine of this methodology, a standard Random Forest was selected for this specific industrial dataset to demonstrate the versatility of the 3P visualization approach and to observe baseline drift dynamics without the influence of active ensemble resampling. The training dataset for this model evolves incrementally: after each chunk is processed,

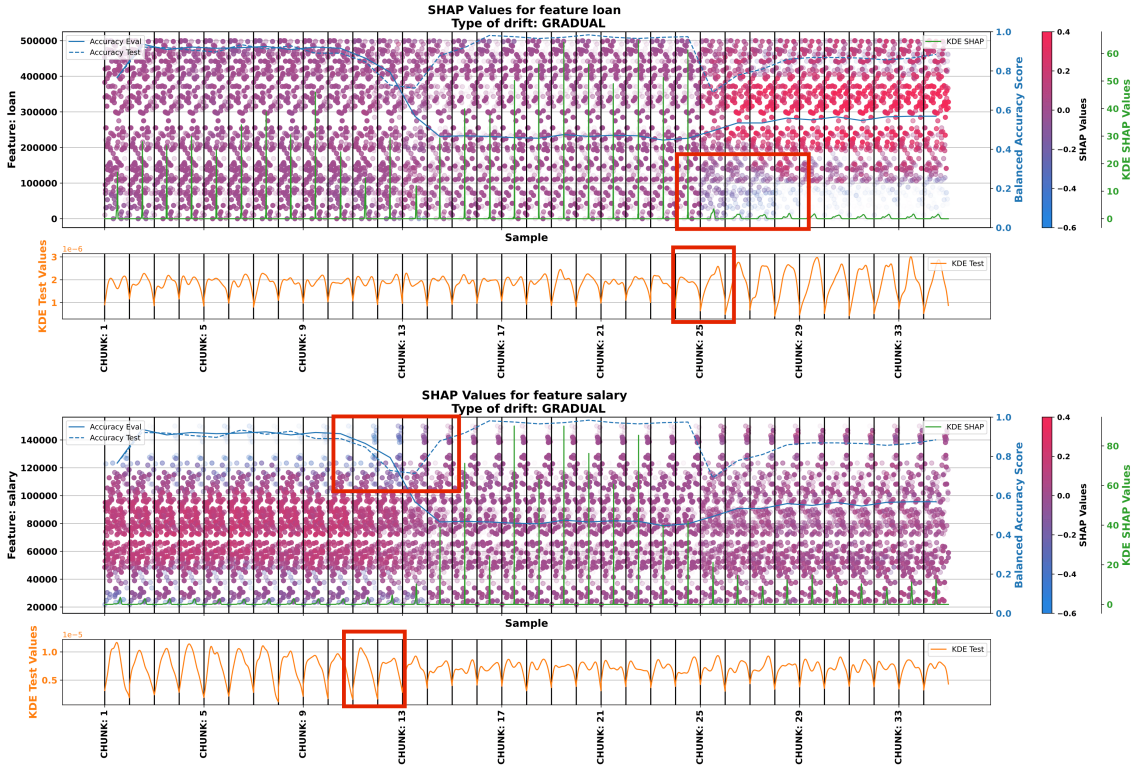


Figure 8.4: *Proposed visualization for AGRAWAL data stream. The plots correspond to gradual concept drift.*

the training set is expanded by including the training data from the current chunk. Consequently, for the first chunk, the model is trained only on data from that chunk, for the second chunk, it is trained on data from both the first and second chunks, and so on, progressively incorporating more data with each iteration.

In the presented dataset, two types of drift are introduced. First, a drift associated with the introduction of previously unseen product types occurs, with the number of rolled products increasing after the 8th chunk. Second, a drift related to increased production capacity is introduced, characterised by heightened power and motor speed, which takes effect after the 16th chunk. Figures 8.5 and 8.6 present the results for features torque_1 and force_1, and motor_power_3 and motor_power_4, respectively. Red boxes highlight the most interesting areas. In Figure 8.5, a shift in the decision boundary becomes evident in chunk 9. Samples with higher values of torque_1 and force_1 begin to lose significance for the model, as indicated by a decrease in their prediction probabilities. Furthermore, a significant shift is visible in the KDE of SHAP values for torque_1. Similarly, force_1 shows a noticeable change in the KDE across test datasets. This change corresponds to the first concept drift, associated with the introduction of a higher number of rolled products. In Figure 8.6, another alteration in the decision boundary is observed in chunk 17. Samples with values in the range of $(4000, 5000)$ for motor_power_3 and $(3250, 3750)$ for motor_power_4 become less important to the model, as reflected by a reduction in their prediction probabilities. A corresponding shift in the KDE of SHAP values for

motor_power_4 further indicates the presence of the second drift, which is linked to an increase in production capacity.

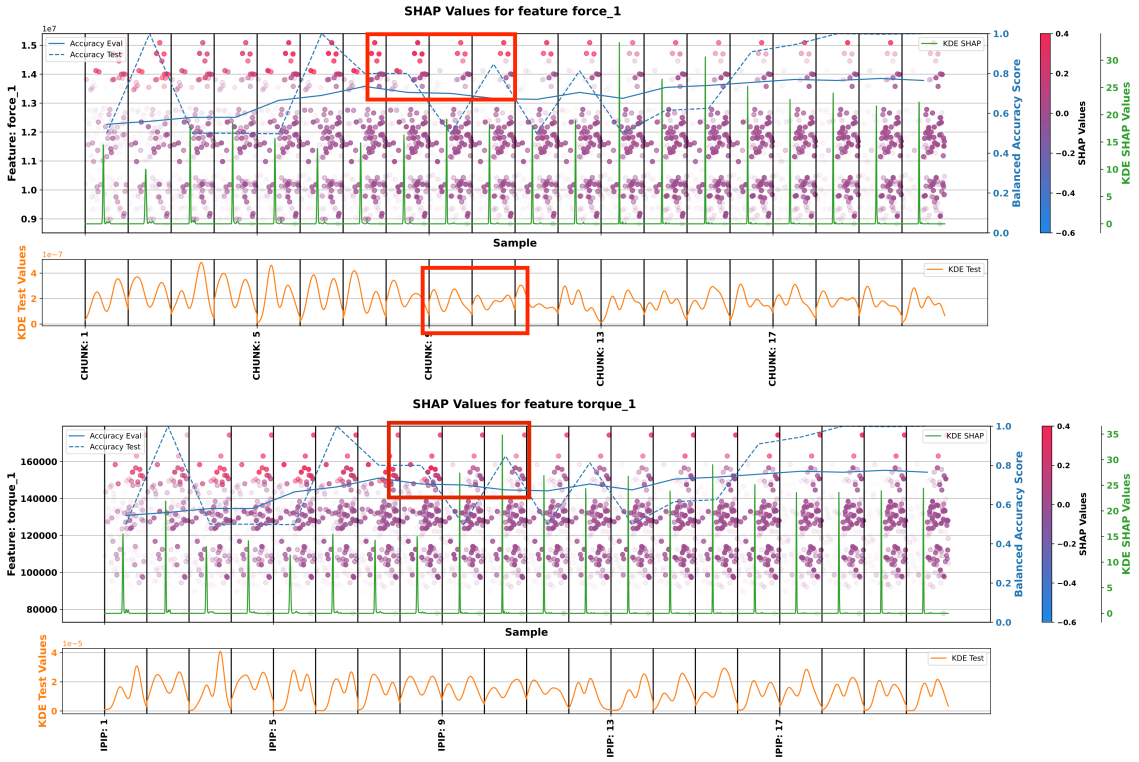


Figure 8.5: *Proposed visualization for tcm5 dataset. The top plot corresponds to feature force_1, and the bottom one to torque_1.*

8.3.3 SMS COVID-19 dataset

In this section, an additional real dataset with data on patients with confirmed cases of COVID-19 is utilized. This dataset contains patient data from COVID-19 over 26 months, with the first record date in March 2020 and 675126 unique patients recorded across all months. A more detailed description of the dataset can be found in [358]. The objective of this dataset is to predict when a patient would need to be hospitalised (class 1) or not (class 0) from the available information. The problem is highly imbalanced, as only 0.75% of the total patients have been hospitalised. To ensure a sufficient number of hospitalised patients in each chunk, the available data is grouped into temporal chunks of two months. This results in 13 chunks of varying sizes. Each chunk was further divided into training and testing subsets, following an 80%-20% split, respectively. Given the importance of temporality in understanding the development of this pandemic, the evaluation dataset X_{eval} consists of 1500 randomly chosen patients from the test subset of the first data chunk. This facilitates the observation of how the importance of variables for patients at the beginning of the pandemic evolves over time in models trained on patients from the subsequent data chunks. The IPIP method is employed for the classification task, as it is designed to address imbalanced problems, handle data chunks

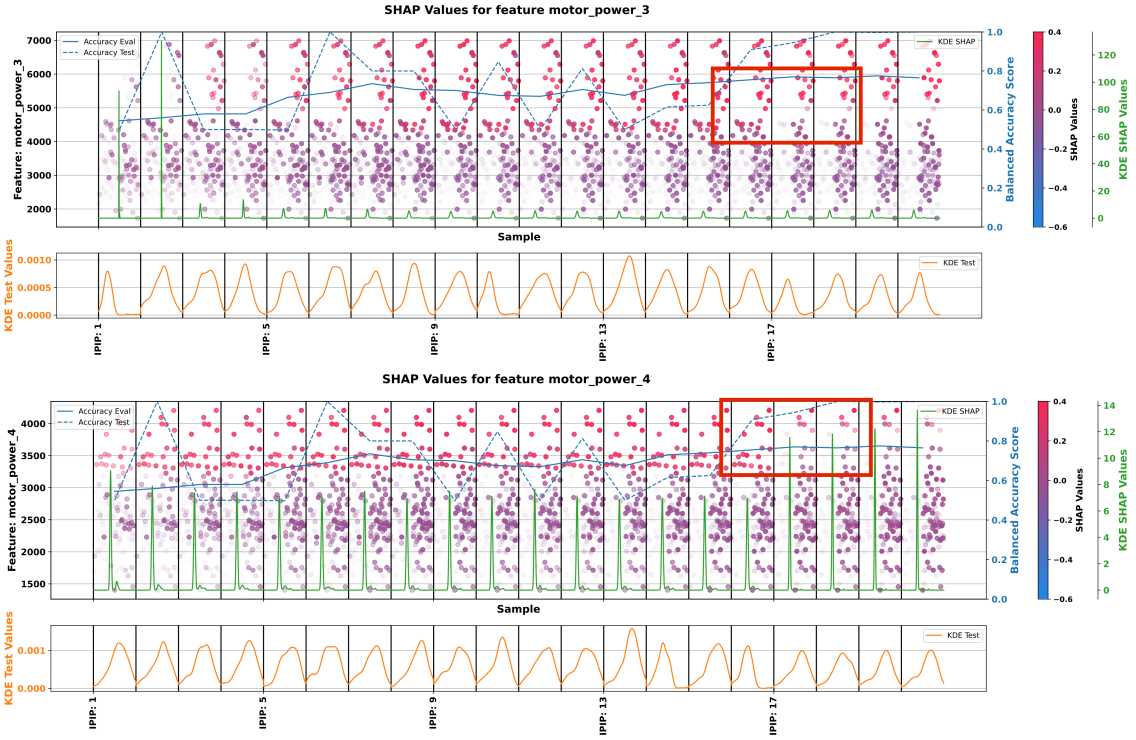


Figure 8.6: *Proposed visualization for tcm5 dataset. The top plot corresponds to feature motor_power_3, and the bottom one to motor_power_4.*

effectively and respond to changes dynamically.

In this dataset, two distinct drifts are evident. The results for the variables chest pain and number of systems affected or age can be observed in the Figure 8.7. Similarly, the results for the variables age and vomits can be seen in Figures 8.8. Red boxes highlight the most interesting chunks of this study. Firstly, due to the worldwide urgency of the pandemic and the high initial infection rate [359], hospital admissions due to the disease in the first months or first wave were for different reasons than for those hospitalized in the second wave or subsequent months, due to the low awareness of the disease. [360, 361]. Consequently, in the initial two data chunks (representing the first four months of the pandemic), as illustrated in Figure 8.7 it can be observed that the variables chest pain and number of systems affected, which are really relevant variables to predict the severity of COVD-19 [358, 362], are not observed with high SHAP values. However, after those first 4 months, the importance is indeed higher, as it is especially shown in the KDE plot, which has a flatter graph and a less pronounced peak in the 0 values of SHAP. This represents the first drift in COVID-19 data. Another notable change can be observed around chunk 9 for the variables age and vomits. This chunk corresponds to September 2021, when the milestone of vaccinating more than 70% of the population of the Region of Murcia with a complete vaccination schedule was reached [358]. This was due to the fact that the vaccination schedules were administered with older people as a priority. In this particular month, the values of the variable age that are pertinent to the prediction of severity are those representing a lower age than in the preceding

chunks. There are higher values in the importance of the age variable between 30 and 50, as these groups of individuals were more susceptible following the vaccination of the elderly. Furthermore, the predominance of the delta variant of COVID-19 occurred between August and December 2021, months in which it is observed (See Figure 8.8) that there are higher values for the variable vomits when showing this symptom, which is associated with the fact that in the delta variant, stomach symptoms such as vomiting are more pronounced and more closely related to a more severe COVID-19 [363, 364]. Those changes in SHAP values represent the second drift in the SMS COVID dataset.

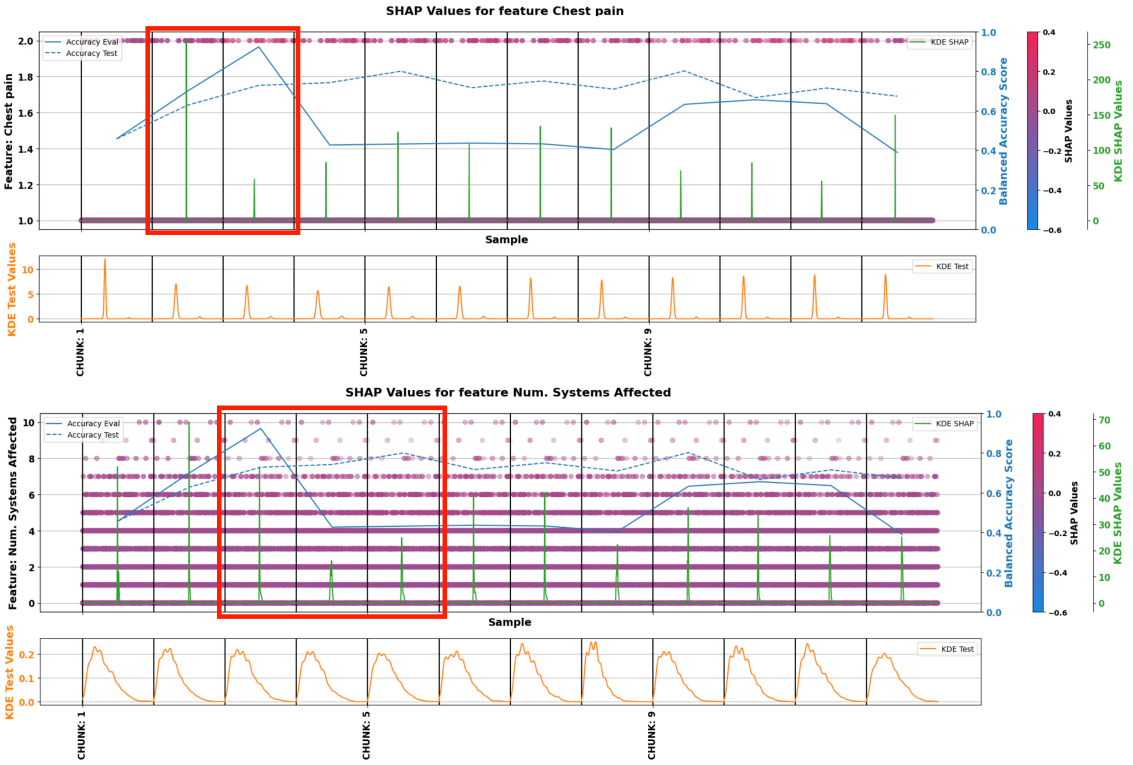


Figure 8.7: *Proposed visualization for SMS COVID-19 dataset. The top plot corresponds to the chest pain symptom, and the bottom one to the number of systems affected by comorbidity.*

8.4 Concluding the chapter

In this work, we show how the three-perspective approach to analysing data streams can be an efficient method for detecting and interpreting concept drift. Focusing solely on either the model or data perspective may be insufficient for understanding concept drift. However, integrating an explainability layer alongside these perspectives provides a much richer insight—not only into how models adapt but also how their performance degrades, how feature importance evolves, and how data distribution changes over time. Moreover, by examining the interactions between these three elements, we gain a deeper and more holistic understanding of drift. While most existing works develop only one of these approaches in isolation, we argue for a unified approach that leverages their combined strengths for a more comprehensive analysis.

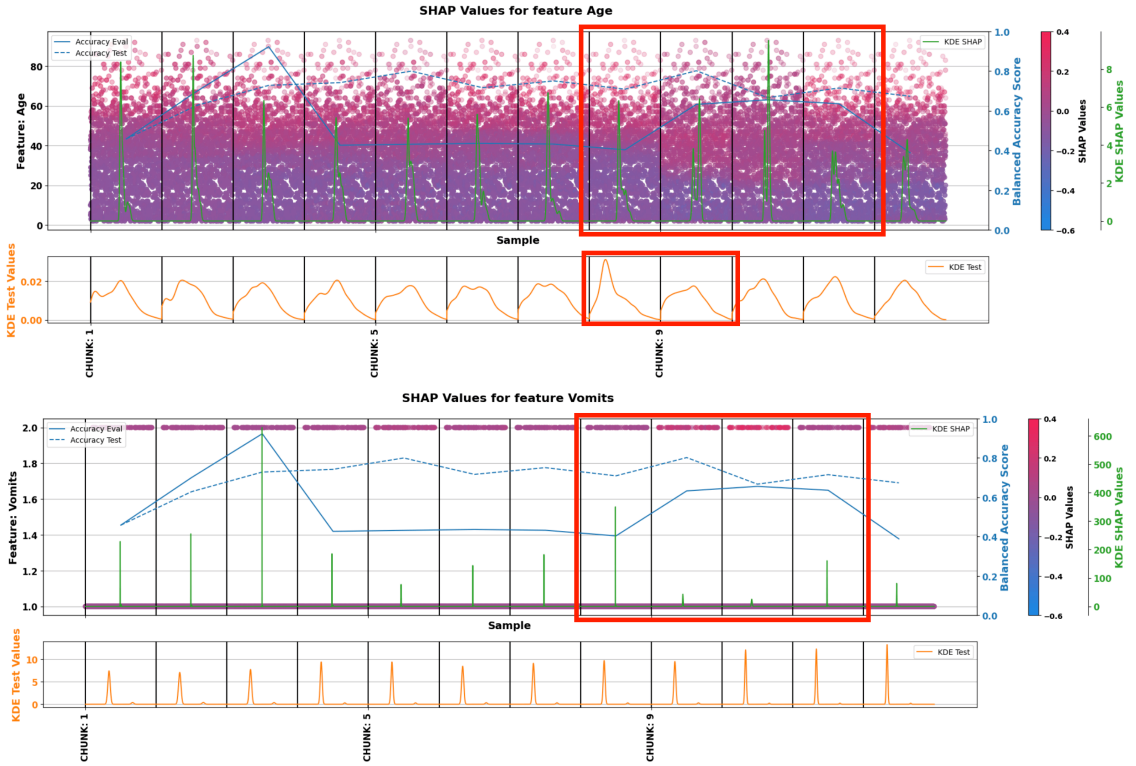


Figure 8.8: *Proposed visualization for SMS COVID-19 dataset. The top plot corresponds to the age feature, and the bottom one to the vomiting symptom.*

The visualization examples on both synthetic and real-world data streams demonstrate the practical utility of this approach. The results of the synthetic AGRAWAL dataset highlight the differences between abrupt and gradual drifts, illustrating how sudden shifts significantly impact model performance, while gradual changes allow a more adaptive transition. The cold rolling mill dataset demonstrates how concept drift arises due to changes in product types and production capacity, affecting model decision boundaries. Similarly, the SMS COVID-19 dataset reveals the evolving importance of features over time, reflecting how the progression of the pandemic influenced hospitalisation risks.

It is important to contextualize the role of this 3P visualization within the broader automated frameworks proposed in this thesis, such as FILM. Methodologies like FILM and its active adaptation mechanisms operate autonomously to maintain predictive performance amidst drift and extreme imbalance. However, this fully automated process inherently acts as a ”black box” to the end-user. The 3P approach is introduced precisely to complement this automation by adding a necessary ”human-in-the-loop” auditing layer. While the model corrects itself automatically, the 3P visual analytics allow domain experts, whether clinical experts or industrial engineers monitoring machinery, to open the black box. It enables them to validate whether the automated model's adaptations align with true, physical real-world phenomena, thereby bridging the gap between automated machine learning and human oversight and trust.

9

Conclusions

This chapter presents the conclusions drawn from the development of this thesis. First, Section 9.1 provides a comprehensive review, interpretation, and discussion of the implications of the results obtained from the methodologies proposed in this research. Next, Section 9.2 addresses the technical and operational limitations encountered during the development and evaluation of these proposals. Section 9.3 outlines the open challenges that remain and proposes promising directions for future research. Finally, Section 9.4 synthesizes the general conclusions, explicitly detailing how the initial objectives of this thesis have been fulfilled.

9.1 Review, interpretation, and implications of the results

The core of this research has been structured around fundamentally interconnected themes: handling static class imbalance, adapting to non-stationary data streams, and ensuring transparent interpretability under such adverse conditions. Below, we interpret the main results obtained in each of these domains and discuss their broader scientific and practical implications.

Initially, this thesis faced the challenge of static datasets exhibiting severe class imbalance through the development of the proposed FILM framework, as detailed in Chapter 5. A primary conclusion derived from our initial experiments regarding model evaluation demonstrates that virtually all standard evaluation metrics exhibit a significant relationship and correlation with the class imbalance ratio. **Consequently, the findings suggest that addressing severe imbalance requires actively mitigating this bias, moving away from simply selecting different traditional metrics towards a robust evaluation framework that explicitly accounts for this problem.**

Moreover, the methodological results of several studies demonstrated that isolated solutions, such as applying simplistic data-level resampling techniques before training, are

insufficient to capture complex decision boundaries in imbalanced problems. To approach this, we developed the FILM framework, which includes a novel method for learning under these situations, and a new evaluation method called BAIC. The idea under our proposed method, IPIP, stems from its resampling idea enforcing balanced representations of classes iteratively across multiple subsets, ensuring that the minority class is fully represented when including all the instances of those resamples, together with an ensemble-based learning approach with base learners such as random forest or logistic regressions. **By enforcing a balanced representation of classes iteratively across multiple base learners, the framework effectively prevents the minority class from being ignored by the overwhelming presence of the other class instances without relying on artificial data generation.**

However, for a correct evaluation strategy in imbalanced problems, as we demonstrated in Chapter 5, using a single metric is insufficient, as most of them are significantly correlated with the imbalanced ratio, and then results may be affected by this problem. By intelligently integrating the basic metrics commonly used in binary classification problems, we proved that BAIC is a new method that allows us to obtain a new metric value that is less correlated with the imbalance factor than 8 baseline metrics such as the Accuracy, Cohen's Kappa, Balanced Accuracy, F1-score, area under the ROC curve, area under the precision-recall curve, MCC, and the Geometric Mean. The arithmetic mean of those 8 metrics was also analysed to prove that a simple integration of them is not valid for our goal. Then, the novel idea of BAIC is to study how the imbalance affects the basic metrics in each specific problem, and to assign them weights associated with the magnitude of that effect. **By aggregating these baseline metrics regulated by their specific weights, BAIC provides a final value that is significantly more reliable for determining the most appropriate method in imbalanced scenarios.**

The development of IPIP and the BAIC evaluation methodology establishes a robust baseline for addressing imbalanced classification in static scenarios. By tackling both the algorithmic and the evaluation facets of the problem, we ensure that models are not only trained to accurately recognise the minority class, but are also fairly assessed without the influence of the imbalanced ratio. Yet, an implication drawn from these static experiments is that real-world environments are rarely invariant. Static assumptions, while essential for establishing these core methodologies, represent a simplification of realities where data distribution and the underlying concept evolve. **Consequently, extending these methodologies to handle non-stationary environments becomes a necessity to ensure their applicability in real-world data streams, motivating the adaptation of both IPIP and BAIC for dynamic scenarios.**

To address the dynamic nature of real-world environments, Chapter 6 extended our proposed methodologies to handle continuous data streams characterized by both concept drift and severe class imbalance. We identified that, as the minority class shifts, static evaluation metrics are affected by this change. By introducing the Adaptive BAIC,

which updates metric integration weights dynamically, we established a robust dynamic performance metric less affected by p_{min} fluctuations compared to the commonly used ones, such as accuracy, balanced accuracy or Cohen's Kappa. Consequently, we developed IPIP-CD, a passive adaptation ensemble methodology. Rather than relying on explicit drift detection mechanisms or synthetic data generation (which risks amplifying noise), IPIP-CD automatically evolves by training new balanced models on incoming data chunks and actively pruning obsolete ones based on their Adaptive BAIC performance. **These findings indicate that continuous adaptation in severely imbalanced streams can be successfully achieved through iterative balanced under-sampling coupled with dynamic model evaluation.**

Moving beyond predictive performance and dynamic adaptation, this thesis identified a critical vulnerability in the widespread adoption of Explainable AI (XAI) methods, specifically SHAP, within imbalanced problems. As explored in Chapter 7, while interpretability is needed in fields like healthcare and industry, post-hoc explanations are frequently assumed to be invariant to the underlying data distribution. To challenge this assumption, we designed a comprehensive empirical framework using an Imbalance-Aware Stratified Sampling (IASS) strategy across multiple datasets and distinct ML model families, including Random Forest, XGBoost, MLP and Polynomial Logistic Regression. **The results demonstrate that class imbalance alters the distribution of SHAP values, meaning that explanations provided to clinicians or engineers can be artificially inflated or destabilised simply due to the distribution of the target feature.**

A detailed review of the results reveals that this imbalance-induced bias does not manifest uniformly. It is highly dependent on the inductive biases of the specific model family. For instance, while neural networks and polynomial models exhibited pronounced shifts in the mean magnitude of SHAP values as imbalance increased, for tree-based methods those means remained relatively stable, but suffered from a significant increase in variance. Furthermore, by evaluating the Shannon entropy of the top-ranked features according to SHAP, we demonstrated that severe imbalance directly degrades the consistency of the feature importance order. When the problem is highly imbalanced, SHAP values are more uniform, leading to unstable explanations. **This highlights that a model might accurately predict a critical minority event, but the reasons it provides for that prediction could be variable under similar data samples.**

The practical implications of these findings prove that high predictive performance is insufficient for evaluating the robustness of explainability. To bridge this gap and provide a diagnostic baseline, this thesis introduced novel stability indicators, namely the Predictive Complexity Score (PCS) for tree-ensembles and the Variance of the Model Parameters (VMP) for neural and polynomial models. These metrics link the instability of SHAP to underlying changes in model complexity and parameter variability under imbalance. **Therefore, to ensure robust explanations using SHAP, the impact of class imbalance must be explicitly evaluated, particularly in non-stationary environments**

where the target variable distribution is subject to continuous change.

Following the study of XAI vulnerabilities in the context of class imbalance, the final phase of this research addressed the challenge of interpreting model behaviour in continuous, non-stationary data streams. As detailed in the subsequent SHAP analysis research, in Chapter 8 we introduced the novel 3P (three-perspective) approach to concept drift detection and explanation. A core interpretation driving this methodology is that relying on a single viewpoint, whether just monitoring incoming data or tracking predictive error, is insufficient to understand data stream dynamics. Instead, the 3P approach captures changes across the data distribution, the model's confidence of the predictions, and the explanation (SHAP) perspectives. **By sequentially projecting the behaviour of successive adaptive models onto a baseline evaluation dataset, this methodology provides a continuous visual representation of how decision boundaries and feature importance evolve over time.**

We evaluated this framework under a controlled synthetic dataset (Agrawal) to prove its capabilities, to later evaluate it in two real-world complex datasets (Cold Rolling Mill and SMS COVID-19 datasets). The 3P approach gives us a visual tool that, after its interpretation, can be directly translated into physical phenomena within the stream. For instance, by continuously monitoring the stability of SHAP values, the approach explicitly identifies when the underlying relevance of specific variables changes, signaling a shift in the problem. Concurrently, tracking the evolution of prediction probabilities across an evaluation baseline provides an early warning mechanism for drops in model confidence, pinpointing exactly when the algorithm begins to struggle with emerging patterns. Simultaneously, assessing statistical variations in the input feature space reveals the onset of distributional drifts even before they fully manifest as predictive errors. **By integrating these three conceptual dimensions, the 3P approach successfully unmasks hidden concept drift patterns and adaptation dynamics that would otherwise remain invisible to standard global performance metrics.**

The most significant practical implication of the 3P approach lies in its relationship with automated machine learning frameworks. While the autonomous adaptation mechanisms developed earlier in this thesis, such as FILM, are strictly necessary to maintain predictive accuracy amidst concept drift and severe imbalance, they inherently operate as "black boxes." The 3P visualization analytics serve as a "human-in-the-loop" auditing layer. This methodology empowers domain experts to open this black box by allowing them to visually validate whether the automated model's continuous adaptations align with true, physical real-world phenomena. **Ultimately, this integration of automated stream learning with visual explainability bridges the gap between algorithmic autonomy and human oversight, ensuring that AI systems deployed in changing environments remain robust and trustworthy.**

To conclude, in this thesis, we have analyzed complex environments by developing and applying advanced ML methodologies and XAI techniques. These novel frameworks

have been utilized to enhance predictive reliability and transparency using both clinical and industrial data. We have addressed specific real-world applications, ranging from COVID-19 patient triage to fault detection in manufacturing processes, while tackling classical machine learning challenges such as severe class imbalance, concept drift, and interpretability biases. **Ultimately, by fulfilling the initial objectives and facilitating the open dissemination of our findings, this research ensures that the proposed algorithms and visual tools, such as the FILM framework and the 3P approach, are accessible, trustworthy, and ready to assist human decision-making in several domains.**

9.2 Limitations

While the methodologies developed in this thesis successfully address the interconnected challenges of class imbalance, concept drift, and explainability, this research is not without its technical and operational limitations. Acknowledging these constraints is essential for a precise interpretation of the results and for guiding future improvements. Below, the specific limitations are detailed according to the chapters of this thesis.

Regarding the static imbalanced learning framework (FILM), including the IPIP algorithm and the novel evaluation metric (BAIC) introduced in Chapter 5, the following limitations were identified:

- The BAIC evaluation metric is inherently a relative and dataset-specific estimate designed strictly for model selection within a single domain. Therefore, it cannot and must not be used to compare the absolute performance of models across entirely different datasets.
- The IPIP algorithm significantly increases the overall training time when compared to simple, data-level resampling techniques. This occurs because it requires building and maintaining an iterative ensemble of base learners.
- Depending on how low the minority class proportion is, the IPIP methodology may be forced to excessively repeat the use of the few available minority instances across the ensemble subsets, which may, in turn, lead to model overfitting.
- In scenarios where the number of minority class instances is extremely small, the balanced resamples generated by the framework will subsequently contain very few data points. Consequently, the individual base models trained on each of these minimal resamples may lack sufficient information to accurately capture the underlying patterns, severely limiting their overall learning capacity.

Regarding the dynamic stream adaptation methodologies and the extension of the framework to address concept drift, as detailed in Chapter 6, the following limitations must be considered:

- The passive chunk-based approach of the IPIP-CD algorithm forces the adaptation

speed to be bounded by the defined chunk size. In scenarios with extremely abrupt and severe concept drifts, the passive pruning mechanism might exhibit a delay in adaptation compared to methods employing active approaches.
- The Adaptive BAIC metric utilizes a one-step lag $(t-1)$ for updating its weights to mitigate the bias of the minority class instances. This design assumes that the correlation between the metrics and the minority class proportion evolves relatively smoothly. Highly sudden and extreme changes in p_{min} within a single step could misalign the applied weights.
- Continuously maintaining, evaluating, and selectively pruning an evolving pool of ensembles across sequential data chunks demands substantial computational and memory resources over time. This could become a bottleneck in extremely high-velocity data streams unless proper memory-management and model-discarding thresholds are set.

Regarding the empirical analysis of explanation biases under imbalanced distributions introduced in Chapter 7, the following limitations must be acknowledged:

- As justified in the methodological design, the investigation of SHAP biases relies on a purely empirical approach. Because SHAP generation implementations are highly model-specific (e.g., TreeSHAP, DeepSHAP, LinearSHAP), formulating a unified theoretical or mathematical bound for these biases across all algorithmic paradigms was unfeasible.
- Consequently, the conclusions regarding how the imbalance ratio affects the magnitude, variance, and entropy of feature attributions are strictly observational. These findings are inherently tied to the specific machine learning model families evaluated during the experiments, namely Random Forest, XGBoost, Multilayer Perceptrons, and Polynomial Logistic Regression, and may vary if applied to entirely different architectures.
- Furthermore, the scope of this evaluation is exclusively centered on SHAP. Other prominent post-hoc interpretability methods, such as Local Interpretable Model-agnostic Explanations (LIME) or Integrated Gradients, were not subjected to this imbalance-aware analysis.
- The empirical evaluation was conducted exclusively on tabular datasets with binary classification problems. Therefore, the observed biases in SHAP distributions under severe imbalance may manifest differently in multi-class classification problems, in regression tasks, or when interpreting complex models trained on unstructured data types, such as images or natural language text.

Regarding the interpretability analysis and the 3P visual monitoring approach developed in Chapter 8, the experimental framework presents the following constraints:

- While SHAP provides valuable insights for explaining model behaviour and inter-

preting concept drift, it is not without flaws and possesses inherent limitations.
- The current SHAP analysis relies on evaluating feature importance over a fixed baseline subset of the data, highlighting the need to enhance the methodology in the future by dynamically selecting evaluation datasets to better adapt to continuous data streams.
- The 3P approach currently functions as a "human-in-the-loop" visual auditing layer, meaning it inherently requires manual interpretation by domain experts to identify changes in the generated plots.
- An operational limitation of the 3P approach is its scalability concerning the feature space dimensionality. For the visual monitoring to be practical and effective, domain experts must have prior knowledge regarding which specific features to observe. Attempting to manually identify SHAP variations across all available variables in high-dimensional datasets would lead to the risk of detecting false negatives.
- Finally, for practical deployment across diverse operational environments, the methodology is currently constrained by the need to develop more user-friendly tools tailored for non-experts.

9.3 Future work

In this section, we will discuss possible improvements or future directions for the research presented in this thesis.

To further enhance the framework for learning in highly imbalanced problems also introduced in Chapter 5, several algorithmic and computational improvements are proposed:

- Implement parallelism and early stopping strategies during the ensemble building process to reduce overall computing time.
- Develop a new version of the IPIP algorithm that trains different model families simultaneously for each balanced resampled subset, selecting the best performing model in each subset instead of training a single ML model.
- Consider and evaluate alternative ensemble approaches for the generated models.

Referring to Chapter 6, in which we presented the extension of the framework to handle concept drift in dynamic data streams, we can point out the following possible improvements and future lines of work:

- Investigate adaptive or dynamic chunk sizing mechanisms for IPIP-CD, allowing the framework to automatically reduce chunk sizes during periods of high variability (abrupt drift) to react faster, and expand them during stable distribution of data over time to improve generalization.

- Explore hybrid adaptation strategies that integrate the passive ensemble pruning of IPIP-CD with active drift detection statistical tests, aiming to accelerate the removal of obsolete models specifically during abrupt drifts.
- Expand the Adaptive BAIC mathematical formulation and the IPIP-CD resampling strategy to address multi-class data streams, ensuring robust model evaluation and adaptation in more complex, non-binary non-stationary environments.

When looking at the analysis of explanation biases in highly imbalanced scenarios, future research should aim to overcome the identified methodological limitations through the following lines of work:

- Develop a unified theoretical and mathematical framework to formally demonstrate and bound the imbalance-induced biases across different algorithmic paradigms, moving beyond purely empirical observations.
- Broaden the empirical evaluation to include a wider variety of machine learning architectures to verify whether the observational conclusions regarding magnitude, variance, and entropy hold across entirely different model families.
- Investigate other prominent post-hoc interpretability methods, such as LIME or Integrated Gradients, to determine if they exhibit similar behaviour when they are exposed to highly imbalanced data.
- Expand the scope of the imbalance-aware analysis to encompass multi-class classification problems and complex models trained on unstructured data domains, such as computer vision and natural language processing.

Referring to the chapter on explainability and visualization in dynamic imbalanced problems, in which we presented the 3P approach to monitor concept drift, we can point out the following possible improvements and future lines of work:

- Explore alternative explainability methods that mitigate SHAP's shortcomings.
- Automate model retraining through techniques like incremental and active learning and explore the 3P framework in those approaches.
- Enhance SHAP analysis by dynamically selecting evaluation datasets.
- Develop user-friendly tools for non-experts, establish benchmarks for drift explanation, and standardize evaluation metrics for practical deployment across diverse environments.
- Develop algorithms for the automatic detection of drift based on the 3P methodology to further complement the human-in-the-loop approach.
- Investigate automated feature-selection or attention mechanisms prior to the 3P visual analytics. Filtering the feature space to highlight only those variables exhibiting the highest statistical variance in their SHAP distributions would guide the domain expert's attention, effectively reducing the risk of false negatives in high-dimensional problems.

9.4 Concluding remarks

The research presented in this thesis represents an effort to bridge the gap between machine learning and the practical challenges of deploying artificial intelligence in evolving environments. Throughout this thesis, the development of novel methodologies has been strictly guided by the initial objectives (See Chapter 2), addressing the interconnected challenges of severe class imbalance, non-stationary data streams, and explainability.

Initially, addressing the first objective, this thesis provided a systematic analysis of how concept drift and non-stationary distributions are handled in medical and epidemiological modelling in Chapter 3. This comprehensive review exposed a critical lack of adaptive frameworks for evolving clinical environments, serving as the motivation for the subsequent methodological developments.

To fulfill the second and third objectives, this research extensively studied the effects of severe class imbalance on traditional performance assessment. By demonstrating that widely used evaluation measures are systematically skewed by class distributions, this work successfully formulated and validated a robust evaluation framework in Chapter 5. Specifically, the introduction of the Bias-Adjusted Integration Coefficient (BAIC) and its adaptive extension provided stable evaluation mechanisms that mathematically mitigate the bias introduced by class imbalance, ensuring reliable model selection in both stationary and dynamic problems, the latter being addressed in Chapter 6 with the AUPC of the Adaptive BAIC.

In response to the fourth objective, this thesis contributed novel adaptive learning approaches designed for environments characterized by severe class imbalance and non-stationary data distributions. The development of the FILM framework, including the IPIP algorithm for stationary datasets (Chapter 5) and the IPIP-CD methodology for non-stationary data streams (Chapter 6), demonstrated that the integration of dynamic resampling with ensemble learning provides a solution for class imbalance learning without relying on artificial data generation techniques.

Furthermore, fulfilling the fifth objective, this research assessed the reliability of model interpretability in dynamic, imbalanced scenarios. The empirical evaluations showed biases in widespread feature importance methods, such as SHAP, when exposed to skewed distributions (Chapter 7). To complement this and provide analytical tools, this thesis introduced the 3P approach (Chapter 8), a visual monitoring framework that unmasks complex concept drifts and aligns model performance and adaptations with true physical phenomena.

Finally, in accordance with the sixth objective, the results and tools developed throughout this thesis have been disseminated to the scientific community. The methodologies, including the FILM framework (`https://github.com/antoniogt/FILM` and `https://`

github.com/antoniogt/FILMDrift), the imbalance-aware SHAP evaluations (https://github.com/antoniogt/Bias-in-SHAP-Values), and the 3P approach (https://github.com/antoniogt/3PApproach), have been made publicly available through open-source repositories to ensure reproducibility. This has been complemented by the communication of these results through a peer-reviewed publication and a presentation at an international conference (See Chapter 10).

10

Contributions of the Thesis

This chapter summarizes the contributions made to the scientific community and the related projects undertaken throughout the development of this thesis. It serves as a comprehensive overview of the actions carried out to fulfill the research dissemination objective established in Chapter 2.

10.1 Publications and dissemination of results

10.1.1 Publication as first author

In this section, I will summarise the scientific paper in which I am listed as the first author:

- **Guillén-Teruel, A.**, Mellina-Andreu, J. L., Reina, G., González-Billalabeitia, E., Rodriguez-Iborra, R., Palma, J., Botía, J. A., & Cisterna-García, A. (2025). Identifying risk factors and predicting long COVID in a Spanish cohort. *Scientific Reports*. `https://doi.org/10.1038/s41598-025-94765-w`
  This article presents the research conducted to identify risk factors and predict the development of long COVID using data from a Spanish cohort. The methodology and results of this study have been thoroughly described in Chapter 4 (See Section 4.1). According to the Scimago Journal Rank (SJR), the journal is Q1 in the *Multidisciplinary* category. The journal has an impact factor of 3.9.

The following three papers, in which I am listed as the first author, are currently under review:

- **Guillén-Teruel, A.**, Caracena, M., Pardo, J. A., de-la-Gándara, F., Palma, J., & Botía,

J. A. (2025). Framework for Imbalanced Learning Machines based on a new bias-adjusted performance measure and a new ensemble-based technique. *Under review*.
This work addresses the critical challenges of handling imbalanced datasets in binary classification tasks, where standard metrics often fail to provide an objective evaluation. We propose a novel metric, the Bias-Adjusted Integration Coefficient (BAIC), designed to aggregate existing measures while penalizing sensitivity to imbalance. Additionally, the research introduces the Identical Partitions for Imbalance Problems (IPIP) algorithm, an ensemble-based technique that enhances model performance in imbalanced data distributions.

- **Guillén-Teruel, A.**, Wojak-Strzelecka, N., Palma, J., Nalepa, G. J., & Botía, J. A. (2025). Artificial Intelligence Approaches for dealing with Concept Drift in Pandemics: A Survey. *Under review*.
  This systematic review provides a comprehensive analysis of Artificial Intelligence and Machine Learning methodologies developed to detect, characterize, and adapt to concept drift during pandemic scenarios. The study categorizes drift based on biological, behavioral, structural, or spatial origins, evaluating the response of classical compartmental models, ensemble methods, and advanced deep learning architectures. This research was conducted as part of the international collaboration with the Jagiellonian University (See Section 10.1.5) and establishes the theoretical foundation for adapting predictive models to non-stationary environments.
- **Guillén-Teruel, A.**, Szelazek, M., Palma, J., Bobek, S., Nalepa, G. J., & Botía, J. A. (2025). Bias in SHAP Values: The Hidden Effect of Class Imbalance. *Under review*.
  This research investigates the reliability of SHapley Additive exPlanations (SHAP) when applied to models trained on imbalanced data. Through a systematic study across different model families, we characterize how class imbalance changes the magnitude and variance of SHAP values, potentially leading to misleading interpretations. This work was developed together with the Jagiellonian University (See Section 10.1.5), and it contributes to the field of eXplainable Artificial Intelligence by providing guidelines for more trustworthy model interpretations in real-world imbalanced scenarios.

10.1.2 Publications as co-author

In this section, I present the scientific publications in which I have contributed as a co-author during the development of this thesis:

- Cisterna-García, A., **Guillén-Teruel, A.**, Caracena, M., Pérez, E., Jiménez, F., Francisco-Verdú, F. J., et al. (2022). A predictive model for hospitalization and survival to COVID-19 in a retrospective population-based study. *Scientific Reports*, 12(1), 18126. `https://doi.org/10.1038/s41598-022-22547-9`
  This paper presents a machine learning approach to predict the risk of hospitaliza-

tion and mortality in COVID-19 patients based on retrospective clinical data. My contribution to this work primarily focused on the development of predictive models using IPIP to provide clinical support. According to the SJR, the journal is Q1 in the *Multidisciplinary* category. The journal has an impact factor of 3.9.

- Cisterna-García, A., González-Vidal, A., Martínez-Ibarra, A., Ye, Y., **Guillén-Teruel, A.**, Bernal-Escobedo, L., & Skarmeta, A. F. (2025). Artificial intelligence for streamflow prediction in river basins: a use case in Mar Menor. *Scientific Reports*, 15(1), 19481. `https://doi.org/10.1038/s41598-025-04524-0`
  This study applies machine and deep learning techniques to forecast streamflow in the Mar Menor basin, addressing the challenge of rapid water increases caused by torrential rains. My contribution to this project is linked to the ThinkInAzul project (See Section 10.1.6). According to the SJR, the journal is Q1 in the *Multidisciplinary* category. The journal has an impact factor of 3.9.

10.1.3 Conference paper

During the development of this thesis, the following work was presented at an international conference:

- Wojak-Strzelecka, N., **Guillén-Teruel, A.**, Bobek, S., Nalepa, G. J., Palma, J., & Botia, J. (2025). Proposing multi-perspective approach for detecting and explaining concepts drifts in evolving data. *TempXAI: Temporal and Sequential Explainable AI Workshop, co-located with ECML-PKDD 2025*. `https://lamarr-institute.github.io/tempxai25/`
  This paper proposes a multi-perspective methodology to detect and explain concept drift in evolving data, aiming to improve transparency in decision-support systems for high-risk domains. I had the opportunity to present this work in person at the workshop. My contribution focused on defining the multi-perspective concept drift analysis problem specifically for imbalanced datasets. Furthermore, I utilised the IPIP-CD framework to address this challenge and conducted extensive experiments on the SMS dataset, which were incorporated into the final publication. This research is a direct result of the collaboration with the Jagiellonian University, which is further detailed in Section 10.1.5.

10.1.4 Collaboration with the University College London

The collaboration with the **University College London** (**UCL**) has been a cornerstone in the expansion of my research toward neurodegenerative diseases. This collaboration focused on the analysis and processing of complex clinical and molecular data, specifically within the field of Parkinson's Disease (PD).

The main contributions and outcomes of this collaboration are summarised below:

- **Exploration of the PPMI Cohort:** This collaboration enabled me to start working with the *Parkinson's Progression Markers Initiative* (PPMI) dataset. I gained experience in handling longitudinal clinical data, which is essential for understanding the progression of PD.
- **Simulation of Concept Drift in Clinical Scenarios:** Building upon the knowledge acquired at UCL, my current research focuses on the development of specialised simulators. These tools are designed to alter PPMI data to generate artificial concept drift, thereby simulating real-world clinical situations.
- **Scientific Publication (Under Review):** As a direct result of the joint efforts with researchers from UCL, I have co-authored the manuscript titled "Parkinson's Disease motor and non-motor progression models emerge from pathway-level transcriptomics".
- **Ongoing Research:** To ensure the progress of these research lines, I maintain regular meetings with the UCL team. These sessions serve to update current investigations and explore future collaborative projects.

10.1.5 Collaboration with the Jagiellonian University

The collaboration with the **Jagiellonian University** (Kraków, Poland), specifically with the *Jagiellonian Human-Centred AI Lab* (*JAHCAI*), has significantly advanced the theoretical and practical results of this thesis. This collaboration focused on two core pillars of modern Machine Learning: Explainable AI (XAI) and Concept Drift detection in non-stationary environments.

The key aspects of this collaboration include:

- **Specialization in XAI and Concept Drift:** My time working with the team in Kraków allowed me to deepen my understanding of SHAP (SHapley Additive exPlanations) values and their application in evolving data streams. This theoretical grounding was essential for developing the methodologies presented in this thesis.
- **State-of-the-Art in Pandemics:** The expertise shared during this collaboration enabled the development of a rigorous survey on AI approaches for concept drift in pandemics. This work provides a critical perspective on how models should adapt to sudden shifts in clinical and epidemiological data.
- **Development of the Multi-perspective Drift Framework:** A major achievement of this collaboration was the proposal of a multi-perspective approach for detecting and explaining concept drift. Unlike traditional methods that rely exclusively on performance metrics, this framework integrates internal model behavior (e.g., SHAP values and model confidence) with external data distributions. This allows for a more robust diagnosis of model degradation.
- **Identification of Bias in Explainability:** A major outcome of this joint research was the discovery of how class imbalance affects the integrity of SHAP values.

This collaboration led to the development of a specific research line focused on bias detection, which is a fundamental contribution of this thesis.

- **Scientific Dissemination:** This collaboration resulted in several international publications, including the work presented at the ECML-PKDD conference and two additional manuscripts currently under review in high-impact journals. Furthermore, I have delivered three presentations at the AIRA (Artificial Intelligence and Robotics) seminar series organised by the GEIST research group, where I shared the findings and progress of the research lines of this thesis (https://www.geist.re/aira:start).

10.1.6 Involvement in other research projects during the PhD

In addition to the main research lines of this thesis, I have actively participated in other funded research projects:

- **PROVIA:** Prognosis prediction in COVID-19 patients based on Artificial Intelligence (*Predicción de pronóstico en pacientes de COVID-19 basado en Inteligencia Artificial*). Project reference: 00007/COVI/20.
- **CALM:** Concept Drift-Aware Machine Learning (*Aprendizaje Automático Sensible a la Deriva de Concepto*). This project focuses on developing robust models capable of adapting to non-stationary environments. ID: PID2022-136306OB-I00.
- **Mar Menor (ThinkInAzul) Project:** Data Science for the Emergence of Intelligence in the Marine and Coastal Environment Through Environmental Monitoring (*Ciencia de Datos para la Emergencia de Inteligencia en el Medio Marino y Litoral a Través de la Monitorización Ambiental*). ID: PRTR-C17.I1, funded by MCIN/AEI and European Union NextGenerationEU.

Bibliography

[1] Geoffrey I Webb et al. "Characterizing concept drift". In: *Data Mining and Knowledge Discovery* 30.4 (2016), pp. 964–994.

[2] Joao Gama et al. "Learning with drift detection". In: *Brazilian symposium on artificial intelligence*. Springer. 2004, pp. 286–295.

[3] Haibo He et al. "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[4] Nathalie Japkowicz et al. "The class imbalance problem: A systematic study". In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

[5] Davide Chicco et al. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), p. 6.

[6] Amalia Luque et al. "The impact of class imbalance in classification performance metrics based on the binary confusion matrix". In: *Pattern recognition* 91 (2019), pp. 216–231.

[7] João Gama et al. "A survey on concept drift adaptation". In: *ACM computing surveys (CSUR)* 46.4 (2014), pp. 1–37.

[8] Jie Lu et al. "Learning under concept drift: A review". In: *IEEE transactions on knowledge and data engineering* 31.12 (2018), pp. 2346–2363.

[9] Katia Koelle et al. "The changing epidemiology of SARS-CoV-2". In: *Science* 375.6585 (2022), pp. 1116–1121.

[10] Bartosz Krawczyk. "Learning from imbalanced data: open challenges and future directions". In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232. DOI: 10.1007/s13748-016-0094-0.

[11] Indrė Žliobaitė. "Learning under concept drift: an overview". In: *arXiv preprint arXiv:1010.4784* (2010).

[12] Albert Bifet et al. “Learning from time-changing data with adaptive windowing”. In: *Proceedings of the 2007 SIAM international conference on data mining*. SIAM. 2007, pp. 443–448.

[13] J Zico Kolter et al. “Dynamic weighted majority: An ensemble method for drifting concepts”. In: *The Journal of Machine Learning Research* 8 (2007), pp. 2755–2790.

[14] Paras Malik et al. “Overview of artificial intelligence in medicine”. In: *Journal of family medicine and primary care* 8.7 (2019), pp. 2328–2331.

[15] Mostafa A Shehata et al. “Deep-learning CT reconstruction in clinical scans of the abdomen: a systematic review and meta-analysis”. In: *Abdominal Radiology* 48.8 (2023), pp. 2724–2756.

[16] Neeraj Beeknoo et al. “A simple method to forecast future bed requirements: a pragmatic alternative to queuing theory”. In: *British Journal of Medicine & Medical Research* 18.4 (2016), pp. 1–20.

[17] Kristina Stuckey et al. “COVID-19 vaccine incentive scheduling using an optimally controlled reinforcement learning model”. In: *Physica D: Nonlinear Phenomena* 445 (2023), p. 133613.

[18] Farah Magrabi et al. “Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications”. In: *Yearbook of medical informatics* 28.01 (2019), pp. 128–134.

[19] Mayur Suresh Gawande et al. “The role of artificial intelligence in pandemic responses: from epidemiological modeling to vaccine development”. In: *Molecular Biomedicine* 6.1 (2025), p. 1.

[20] Sam Abbott et al. “Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts”. In: *Wellcome Open Research* 5.112 (2020), p. 112.

[21] Taylor M Oeschger et al. “Early warning diagnostics for emerging infectious diseases in developing into late-stage pandemics”. In: *Accounts of Chemical Research* 54.19 (2021), pp. 3656–3666.

[22] Christo El Morr et al. “AI-based epidemic and pandemic early warning systems: A systematic scoping review”. In: *Health Informatics Journal* 30.3 (2024), p. 14604582241275844.

[23] Dileep Kumar Chikwarti et al. “Deep Learning for Predictive Healthcare Resource Allocation During Pandemics”. In: *International Journal of Advanced Engineering Technologies and Innovations* (2025).

[24] Hassan Ali. “AI for pandemic preparedness and infectious disease surveillance: predicting outbreaks, modeling transmission, and optimizing public health interventions”. In: *Int J Res Publ Rev* 5.8 (2024), pp. 4605–19.

[25] Geoffrey Fairchild et al. "Epidemiological data challenges: planning for a more robust future through data standards". In: *Frontiers in Public Health* 6 (2018), p. 336.

[26] Mustafa Alhasan et al. "Digital imaging, technologies and artificial intelligence applications during COVID-19 pandemic". In: *Computerized Medical Imaging and Graphics* 91 (2021), p. 101933.

[27] Yang Xu et al. "Early alert systems during a pandemic: A simulation study on the impact of concept drift". In: *LAK21: 11th International Learning Analytics and Knowledge Conference*. 2021, pp. 504–510.

[28] Martin Müller et al. "Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic". In: *arXiv preprint arXiv:2012.02197* (2020).

[29] Pierre Nouvellet et al. "Reduction in mobility and COVID-19 transmission". In: *Nature communications* 12.1 (2021), p. 1090.

[30] Ananth Mahadevan et al. "Cost-aware retraining for machine learning". In: *Knowledge-Based Systems* 293 (2024), p. 111610.

[31] Firas Bayram et al. "From concept drift to model degradation: An overview on performance-aware drift detectors". In: *Knowledge-Based Systems* 245 (2022), p. 108632.

[32] Roberto Souto Maior Barros et al. "A large-scale comparison of concept drift detectors". In: *Information Sciences* 451 (2018), pp. 348–370.

[33] Zeng Li et al. "Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm". In: *Knowledge-Based Systems* 195 (2020), p. 105694.

[34] Vikas Benhur Chitla et al. "Multivariate Analysis and Detection of Concept Drift in Longitudinal COVID-19 Data: Implications for Adaptive Healthcare Strategies". In: *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*. 2024, pp. 1–11. DOI: 10.1109/NMITCON62075.2024.10699076.

[35] Weike Liu et al. "A comprehensive active learning method for multiclass imbalanced data streams with concept drift". In: *Knowledge-Based Systems* 215 (2021), p. 106778.

[36] Samuel Ackerman et al. "Automatically detecting data drift in machine learning classifiers". In: *arXiv preprint arXiv:2111.05672* (2021).

[37] Rajesh Siraskar et al. "Reinforcement learning for predictive maintenance: A systematic technical review". In: *Artificial Intelligence Review* 56.11 (2023), pp. 12885–12947.

[38] William Ogilvy Kermack et al. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pp. 700–721.

[39] Herbert W Hethcote. "The mathematics of infectious diseases". In: *SIAM review* 42.4 (2000), pp. 599–653.

[40] Giulia Giordano et al. "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy". In: *Nature medicine* 26.6 (2020), pp. 855–860.

[41] Jeffrey L Elman. "Finding structure in time". In: *Cognitive science* 14.2 (1990), pp. 179–211.

[42] Sepp Hochreiter et al. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[43] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[44] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[45] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (2002), pp. 2278–2324.

[46] Sina F Ardabili et al. "Covid-19 outbreak prediction with machine learning". In: *Algorithms* 13.10 (2020), p. 249.

[47] Farah Shahid et al. "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM". In: *Chaos, Solitons & Fractals* 140 (2020), p. 110212.

[48] Narinder Singh Punn et al. "COVID-19 epidemic analysis using machine learning and deep learning algorithms". In: *MedRxiv* (2020), pp. 2020–04.

[49] Corinna Cortes et al. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[50] Zhenyu Tang et al. "Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images". In: *arXiv preprint arXiv:2003.11988* (2020).

[51] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[52] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[53] Yazeed Zoabi et al. "Machine learning-based prediction of COVID-19 diagnosis based on symptoms". In: *npj digital medicine* 4.1 (2021), pp. 1–5.

[54] Tianqi Chen et al. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[55] Li Yan et al. "An interpretable mortality prediction model for COVID-19 patients". In: *Nature machine intelligence* 2.5 (2020), pp. 283–288.

[56] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[57] Alexia Couture et al. "Estimating COVID-19 hospitalizations in the United States with surveillance data using a Bayesian hierarchical model: modeling study". In: *JMIR Public Health and Surveillance* 8.6 (2022), e34296.

[58] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[59] Sanjay Kumar. "Monitoring novel corona virus (COVID-19) infections in India by cluster analysis". In: *Annals of Data Science* 7.3 (2020), pp. 417–425.

[60] David M Morens et al. "What is a pandemic?" In: *The Journal of infectious diseases* 200.7 (2009), pp. 1018–1021.

[61] World Health Organization. *Managing epidemics: key facts about major deadly diseases*. World Health Organization, 2023.

[62] Ensheng Dong et al. "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet infectious diseases* 20.5 (2020), pp. 533–534.

[63] Seth Flaxman et al. "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe". In: *Nature* 584.7820 (2020), pp. 257–261.

[64] Adam J Kucharski et al. "Early dynamics of transmission and control of COVID-19: a mathematical modelling study". In: *The lancet infectious diseases* 20.5 (2020), pp. 553–558.

[65] Aniruddha Adiga et al. "Mathematical models for covid-19 pandemic: a comparative analysis". In: *Journal of the Indian Institute of Science* 100.4 (2020), pp. 793–807.

[66] Raju Vaishya et al. "Artificial Intelligence (AI) applications for COVID-19 pandemic". In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14.4 (2020), pp. 337–339.

[67] Jerzy Stefanowski et al. "Stream Classification." In: *Encyclopedia of Machine learning*. Springer, 2017, pp. 1191–1195.

[68] Bartosz Krawczyk et al. "Ensemble learning for data stream analysis: A survey". In: *Information Fusion* 37 (2017), pp. 132–156.

[69] Albert Bifet et al. "Learning from time-changing data with adaptive windowing". In: *Proceedings of the 2007 SIAM international conference on data mining*. SIAM. 2007, pp. 443–448.

[70] E. S. PAGE. "Continuous inspection schemes". In: *Biometrika* 41.1-2 (June 1954), pp. 100–115. ISSN: 0006-3444. URL: https://doi.org/10.1093/biomet/41.1-2.100.

[71] Manuel Baena-Garc et al. "Early Drift Detection Method". In: 2005. URL: https://api.semanticscholar.org/CorpusID:15672006.

[72] Supriya Agrahari et al. "Concept Drift Detection in Data Stream Mining : A literature review". In: *Journal of King Saud University - Computer and Information Sciences* 34.10, Part B (2022), pp. 9523–9540. ISSN: 1319-1578. URL: https://www.sciencedirect.com/science/article/pii/S1319157821003062.

[73] Christian Seiffer et al. "Detection of Concept Drift in Manufacturing Data with SHAP Values to Improve Error Prediction". en. In: *DATA ANALYTICS 2021 : The Tenth International Conference on Data Analytics, October 3 - 7, 2021, Barcelona, Spain*. 2021, pp. 51–60. ISBN: 978-1-61208-891-4. URL: https://www.thinkmind.org/index.php?view=article%5C&articleid=data_analytics_2021_3_40_60034.

[74] Christopher Duckworth et al. "Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19". In: *Scientific Reports* 11.1 (Nov. 2021), p. 23017. ISSN: 2045-2322. URL: https://doi.org/10.1038/s41598-021-02481-y.

[75] Scott M Lundberg et al. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[76] Maciej Mozolewski et al. "Towards Evaluation of Explainable Artificial Intelligence in Streaming Data". In: *Explainable Artificial Intelligence*. Ed. by Luca Longo et al. Cham: Springer Nature Switzerland, 2024, pp. 145–168. ISBN: 978-3-031-63803-9.

[77] Thomas Rojat et al. *Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey*. 2021. arXiv: 2104.00950 [cs.LG]. URL: https://arxiv.org/abs/2104.00950.

[78] Narayanan U. Edakunni et al. *Explaining Drift using Shapley Values*. 2024. arXiv: 2401.09756 [cs.LG]. URL: https://arxiv.org/abs/2401.09756.

[79] Maximilian Muschalik et al. "iPDP: On Partial Dependence Plots in Dynamic Modeling Scenarios". In: *Explainable Artificial Intelligence*. Springer Nature Switzerland, 2023, pp. 177–194. ISBN: 9783031440649. URL: http://dx.doi.org/10.1007/978-3-031-44064-9_11.

[80] Aalok Shanbhag et al. *Unified Shapley Framework to Explain Prediction Drift*. 2021. arXiv: 2102.07862 [cs.LG]. URL: https://arxiv.org/abs/2102.07862.

[81] Matthew J Page et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". In: *BMJ* 372 (2021). URL: https://www.bmj.com/content/372/bmj.n71.

[82] Maria J Grant et al. "A typology of reviews: an analysis of 14 review types and associated methodologies". In: *Health information & libraries journal* 26.2 (2009), pp. 91–108.

[83] Marcos A. Capistrán et al. "Disentangling the role of virus infectiousness and awareness-based human behavior during the early phase of the COVID-19 pandemic in the European Union". In: *Applied Mathematical Modelling* 122 (2023). Type: Article, pp. 187–199. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85161318492%5C&%20doi=10.1016%5C%2fj.apm.2023.05.027%5C&%20partnerID=40%5C&%20md5=ea9c0c2bcd018f07f090b021ad722689.

[84] Mario Nosvelli. "Adaptive learning in containment measures: evaluation of policy interventions during the 2020 waves of Covid-19 in Italy". In: *International Review of Applied Economics* 38.3 (2024). Type: Article, pp. 305–317. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152966639%5C&doi=10.1080%5C%2f02692171.2023.2203473%5C&partnerID=40%5C&md5=934fa6474417b7d85923a2a5910b54fe.

[85] Xiaoyi Fu et al. "An event-centric prediction system for COVID-19". In: *Proceedings - 11th IEEE International Conference on Knowledge Graph, ICKG 2020*. Type: Conference paper. 2020, pp. 195–202. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092462108%5C&doi=10.1109%5C%2fICBK50248.2020.00037%5C&partnerID=40%5C&md5=604cd738db798464044cdc8f226ed17f.

[86] Jaco Tetteroo et al. "Automated Machine Learning for COVID-19 Forecasting". In: *IEEE Access* 10 (2022). Type: Article, pp. 94718–94737. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137598718%5C&doi=10.1109%5C%2fACCESS.2022.3202220%5C&partnerID=40%5C&md5=b08536c91845bb3b439be47c86e73aea.

[87] Ahmad Sedaghat et al. "COVID-19 (Coronavirus Disease) Outbreak Prediction Using a Susceptible-Exposed-Symptomatic Infected-Recovered-Super Spreaders-Asymptomatic Infected-Deceased-Critical (SEIR-PADC) Dynamic Model". In: *CANDO-EPE 2020 - Proceedings, IEEE 3rd International Conference and Workshop in Obuda on Electrical and Power Engineering*. Type: Conference paper. 2020, pp. 275–282. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101077207%5C&doi=10.1109%5C%2fCANDO-EPE51100.2020.9337775%5C&partnerID=40%5C&md5=19cb4a2270523d0efe3463e8b09bbb89.

[88] Hamid Reza Pourghasemi et al. "Assessment of the outbreak risk, mapping and infection behavior of COVID-19: Application of the autoregressive integrated-moving average (ARIMA) and polynomial models". In: *PLoS ONE* 15.7 July (2020). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088811508%5C&doi=10.1371%5C%2fjournal.pone.0236238%5C&partnerID=40%5C&md5=7630dff934e784841114fdaaab63bfe6.

[89] Henrique Mohallem Paiva et al. "A data-driven model to describe and forecast the dynamics of COVID-19 transmission". In: *PLoS ONE* 15.7 July (2020). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089129063%5C&doi=10.1371%5C%2fjournal.pone.0236386%5C&partnerID=40%5C&md5=4bca7d2c2aaaaf036186844d0d5f9d15.

[90] Mengqi He et al. "Combining the dynamic model and deep neural networks to identify the intensity of interventions during COVID-19 pandemic". In: *PLoS Computational Biology* 19.10 October (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174749011%5C&doi=10.1371%5C%2fjournal.pcbi.1011535%5C&partnerID=40%5C&md5=438e22fcf7d5ab95ee9e53c0d486dbaa.

[91] E. Camargo et al. "An incremental learning approach to prediction models of SEIRD variables in the context of the COVID-19 pandemic". In: *Health and Technology* 12.4 (2022). Type: Article, pp. 867–877. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128607843%5C&%20doi=10.1007%5C%2fs12553-022-00668-5%5C&%20partnerID=40%5C&%20md5=6d7d7c575ef9b22010a5f421c7cd74ad.

[92] Qianqian Cui et al. "Dynamic variations in COVID-19 with the SARS-CoV-2 Omicron variant in Kazakhstan and Pakistan". In: *Infectious Diseases of Poverty* 12.1 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150281056%5C&%20doi=10.1186%5C%2fs40249-023-01072-5%5C&%20partnerID=40%5C&%20md5=8c859c0244c7470beffdc873bc477eea.

[93] Yuxuan Zhang et al. "A prognostic dynamic model applicable to infectious diseases providing easily visualized guides: a case study of COVID-19 in the UK". In: *Scientific Reports* 11.1 (2021). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104461695%5C&doi=10.1038%5C%2fs41598-021-87882-9%5C&partnerID=40%5C&md5=ff888b59a8a44ffe21d1cfa755bb6cda.

[94] Shan Gao et al. "Scenario prediction of public health emergencies using infectious disease dynamics model and dynamic Bayes". In: *Future Generation Computer Systems* 127 (2022). Type: Article, pp. 334–346. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115925104%5C&doi=10.1016%5C%2fj.future.2021.09.028%5C&partnerID=40%5C&md5=e9c0568838432235e5ea16d8f2bba2e9.

[95] Ricardo Manuel Arias Velásquez et al. "Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression". In: *Chaos,*

Solitons & Fractals 136 (July 2020). Publisher: Elsevier BV, p. 109924. ISSN: 0960-0779. URL: http://dx.doi.org/10.1016/j.chaos.2020.109924.

[96] Jun Sun et al. "Nonlinear dynamic model for COVID-19 epidemics using the Gaussian distributed wiring small-world network technique". In: *Alexandria Engineering Journal* 81 (2023). Type: Article, pp. 243–255. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171444683%5C&doi=10.1016%5C%2fj.aej.2023.09.015%5C&partnerID=40%5C&md5=14605ed52c98cc0fba0a4035b4114419.

[97] Tongtong Huang et al. "Population stratification enables modeling effects of reopening policies on mortality and hospitalization rates". In: *Journal of Biomedical Informatics* 119 (2021). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108720047%5C&doi=10.1016%5C%2fj.jbi.2021.103818%5C&partnerID=40%5C&md5=641ded708e7d01cb4545cde5488cb1a5.

[98] Ting Zhang et al. "Methods on COVID-19 Epidemic Curve Estimation During Emergency Based on Baidu Search Engine and ILI Traditional Surveillance in Beijing, China". In: *Engineering* 31 (2023). Type: Article, pp. 112–119. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85180340508%5C&doi=10.1016%5C%2fj.eng.2023.08.006%5C&partnerID=40%5C&md5=1553afa11208192f616f4b996e12d168.

[99] Jayanthi Devaraj et al. "Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?" In: *Results in Physics* 21 (2021). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100104393%5C&doi=10.1016%5C%2fj.rinp.2021.103817%5C&partnerID=40%5C&md5=c61c0ccee897a79f23b98f6f077850f3.

[100] Maher Ala'raj et al. "Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections". In: *Infectious Disease Modelling* 6 (2021). Type: Article, pp. 98–111. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097468552%5C&%20doi=10.1016%5C%2fj.idm.2020.11.007%5C&%20partnerID=40%5C&%20md5=8a3c86b18fb4bcbca308e04205536da3.

[101] Tchavdar T. Marinov et al. "Dynamics of COVID-19 using inverse problem for coefficient identification in SIR epidemic models". In: *Chaos, Solitons and Fractals: X* 5 (2020). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089338761%5C&doi=10.1016%5C%2fj.csfx.2020.100041%5C&partnerID=40%5C&md5=0b7e74671774229f8a5653420a9b2978.

[102] Joao Alexandre Lobo Marques et al. "Epidemiology Compartmental Models—SIR, SEIR, and SEIR with Intervention". In: *SpringerBriefs in Applied Sciences and Technology* (2021). Type: Book chapter, pp. 15–39. DOI: 10.1007/978-3-030-61913-8_2.

[103] Amy L. Greer et al. "Using a Dynamic Model to Consider Optimal Antiviral Stockpile Size in the Face of Pandemic Influenza Uncertainty". In: *PLoS ONE* 8.6 (2013). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84879272128%5C&doi=10.1371%5C%2fjournal.pone.0067253%5C&partnerID=40%5C&md5=5539570dc30b81a32c377a715f6c5611.

[104] Yuexin Li et al. "Toward the Impact of Non-pharmaceutical Interventions and Vaccination on the COVID-19 Pandemic With Time-Dependent SEIR Model". In: *Frontiers in Artificial Intelligence* 4 (2021). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107966441%5C&doi=10.3389%5C%2ffrai.2021.648579%5C&partnerID=40%5C&md5=524a26d6c2f5d068b49334221d75ed90.

[105] Vinayaka Gude. "Modeling a decision support system for Covid-19 using systems dynamics and fuzzy inference". In: *Health Informatics Journal* 28.3 (2022). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136464421%5C&doi=10.1177%5C%2f14604582221120344%5C&partnerID=40%5C&md5=375838eeee481fb4ba222f271a530787.

[106] Berik Koichubekov et al. "Is It Possible to Predict COVID-19? Stochastic System Dynamic Model of Infection Spread in Kazakhstan". In: *Healthcare (Switzerland)* 11.5 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85149766205%5C&doi=10.3390%5C%2fhealthcare11050752%5C&partnerID=40%5C&md5=e9a34f2b4bb02a6bdce8d49d3ad4901c.

[107] Xiaoyan Li et al. "Real-Time Epidemiology and Acute Care Need Monitoring and Forecasting for COVID-19 via Bayesian Sequential Monte Carlo-Leveraged Transmission Models". In: *International Journal of Environmental Research and Public Health* 21.2 (2024). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85185900642%5C&doi=10.3390%5C%2fijerph21020193%5C&partnerID=40%5C&md5=07a4528d28e439bea123a61edd5002f1.

[108] Yuansan Liu et al. "Pandemic model with data-driven phase detection, a study using COVID-19 data". In: *Journal of the Operational Research Society* 74.2 (2023). Type: Article, pp. 450–464. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121690251%5C&doi=10.1080%5C%2f01605682.2021.1982652%5C&partnerID=40%5C&md5=49ab716c871e4c62ab6c74204d622dbf.

[109] Andrés L. Suárez-Cetrulo et al. "Modelling the COVID-19 Virus Evolution With Incremental Machine Learning". In: *CEUR Workshop Proceedings*. Vol. 3105. Type: Conference paper. 2021, pp. 1–12. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85126564475%5C&partnerID=40%5C&md5=546f2fa59a441f53b19d783d87f5

[110] Giuseppe Quaranta et al. "Understanding COVID-19 nonlinear multi-scale dynamic spreading in Italy". In: *Nonlinear Dynamics* 101.3 (2020). Type: Article, pp. 1583–1619. URL: https://www.scopus.com/inward/record.uri?eid=2-

s2.0-85090140447%5C&doi=10.1007%5C%2fs11071-020-05902-1%5C&partnerID=40%5C&md5=5d579148f262474fd5e49c32149d3ffb.

[111] Manzoor A. Wani et al. "Risk assessment of COVID-19 pandemic using deep learning model for J& amp;K in India: a district level analysis". In: *Environmental Science and Pollution Research* 29.12 (Oct. 2021). Publisher: Springer Science and Business Media LLC, 18271â€"18281. ISSN: 1614-7499. URL: http://dx.doi.org/10.1007/s11356-021-17046-9.

[112] Yuan Zhang et al. "Prediction of the COVID-19 outbreak in China based on a new stochastic dynamic model". In: *Scientific Reports* 10.1 (2020). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097418795%5C&doi=10.1038%5C%2fs41598-020-76630-0%5C&partnerID=40%5C&md5=c77a386e0f21c97f0ecbd7be427d81b2.

[113] Xin Wang et al. "Short-term local predictions of COVID-19 in the United Kingdom using dynamic supervised machine learning algorithms". In: *Communications Medicine* 2.1 (2022). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85190646242%5C&doi=10.1038%5C%2fs43856-022-00184-7%5C&partnerID=40%5C&md5=8ce6d69c7fe559a97c0dcaa32420bc0b.

[114] Jike Ge et al. "Simulation analysis of epidemic trend for Covid-19 based on SEIRS model". In: *Proceedings of 2020 IEEE 19th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2020*. Type: Conference paper. 2020, pp. 158–161. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85112864710%5C&doi=10.1109%5C%2fICCICC50026.2020.09450226%5C&partnerID=40%5C&md5=9bccbe61489321b969e0931f621358cb.

[115] Vasyl Martsenyuk et al. "Two-Strain COVID-19 Model Using Delayed Dynamic System and Big Data". In: *IEEE Access* 9 (2021). Type: Article, pp. 113866–113878. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113327401%5C&doi=10.1109%5C%2fACCESS.2021.3104519%5C&partnerID=40%5C&md5=bd97946202ebab3a4d0bf044a6c45657.

[116] Xuesong Wang et al. "Uncertainty Estimation With Neural Processes for Meta-Continual Learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 34.10 (2023). Type: Article, pp. 6887–6897. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141612080%5C&doi=10.1109%5C%2fTNNLS.2022.3215633%5C&partnerID=40%5C&md5=869857e7e91123c03ed4b6ef4a4e1a74.

[117] Rashmi Bhardwaj et al. "Development of Epidemiological Modeling RD-Covid-19 of Coronavirus Infectious Disease and Its Numerical Simulation". In: *Infosys Science Foundation Series in Mathematical Sciences* (2021). Type: Book chapter, pp. 245–277. DOI: 10.1007/978-981-16-2450-6_12.

[118] Khaled Jouini et al. "DRIFT-DRIVEN REGRESSION FOR PREDICTING THE EVOLUTION OF PANDEMICS". In: *International Conferences on Applied Computing 2023, AC 2023 and WWW/Internet 2023, ICWI 2023*. Type: Conference paper. 2023, pp. 77–84. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181770561%5C&partnerID=40%5C&md5=c93d81e728e5c9744a3b31b14a1c876b.

[119] Linder Yaroslav et al. "Modeling and prediction of COVID-19 using hybrid dynamic model based on SEIRD with ARIMA corrections". In: *CEUR Workshop Proceedings*. Vol. 2845. Type: Conference paper. 2021, pp. 204–216. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104041535%5C&partnerID=40%5C&md5=31851bf4046a9d84f700cda8fc69a7bf.

[120] Kai Huang et al. "A new online learning algorithm for streaming data and decision support with a Bayesian approach". In: *Journal of Statistical Computation and Simulation* 94.11 (2024). Type: Article, pp. 2483–2499. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85191016185%5C&doi=10.1080%5C%2f00949655.2024.2340581%5C&partnerID=40%5C&md5=7a83f10e2d82e0d922b83b60d18e1d83.

[121] Ashish R. Hota et al. "A Closed-Loop Framework for Inference, Prediction, and Control of SIR Epidemics on Networks". In: *IEEE Transactions on Network Science and Engineering* 8.3 (2021). Type: Article, pp. 2262–2278. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107364425%5C&doi=10.1109%5C%2fTNSE.2021.3085866%5C&partnerID=40%5C&md5=f205671f980a53a6e857ac7cdaaf5b2f.

[122] Jiecheng Song et al. "A Detecting System for Abrupt Changes in Temporal Incidence Rate of COVID-19 and Other Pandemics". In: *Stats* 6.3 (2023). Type: Article, pp. 931–941. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85172233067%5C&doi=10.3390%5C%2fstats6030058%5C&partnerID=40%5C&md5=2740ab549ea2890bbdc004934a3358de.

[123] Seyed Ali Rakhshan et al. "Global analysis and prediction scenario of infectious outbreaks by recurrent dynamic model and machine learning models: A case study on COVID-19". In: *Computers in Biology and Medicine* 158 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85151019085%5C&doi=10.1016%5C%2fj.compbiomed.2023.106817%5C&partnerID=40%5C&md5=3a4263e5d1ef59b028e9dbe690010cd1.

[124] Yifei Ma et al. "Coronavirus disease 2019 epidemic prediction in Shanghai under the "dynamic zero-COVID policy" using time-dependent SEAIQR model". In: *Journal of Biosafety and Biosecurity* 4.2 (2022). Type: Article, pp. 105–113. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133720200%5C&doi=10.1016%5C%2fj.jobb.2022.06.002%5C&partnerID=40%5C&md5=796a1764836a329eb54d7786c0fdd0d1.

[125] Lei Zhang. "A Bio-Inspired Neural Network for Modelling COVID-19 Transmission in Canada". In: *Canadian Conference on Electrical and Computer Engineering*.

Vol. 2022-September. Type: Conference paper. 2022, pp. 281–287. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140998279%5C&doi=10.1109%5C%2fCCECE49351.2022.9918369%5C&partnerID=40%5C&md5=178b93c6cde5bcfaa06f019d3a47a3bf.

[126] Balakrishnama Manohar et al. "A hybridized LSTM-ANN-RSA based deep learning models for prediction of COVID-19 cases in Eastern European countries". In: *Expert Systems with Applications* 256 (2024). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200823915%5C&doi=10.1016%5C%2fj.eswa.2024.124977%5C&partnerID=40%5C&md5=f831a5b6a46c57bd4eea9dfb5dcf7a1f.

[127] David Romero et al. "Adaptive Learning in Agent-Based Models: An Approach for Analyzing Human Behavior in Pandemic Crowding". In: *Applied System Innovation* 6.6 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85180260637%5C&doi=10.3390%5C%2fasi6060113%5C&partnerID=40%5C&md5=ebe3539dfd407315252a0fba3f0dc931.

[128] Vaibhav Vaishnav et al. "Assessment of impact of relaxation in lockdown and forecast of preparation for combating COVID-19 pandemic in India using Group Method of Data Handling". In: *Chaos, Solitons and Fractals* 140 (2020). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089348039%5C&doi=10.1016%5C%2fj.chaos.2020.110191%5C&partnerID=40%5C&md5=e464bc1ee58345dc3ab29a82d46c3e68.

[129] Gustavo A. Muñoz-Fernández et al. "A SIR-type model describing the successive waves of COVID-19". In: *Chaos, Solitons and Fractals* 144 (2021). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85099513137%5C&doi=10.1016%5C%2fj.chaos.2021.110682%5C&partnerID=40%5C&md5=e1e7c425db4b308754d370c578f19c98.

[130] Junaid Farooq et al. "A deep learning algorithm for modeling and forecasting of COVID-19 in five worst affected states of India". In: *Alexandria Engineering Journal* 60.1 (2021). Type: Article, pp. 587–596. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092105109%5C&doi=10.1016%5C%2fj.aej.2020.09.037%5C&partnerID=40%5C&md5=7f451816be5550a078de675a8ed27bfd.

[131] Jiang Long et al. "A switching dynamic model based on phased COVID-19 data in Chongqing and its evaluation". In: *Infection, Genetics and Evolution* 100 (2022). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85126522532%5C&doi=10.1016%5C%2fj.meegid.2022.105270%5C&partnerID=40%5C&md5=93efd7f7b29f092638dcde8d251031b9.

[132] Kraichat Tantrakarnapa et al. "Challenging the spread of COVID-19 in Thailand". In: *One Health* 11 (2020). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092622392%5C&doi=10.1016%5C%2fj.onehlt.2020.100173%5C&partnerID=40%5C&md5=b257f0bb63c1574268e8bc7b1b610f68.

[133] Zipeng Wu et al. "A novel online multi-task learning for COVID-19 multi-output spatio-temporal prediction". In: *Heliyon* 9.8 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168378562%5C&doi=10.1016%5C%2fj.heliyon.2023.e18771%5C&partnerID=40%5C&md5=93caf01bd8da291a3411c908c01a31a8.

[134] Jyotirmay Gadewadikar et al. "A methodology for parameter estimation in system dynamics models using artificial intelligence". In: *Systems Engineering* 27.2 (2024). Type: Article, pp. 253–266. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171277032%5C&doi=10.1002%5C%2fsys.21718%5C&partnerID=40%5C&md5=b7f070633b8bd75852095c41048c9f23.

[135] S Sepehr Tabatabaei et al. "Contagious disease outbreaks: Distinguishing between constant and variable order using the SEIAR model". In: *Results in Engineering* 25 (2025), p. 104125.

[136] Jie Yang et al. "COVID-19 reinfection processes: mathematical modeling, data fitting with time-varying parameters, and optimal control studies". In: *Journal of Biological Systems* 33.02 (2025), pp. 423–449.

[137] Akilashri PSS et al. "Improved Heuristic Algorithm and Optimal Incremental Learning for Dashboard Visualization and Healthcare Data Analysis on COVID Outbreaks". In: *Operations Research Forum*. Vol. 6. 3. Springer. 2025, pp. 1–37.

[138] Naresh Kumar et al. "Non-stationary fuzzy time series modeling and forecasting using deep learning with swarm optimization". In: *International Journal of Machine Learning and Cybernetics* (2025), pp. 1–19.

[139] Naveed Iqbal Shah Hussain et al. "On the stochastic modeling and forecasting of the SVIR epidemic dynamic model under environmental white noise". In: *AIMS Mathematics* 10.2 (2025), pp. 3983–3999.

[140] Yulong An et al. "Parameter Estimation and Prediction of Infectious Disease Model Based on VMD and Transformer". In: *2025 10th International Conference on Intelligent Computing and Signal Processing (ICSP)*. IEEE. 2025, pp. 243–246.

[141] Tianan Yang et al. "Comparison of health-oriented cross-regional allocation strategies for the COVID-19 vaccine: a mathematical modelling study". In: *Annals of Medicine* 54.1 (2022). Type: Article, pp. 941–952. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128437904%5C&doi=10.1080%5C%2f07853890.2022.2060522%5C&partnerID=40%5C&md5=725aab6df7b28120fbefd0aa23625fc4.

[142] A. Fierro et al. "A simple stochastic lattice gas model for H1N1 pandemic. Application to the Italian epidemiological data". In: *European Physical Journal E* 34.2 (2011). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-79951974133%5C&doi=10.1140%5C%2fepje%5C%2fi2011-11011-2%5C&partnerID=40%5C&md5=ffaab0f1e9e8193475a1ac807fe89590.

[143] Yunbo Tu et al. "Modeling and multi-objective optimal control of reaction-diffusion COVID-19 system due to vaccination and patient isolation". In: *Applied Mathematical Modelling* 118 (2023). Type: Article, pp. 556–591. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147852011%5C&doi=10.1016%5C%2fj.apm.2023.02.002%5C&partnerID=40%5C&md5=77fa0da169ee4e30a84cfdbcb32848a1.

[144] Annabelle Collin et al. "Using a population-based Kalman estimator to model the COVID-19 epidemic in France: estimating associations between disease transmission and non-pharmaceutical interventions". In: *The International Journal of Biostatistics* 20.1 (Jan. 2023). Publisher: Walter de Gruyter GmbH, pp. 13–41. ISSN: 1557-4679. URL: http://dx.doi.org/10.1515/ijb-2022-0087.

[145] Yuan Bai et al. "Event-based structural change detection in urban-scale contact network". In: *AAAI Workshop - Technical Report*. Vol. WS-17-01 - WS-17-15. Type: Conference paper. 2017, pp. 529–532. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046102578%5C&%20partnerID=40%5C&%20md5=d683935b083963e2bcf921d29a00232f.

[146] Junyi Ma et al. "RNN enhanced Compartmental Model for Infectious Disease Prediction". In: *Proceedings - 2024 IEEE International Conference on Digital Health, ICDH 2024*. Type: Conference paper. 2024, pp. 225–236. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203828232%5C&doi=10.1109%5C%2fICDH62654.2024.00045%5C&partnerID=40%5C&md5=f4476aa1b4faa2b1f27f9408504defc

[147] Giovanna Nicora et al. "Dynamic Prediction of Non-Neutral SARS-Cov-2 Variants Using Incremental Machine Learning". In: *Studies in Health Technology and Informatics* 294 (2022). Type: Conference paper, pp. 654–658. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85131105950%5C&doi=10.3233%5C%2fSHTI220550%5C&partnerID=40%5C&md5=3356ebe13f3498c5f61a0dd2a7ee7f36.

[148] Lei Li et al. "Multi-task learning sparse group lasso: a method for quantifying antigenicity of influenza A(H1N1) virus using mutations and variations in glycosylation of Hemagglutinin". In: *BMC Bioinformatics* 21.1 (May 2020). Publisher: Springer Science and Business Media LLC. ISSN: 1471-2105. URL: http://dx.doi.org/10.1186/s12859-020-3527-5.

[149] Corrado Loglisci et al. "Discovering variability patterns for change detection in complex phenotype data". In: *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*) 9384 (2015). Type: Conference paper, pp. 9–18. DOI: 10.1007/978-3-319-25252-0_2.

[150] Zhao Wang et al. "Contrastive Cross-Site Learning with Redesigned Net for COVID-19 CT Classification". In: *IEEE Journal of Biomedical and Health Informatics* 24.10 (2020). Type: Article, pp. 2806–2813. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092749906%5C&doi=10.1109%5C%2fJBHI.2020.3023246%5C&partnerID=40%5C&md5=125f2c56d5594138f275ad40cae1e466.

[151] Jianpeng An et al. "COVID-19 Screening in Chest X-Ray Images Using Lung Region Priors". In: *IEEE Journal of Biomedical and Health Informatics* 25.11 (2021). Type: Article, pp. 4119–4127. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113901619%5C&%20doi=10.1109%5C%2fJBHI.2021.3104629%5C&%20partnerID=40%5C&%20md5=530179dbf57726cad9bfae1b4d2a99bb.

[152] Xinpa Ying et al. "Big data analysis for Covid-19 in hospital information systems". In: *PLoS ONE* 19.5 May (2024). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85194018334%5C&doi=10.1371%5C%2fjournal.pone.0294481%5C&partnerID=40%5C&md5=adb0656faf33c4ab56599ba7bcd592fc.

[153] Theresa Roland et al. "Domain Shifts in Machine Learning Based Covid-19 Diagnosis From Blood Tests". In: *Journal of Medical Systems* 46.5 (2022). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127227373%5C&doi=10.1007%5C%2fs10916-022-01807-1%5C&partnerID=40%5C&md5=d76d8e6da9802e1852b8b1510fa6dbe7.

[154] Jonathan Handler et al. "Can a 5-to-90-day Mortality Predictor Perform Consistently Across Time and Equitably Across Populations?" In: *Journal of Medical Systems* 47.1 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163824765%5C&doi=10.1007%5C%2fs10916-023-01962-z%5C&partnerID=40%5C&md5=88b9274237ba3c1c18cb4aaf9faac193.

[155] Kamaryn T. Tanner et al. "Dynamic updating of clinical survival prediction models in a changing environment". In: *Diagnostic and Prognostic Research* 7.1 (Dec. 2023). Publisher: Springer Science and Business Media LLC. ISSN: 2397-7523. URL: http://dx.doi.org/10.1186/s41512-023-00163-z.

[156] Dong Yang et al. "Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan". In: *Medical Image Analysis* 70 (2021). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101010488%5C&doi=10.1016%5C%2fj.media.2021.101992%5C&partnerID=40%5C&md5=b7dd8e290e5ece2833e5bf186972c994.

[157] Qinghao Ye et al. "Robust weakly supervised learning for COVID-19 recognition using multi-center CT images". In: *Applied Soft Computing* 116 (2022). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121314601%5C&doi=10.1016%5C%2fj.asoc.2021.108291%5C&partnerID=40%5C&md5=d26b6a6c16b36e2c9d2e02ca71b8f082.

[158] Kun Xiang et al. "Toward Robust Diagnosis: A Contour Attention Preserving Adversarial Defense for COVID-19 Detection". In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*. Vol. 37. Type: Conference paper. 2023, pp. 2928–2937. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85167962782%5C&partnerID=40%5C&md5=d0d4fa83d70752c50fe89c95fe8c24eb.

[159] Stefanos Zervoudakis et al. "Prediction of COVID-19 Infection Based on Symptoms and Social Life Using Machine Learning Techniques". In: *ACM International Conference Proceeding Series*. Conference paper. 2021, pp. 277–283. DOI: 10.1145/3453892.3462696.

[160] Jean Feng et al. "Monitoring machine learning-based risk prediction algorithms in the presence of performativity". In: *Proceedings of Machine Learning Research*. Vol. 238. Type: Conference paper. 2024, pp. 919–927. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85194159877%5C&partnerID=40%5C&md5=183ca9085a57fe82a47bc1b63ad855b9.

[161] Kueiyu Joshua Lin et al. "Prospective validation of a dynamic prognostic model for identifying <scp>COVID</scp>-19 patients at high risk of rapid deterioration". In: *Pharmacoepidemiology and Drug Safety* 32.5 (Dec. 2022). Publisher: Wiley, pp. 545–557. ISSN: 1099-1557. URL: http://dx.doi.org/10.1002/pds.5580.

[162] Teo Susnjak et al. "Forecasting patient flows with pandemic induced concept drift using explainable machine learning". In: *EPJ Data Science* 12.1 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85154069744%5C&doi=10.1140%5C%2fepjds%5C%2fs13688-023-00387-5%5C&partnerID=40%5C&md5=1f003c157d2e605ce8fddf87248c2660.

[163] Christopher Duckworth et al. "Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19". In: *Scientific Reports* 11.1 (2021). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119841842%5C&doi=10.1038%5C%2fs41598-021-02481-y%5C&partnerID=40%5C&md5=f07edfd2e428a9aaca4ecef8d5cf8516.

[164] Andrew B. Sellergren et al. "Simplified Transfer Learning for Chest Radiography Models Using Less Data". In: *Radiology* 305.2 (2022). Type: Article, pp. 454–465. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138217311%5C&doi=10.1148%5C%2fradiol.212482%5C&partnerID=40%5C&md5=2f5af1f1d25bbb67727ab5b451824fd9.

[165] Yuanyi Feng et al. "Cross-platform privacy-preserving CT image COVID-19 diagnosis based on source-free domain adaptation". In: *Knowledge-Based Systems* 264 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147848362%5C&doi=10.1016%5C%2fj.knosys.2023.110324%5C&partnerID=40%5C&md5=80c1e6a36d3680e009b9b80dc9f6902f.

[166] Theofanis Ganitidis et al. "A Comprehensive Drift-Adaptive Framework for Sustaining Model Performance in COVID-19 Detection From Dynamic Cough Audio Data: Model Development and Validation". In: *Journal of Medical Internet Research* 27 (2025), e66919.

[167] Vallijah Subasri et al. “Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models”. In: *JAMA Network Open* 8.6 (2025), e2513685–e2513685.

[168] Giuseppe Tricomi et al. “HERALD: A Hybrid distributEd leaRning incrementAL & feDerated solution for knowledge distillation in COVID-19 classification”. In: *Future Generation Computer Systems* 174 (2026), p. 107991.

[169] Fernanda Cristina Barbosa Lana et al. “Unraveling relevant cross-waves pattern drifts in patient-hospital risk factors among hospitalized COVID-19 patients using explainable machine learning methods”. In: *BMC Infectious Diseases* 25.1 (2025), p. 537.

[170] Ru Geng et al. “Analysis of the Spatio-Temporal Dynamics of COVID-19 in Massachusetts via Spectral Graph Wavelet Theory”. In: *IEEE Transactions on Signal and Information Processing over Networks* 8 (2022). Type: Article, pp. 670–683. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135242371%5C&doi=10.1109%5C%2fTSIPN.2022.3193252%5C&partnerID=40%5C&md5=2839e4891d8e6732db9039664905721f.

[171] Helena Baptista et al. “Similarity- and neighbourhood-based dynamic models for infection data: Uncovering the complexities of the COVID-19 infection risks”. In: *Spatial and Spatio-temporal Epidemiology* 51 (2024). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203145530%5C&%20doi=10.1016%5C%2fj.sste.2024.100681%5C&%20partnerID=40%5C&%20md5=fadd35fca2b9aaae545aff3811a7a974.

[172] Kang Heng Lim et al. “Enhancing Emergency Department Management: A Data-Driven Approach to Detect and Predict Surge Persistence”. In: *Healthcare (Switzerland)* 12.17 (2024). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203694230%5C&doi=10.3390%5C%2fhealthcare12171751%5C&partnerID=40%5C&md5=e1ed5738a97ed385aece30f0358b8626.

[173] Yuchen Liang et al. “Non-parametric quickest detection of a change in the mean of an observation sequence”. In: *2021 55th Annual Conference on Information Sciences and Systems, CISS 2021*. Type: Conference paper. 2021. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104981309%5C&doi=10.1109%5C%2fCISS50987.2021.9400252%5C&partnerID=40%5C&md5=b83779dc920ff4a170f14f448c490712.

[174] Ari J Smith et al. “Multisite evaluation of prediction models for emergency department crowding before and during the COVID-19 pandemic”. In: *Journal of the American Medical Informatics Association* 30.2 (Oct. 2022). Publisher: Oxford University Press (OUP), 292â€“300. ISSN: 1527-974X. URL: http://dx.doi.org/10.1093/jamia/ocac214.

[175] Alessio Bernardo et al. "Predict COVID-19 Spreading with C-SMOTE". In: *Business Information Systems*. Vol. 1. Type: Conference paper. 2021, pp. 27–38. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129518854%5C&%20doi=10.52825%5C%2fbis.v1i.45%5C&%20partnerID=40%5C&%20md5=78a8bba7660ebd31ed00e330968b5048.

[176] Atiye Sadat Hashemi et al. "Surveillance of Disease Outbreaks Using Unsupervised Uni-Multivariate Anomaly Detection of Time-Series Symptoms". In: *Studies in health technology and informatics* 316 (2024). Type: Article, pp. 1916–1920. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85202005899%5C&doi=10.3233%5C%2fSHTI240807%5C&partnerID=40%5C&md5=4c5a38685e48f56b458b8921d587d

[177] Paola Sebastiani et al. "A Bayesian dynamic model for influenza surveillance". In: *Statistics in Medicine* 25.11 (2006). Type: Article, pp. 1803–1816. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-33744825430%5C&doi=10.1002%5C%2fsim.2566%5C&partnerID=40%5C&md5=da4aafcf43c892b78d5bc01ff55a9aa8.

[178] Masoud Asadi-Zeydabadi et al. "Analysis of COVID-19 pandemic in USA, using Topological Weighted Centroid". In: *Comput. Biol. Med.* 136.C (Sept. 2021). Place: USA Publisher: Pergamon Press, Inc. ISSN: 0010-4825. URL: https://doi.org/10.1016/j.compbiomed.2021.104670.

[179] Lander Willem et al. "Active Learning to Understand Infectious Disease Models and Improve Policy Making". In: *PLoS Computational Biology* 10.4 (2014). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84901335905%5C&doi=10.1371%5C%2fjournal.pcbi.1003563%5C&partnerID=40%5C&md5=00e6914cf51b1e5c16677335f4dd9a10.

[180] Anna K. Lugnér et al. "Dynamic versus static models in cost-effectiveness analyses of anti-viral drug therapy to mitigate an influenza pandemic". In: *Health Economics* 19.5 (2010). Type: Article, pp. 518–531. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77951727815%5C&doi=10.1002%5C%2fhec.1485%5C&partnerID=40%5C&md5=fddb8740f3d46f1b2020854d3d0b263e.

[181] Yiruo Lu et al. "Hospital Beds Planning and Admission Control Policies for COVID-19 Pandemic: A Hybrid Computer Simulation Approach". In: *IEEE International Conference on Automation Science and Engineering*. Vol. 2021-August. Type: Conference paper. 2021, pp. 956–961. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116996604%5C&doi=10.1109%5C%2fCASE49439.2021.9551589%5C&partnerID=40%5C&md5=20f20bb4c266c255ceb15b6695bbd56b.

[182] Zohreh Abbasi et al. "Vaccination and isolation based control design of the COVID-19 pandemic based on adaptive neuro fuzzy inference system optimized with the genetic algorithm". In: *Evolving Systems* 14.3 (Sept. 2022). Publisher: Springer Science and Business Media LLC, 413â€"435. ISSN: 1868-6486. URL: http://dx.doi.org/10.1007/s12530-022-09459-9.

[183] Lin-Tao Fu et al. "Spatiotemporal dynamics of confirmed case distribution during the COVID-19 pandemic in China: data comparison between 2020/04–2020/08 and 2021/04–2021/08". In: *Scientific Reports* 13.1 (2023). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165550736%5C&doi=10.1038%5C%2fs41598-023-39139-w%5C&partnerID=40%5C&md5=1381ec4a9f4ac573863259f7f5943d84.

[184] Stephen Edward Moore et al. "CONTROLLING THE TRANSMISSION DYNAMICS OF COVID-19". In: *Communications in Mathematical Biology and Neuroscience* 2022 (2022). Type: Article. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122991096%5C&doi=10.28919%5C%2fcmbn%5C%2f6792%5C&partnerID=40%5C&md5=fb3e206f2beb2f70de1278e44bcef0cb.

[185] Sarasij Banerjee et al. "A Data-Driven Hybrid Model Predictive Control Framework for Managing Epidemics Using 3DoF-KF HMPC". In: *2025 American Control Conference (ACC)*. IEEE. 2025, pp. 2023–2028.

[186] Zuiyuan Guo et al. "Quantitatively analyzing the relationship between non-pharmaceutical interventions and the direction of virus evolution using a dynamic model". In: *Frontiers in Public Health* 13 (2025), p. 1542759.

[187] John PA Ioannidis et al. "Forecasting for COVID-19 has failed". In: *International journal of forecasting* 38.2 (2022), pp. 423–438.

[188] Giulio Viceconte et al. "COVID-19 R0: Magic number or conundrum?" In: *Infectious disease reports* 12.1 (2020), p. 8516.

[189] Lin Liu. "Fields Virology, 6th Edition". In: *Clinical Infectious Diseases* 59.4 (May 2014), pp. 613–613. ISSN: 1058-4838. eprint: https://academic.oup.com/cid/article-pdf/59/4/613/9412985/ciu346.pdf. URL: https://doi.org/10.1093/cid/ciu346.

[190] Ian Goodfellow et al. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[191] Andriy Burkov. *Machine learning engineering*. Vol. 1. True Positive Incorporated Montreal, QC, Canada, 2020.

[192] Fred Brauer. "Compartmental models in epidemiology". In: *Mathematical epidemiology* (2008), pp. 19–79.

[193] Hugh Durrant-Whyte et al. "Introduction to estimation and the Kalman filter". In: *Australian Centre for Field Robotics* 28.3 (2001), pp. 65–94.

[194] Andreas Holzinger et al. "Causability and explainability of artificial intelligence in medicine". In: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 9.4 (2019), e1312.

[195] Scott Lundberg. "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874* (2017).

[196] Suruchi Kumari et al. "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives". In: *Computers in Biology and Medicine* 170 (2024), p. 107912.

[197] Berkman Sahiner et al. "Data drift in medical machine learning: implications and potential remedies". In: *The British Journal of Radiology* 96.1150 (2023), p. 20220878.

[198] Ketan Sanjay Desale et al. "Addressing concept drifts using deep learning for heart disease prediction: a review". In: *Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021*. Springer. 2021, pp. 157–167.

[199] Antonio Guillén-Teruel et al. "Identifying risk factors and predicting long COVID in a Spanish cohort". In: *Scientific Reports* 15.1 (2025), p. 10758.

[200] Erik Von Elm et al. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies". In: *The lancet* 370.9596 (2007), pp. 1453–1457.

[201] Centers for Disease Control et al. *Post-COVID Conditions*. 2023. URL: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html.

[202] César Fernández-de-las-Peñas et al. "Defining Post-COVID Symptoms (Post-Acute COVID, Long COVID, Persistent Post-COVID): An Integrative Classification". In: *International Journal of Environmental Research and Public Health* 18.5 (Mar. 5, 2021), p. 2621. ISSN: 1660-4601. URL: https://www.mdpi.com/1660-4601/18/5/2621 (visited on 02/05/2024).

[203] Alejandro Cisterna-García et al. "A predictive model for hospitalization and survival to COVID-19 in a retrospective population-based study". In: *Scientific Reports* 12.1 (2022), p. 18126.

[204] Francesca Bai et al. "Female gender is associated with long COVID syndrome: a prospective cohort study". In: *Clinical Microbiology and Infection* 28.4 (Apr. 2022), 611.e9–611.e16. ISSN: 1198743X. URL: https://linkinghub.elsevier.com/retrieve/pii/S1198743X21006297 (visited on 02/28/2024).

[205] César Fernández-de-las-Peñas et al. "Female Sex Is a Risk Factor Associated with Long-Term Post-COVID Related-Symptoms but Not with COVID-19 Symptoms: The LONG-COVID-EXP-CM Multicenter Study". In: *Journal of Clinical Medicine* 11.2 (Jan. 14, 2022), p. 413. ISSN: 2077-0383. URL: https://www.mdpi.com/2077-0383/11/2/413 (visited on 05/21/2024).

[206] A. Rodríguez Onieva et al. "Long COVID: Factors influencing persistent symptoms and the impact of gender". In: *Medicina de Familia. SEMERGEN* 50.5 (July

2024), p. 102208. ISSN: 11383593. URL: https://linkinghub.elsevier.com/retrieve/pii/S1138359324000182 (visited on 05/21/2024).

[207] Jennifer Cohen et al. "An intersectional analysis of long COVID prevalence". In: *International Journal for Equity in Health* 22.1 (Dec. 13, 2023), p. 261. ISSN: 1475-9276. URL: https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-023-02072-5 (visited on 05/21/2024).

[208] Ali Akbar Asadi-Pooya et al. "Risk Factors Associated with Long COVID Syndrome: A Retrospective Study". In: *Iranian Journal of Medical Sciences* 46.6 (Nov. 2021). URL: https://doi.org/10.30476/ijms.2021.92080.2326 (visited on 05/21/2024).

[209] Michał Chudzik et al. "Long-COVID clinical features and risk factors: a retrospective analysis of patients from the STOP-COVID registry of the PoLoCOV study". In: *Viruses* 14.8 (2022), p. 1755.

[210] Virasakdi Chongsuvivatwong. *epiDisplay: Epidemiological Data Display Package*. R package version 3.5.0.2. 2022. URL: https://CRAN.R-project.org/package=epiDisplay.

[211] Max Kuhn. "Building predictive models in R using the caret package". In: *Journal of statistical software* 28 (2008), pp. 1–26. DOI: 10.18637/jss.v028.i05.

[212] Hyun Kang. "The prevention and handling of the missing data". In: *Korean Journal of Anesthesiology* 64.5 (May 2013), pp. 402–406. ISSN: 2005-6419. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/ (visited on 02/12/2025).

[213] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[214] Diego Fano-Sizgorich et al. "Risk of death, hospitalization and intensive care unit admission by SARS-CoV-2 variants in Peru: a retrospective study". In: *International Journal of Infectious Diseases* 127 (Feb. 2023), pp. 144–149. ISSN: 12019712. URL: https://linkinghub.elsevier.com/retrieve/pii/S1201971222006579 (visited on 05/24/2024).

[215] Pontus Hedberg et al. "In-hospital mortality during the wild-type, alpha, delta, and omicron SARS-CoV-2 waves: a multinational cohort study in the EuCARE project". In: *The Lancet Regional Health - Europe* 38 (Mar. 2024), p. 100855. ISSN: 26667762. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666776224000218 (visited on 05/24/2024).

[216] Thomas D. Hughes et al. "The effect of SARS-CoV-2 variant on respiratory features and mortality". In: *Scientific Reports* 13.1 (Mar. 18, 2023), p. 4503. ISSN: 2045-2322. URL: https://www.nature.com/articles/s41598-023-31761-y (visited on 05/24/2024).

[217] Siwen Wang et al. "Associations of Depression, Anxiety, Worry, Perceived Stress, and Loneliness Prior to Infection With Risk of Post–COVID-19 Conditions". In: *JAMA Psychiatry* 79.11 (Nov. 1, 2022), p. 1081. ISSN: 2168-622X. URL: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2796097 (visited on 05/21/2024).

[218] Luigi Vimercati et al. "Association between Long COVID and Overweight/Obesity". In: *Journal of Clinical Medicine* 10.18 (Sept. 14, 2021), p. 4143. ISSN: 2077-0383. URL: https://www.mdpi.com/2077-0383/10/18/4143 (visited on 05/21/2024).

[219] Maciej Debski et al. "Post-COVID-19 syndrome risk factors and further use of health services in East England". In: *PLOS Global Public Health* 2.11 (Nov. 30, 2022). Ed. by Janelisa Musaya, e0001188. ISSN: 2767-3375. URL: https://dx.plos.org/10.1371/journal.pgph.0001188 (visited on 05/21/2024).

[220] Mark W. Tenforde et al. "Symptom Duration and Risk Factors for Delayed Return to Usual Health Among Outpatients with COVID-19 in a Multistate Health Care Systems Network — United States, March–June 2020". In: *MMWR. Morbidity and Mortality Weekly Report* 69.30 (July 31, 2020), pp. 993–998. ISSN: 0149-2195, 1545-861X. URL: http://www.cdc.gov/mmwr/volumes/69/wr/mm6930e1.htm?s_cid=mm6930e1_w (visited on 05/21/2024).

[221] Anuradhaa Subramanian et al. "Symptoms and risk factors for long COVID in non-hospitalized adults". In: *Nature Medicine* 28.8 (Aug. 2022), pp. 1706–1714. ISSN: 1078-8956, 1546-170X. URL: https://www.nature.com/articles/s41591-022-01909-w (visited on 02/05/2024).

[222] Kim Daniel Jakobsen et al. "Machine learning identifies risk factors associated with long-term sick leave following COVID-19 in Danish population". In: *Communications Medicine* 3.1 (Dec. 20, 2023), p. 188. ISSN: 2730-664X. URL: https://www.nature.com/articles/s43856-023-00423-5 (visited on 05/23/2024).

[223] Qiang Zhang et al. "COVID-19 Case Fatality and Alzheimer's Disease". In: *Journal of Alzheimer's Disease* 84.4 (Dec. 7, 2021), pp. 1447–1452. ISSN: 13872877, 18758908. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JAD-215161 (visited on 05/23/2024).

[224] Ana C. Tahira et al. "Dementia is an age-independent risk factor for severity and death in COVID-19 inpatients". In: *Alzheimer's & Dementia* 17.11 (Nov. 2021), pp. 1818–1831. ISSN: 1552-5260, 1552-5279. URL: https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12352 (visited on 05/23/2024).

[225] Naciye Magusali et al. "A genetic link between risk for Alzheimer's disease and severe COVID-19 outcomes via the *OAS1* gene". In: *Brain* 144.12 (Dec. 31, 2021), pp. 3727–3741. ISSN: 0006-8950, 1460-2156. URL: https://academic.oup.com/brain/article/144/12/3727/6382473 (visited on 05/23/2024).

[226] the Management Group of the EAN Dementia and Cognitive Disorders Scientific Panel et al. "Dementia and COVID-19, a Bidirectional Liaison: Risk Factors, Biomarkers, and Optimal Health Care". In: *Journal of Alzheimer's Disease* 82.3 (Aug. 3, 2021). Ed. by Laura Bonanni, pp. 883–898. ISSN: 13872877, 18758908. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JAD-210335 (visited on 05/23/2024).

[227] Dan Shan et al. *Temporal Association between COVID-19 Infection and Subsequent New-Onset Dementia in Older Adults: A Systematic Review and Meta-Analysis*. 2024. URL: https://www.ssrn.com/abstract=4716751 (visited on 05/23/2024).

[228] Nancy A. Hodgson et al. "Undiagnosed Illness and Neuropsychiatric Behaviors in Community Residing Older Adults With Dementia". In: *Alzheimer Disease & Associated Disorders* 25.2 (Apr. 2011), pp. 109–115. ISSN: 0893-0341. URL: https://journals.lww.com/00002093-201104000-00002 (visited on 05/23/2024).

[229] Minna K. Löppönen et al. "Undiagnosed Diseases in Patients with Dementia – A Potential Target Group for Intervention". In: *Dementia and Geriatric Cognitive Disorders* 18.3 (2004), pp. 321–329. ISSN: 1420-8008, 1421-9824. URL: https://karger.com/DEM/article/doi/10.1159/000080126 (visited on 05/23/2024).

[230] Carolyn T. Thorpe et al. "Receipt of Monitoring of Diabetes Mellitus in Older Adults with Comorbid Dementia". In: *Journal of the American Geriatrics Society* 60.4 (Apr. 2012), pp. 644–651. ISSN: 0002-8614, 1532-5415. URL: https://agsjournals.onlinelibrary.wiley.com/doi/10.1111/j.1532-5415.2012.03907.x (visited on 05/23/2024).

[231] Carole H Sudre et al. "Attributes and predictors of long COVID". In: *Nature medicine* 27.4 (2021), pp. 626–631.

[232] Rosa Morello et al. "Risk factors for post-COVID-19 condition (Long Covid) in children: a prospective cohort study". In: *EClinicalMedicine* 59 (2023).

[233] Liam Townsend et al. "Persistent fatigue following SARS-CoV-2 infection is common and independent of severity of initial infection". In: *PLOS ONE* 15.11 (Nov. 9, 2020). Ed. by Giordano Madeddu, e0240784. ISSN: 1932-6203. URL: https://dx.plos.org/10.1371/journal.pone.0240784 (visited on 05/24/2024).

[234] Oscar Moreno-Pérez et al. "Post-acute COVID-19 syndrome. Incidence and risk factors: A Mediterranean cohort study". In: *Journal of Infection* 82.3 (Mar. 2021), pp. 378–383. ISSN: 01634453. URL: https://linkinghub.elsevier.com/retrieve/pii/S0163445321000098 (visited on 05/24/2024).

[235] Yu-Hui Liu et al. "One-Year Trajectory of Cognitive Changes in Older Survivors of COVID-19 in Wuhan, China: A Longitudinal Cohort Study". In: *JAMA Neurology* 79.5 (May 1, 2022), p. 509. ISSN: 2168-6149. URL: https://jamanetwork.com/journals/jamaneurology/fullarticle/2789919 (visited on 05/24/2024).

[236] Ziyad Al-Aly et al. "High-dimensional characterization of post-acute sequelae of COVID-19". In: *Nature* 594.7862 (June 10, 2021), pp. 259–264. ISSN: 0028-0836, 1476-4687. URL: https://www.nature.com/articles/s41586-021-03553-9 (visited on 05/24/2024).

[237] Ziyad Al-Aly et al. "Long COVID after breakthrough SARS-CoV-2 infection". In: *Nature Medicine* 28.7 (July 2022), pp. 1461–1467. ISSN: 1078-8956, 1546-170X. URL: https://www.nature.com/articles/s41591-022-01840-0 (visited on 05/24/2024).

[238] Yoonjung Kim et al. "Long COVID prevalence and impact on quality of life 2 years after acute COVID-19". In: *Scientific Reports* 13.1 (July 11, 2023), p. 11207. ISSN: 2045-2322. URL: https://www.nature.com/articles/s41598-023-36995-4 (visited on 05/24/2024).

[239] Vasiliki Tsampasian et al. "Risk Factors Associated With Post- COVID-19 Condition: A Systematic Review and Meta-analysis". In: *JAMA Internal Medicine* 183.6 (June 1, 2023), p. 566. ISSN: 2168-6106. URL: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2802877 (visited on 05/24/2024).

[240] Kin Israel Notarte et al. "Impact of COVID-19 vaccination on the risk of developing long-COVID and on existing long-COVID symptoms: A systematic review". In: *eClinicalMedicine* 53 (Nov. 2022), p. 101624. ISSN: 25895370. URL: https://linkinghub.elsevier.com/retrieve/pii/S2589537022003546 (visited on 02/23/2024).

[241] Atsuyuki Watanabe et al. "Protective effect of COVID-19 vaccination against long COVID syndrome: A systematic review and meta-analysis". In: *Vaccine* 41.11 (Mar. 2023), pp. 1783–1790. ISSN: 0264410X. URL: https://linkinghub.elsevier.com/retrieve/pii/S0264410X23001342 (visited on 05/16/2024).

[242] Felicia Ceban et al. "COVID-19 vaccination for the prevention and treatment of long COVID: A systematic review and meta-analysis". In: *Brain, Behavior, and Immunity* 111 (July 2023), pp. 211–229. ISSN: 08891591. URL: https://linkinghub.elsevier.com/retrieve/pii/S088915912300079X (visited on 05/16/2024).

[243] M. Daniel Brannock et al. "Long COVID risk and pre-COVID vaccination in an EHR-based cohort study from the RECOVER program". In: *Nature Communications* 14.1 (May 22, 2023), p. 2914. ISSN: 2041-1723. URL: https://www.nature.com/articles/s41467-023-38388-7 (visited on 05/16/2024).

[244] Martí Català et al. "The effectiveness of COVID-19 vaccines to prevent long COVID symptoms: staggered cohort study of data from the UK, Spain, and Estonia". In: *The Lancet Respiratory Medicine* (Jan. 2024), S2213260023004149. ISSN: 22132600. URL: https://linkinghub.elsevier.com/retrieve/pii/S2213260023004149 (visited on 02/23/2024).

[245] Nhung Th Trinh et al. "Effectiveness of COVID-19 vaccines to prevent long COVID: data from Norway". In: *The Lancet Respiratory Medicine* 12.5 (May 2024), e33–e34. ISSN: 22132600. URL: https://linkinghub.elsevier.com/retrieve/pii/S2213260024000821 (visited on 05/16/2024).

[246] Alexandre R. Marra et al. "The effectiveness of COVID-19 vaccine in the prevention of post-COVID conditions: a systematic literature review and meta-analysis of the latest research". In: *Antimicrobial Stewardship & Healthcare Epidemiology* 3.1 (2023), e168. ISSN: 2732-494X. URL: https://www.cambridge.org/core/product/identifier/S2732494X23004473/type/journal_article (visited on 05/20/2024).

[247] Lisa Lundberg-Morris et al. "Covid-19 vaccine effectiveness against post-covid-19 condition among 589 722 individuals in Sweden: population based cohort study". In: *BMJ* (Nov. 22, 2023), e076990. ISSN: 1756-1833. URL: https://www.bmj.com/lookup/doi/10.1136/bmj-2023-076990 (visited on 05/20/2024).

[248] Marit J. Van Gils et al. "Antibody responses against SARS-CoV-2 variants induced by four different SARS-CoV-2 vaccines in health care workers in the Netherlands: A prospective cohort study". In: *PLOS Medicine* 19.5 (May 17, 2022). Ed. by James G. Beeson, e1003991. ISSN: 1549-1676. URL: https://dx.plos.org/10.1371/journal.pmed.1003991 (visited on 05/20/2024).

[249] Marzieh Soheili et al. "The efficacy and effectiveness of COVID-19 vaccines around the world: a mini-review and meta-analysis". In: *Annals of Clinical Microbiology and Antimicrobials* 22.1 (May 19, 2023), p. 42. ISSN: 1476-0711. URL: https://annclinmicrob.biomedcentral.com/articles/10.1186/s12941-023-00594-y (visited on 05/20/2024).

[250] Elena Azzolini et al. "Association Between BNT162b2 Vaccination and Long COVID After Infections Not Requiring Hospitalization in Health Care Workers". In: *JAMA* 328.7 (Aug. 2022), pp. 676–678. ISSN: 0098-7484. eprint: https://jamanetwork.com/journals/jama/articlepdf/2794072/jama_azzolini_2022_ld_220046_1660585244.20049.pdf. URL: https://doi.org/10.1001/jama.2022.11691.

[251] Lisa Lundberg-Morris et al. "Covid-19 vaccine effectiveness against post-covid-19 condition among 589 722 individuals in Sweden: population based cohort study". In: *BMJ* 383 (2023). eprint: https://www.bmj.com/content/383/bmj-2023-076990.full.pdf. URL: https://www.bmj.com/content/383/bmj-2023-076990.

[252] Blessy Antony et al. "Predictive models of long COVID". In: *eBioMedicine* 96 (Oct. 2023), p. 104777. ISSN: 23523964. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352396423003432 (visited on 05/23/2024).

[253] Roman Kessler et al. "Predictive Attributes for Developing Long COVID—A Study Using Machine Learning and Real-World Data from Primary Care Physicians

in Germany". In: *Journal of Clinical Medicine* 12.10 (May 17, 2023), p. 3511. ISSN: 2077-0383. URL: https://www.mdpi.com/2077-0383/12/10/3511 (visited on 05/23/2024).

[254] Jock Blackard. *Covertype*. UCI Machine Learning Repository. 1998. DOI: 10.24432/C50K5N.

[255] W Nick Street et al. "A streaming ensemble algorithm (SEA) for large-scale classification". In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 2001, pp. 377–382.

[256] Geoff Hulten et al. "Mining time-changing data streams". In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 2001, pp. 97–106.

[257] Albert Bifet et al. "MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering". In: *Proceedings of the First Workshop on Applications of Pattern Analysis*. Ed. by Tom Diethe et al. Vol. 11. Proceedings of Machine Learning Research. Cumberland Lodge, Windsor, UK: PMLR, 2010, pp. 44–50. URL: https://proceedings.mlr.press/v11/bifet10a.html.

[258] Albert Bifet et al. *Machine Learning for Data Streams with Practical Examples in MOA*. https://moa.cms.waikato.ac.nz/book/. MIT Press, 2018.

[259] R. Agrawal et al. "Database mining: a performance perspective". In: *IEEE Transactions on Knowledge and Data Engineering* 5.6 (1993), pp. 914–925. DOI: 10.1109/69.250074.

[260] URL: https://zenodo.org/records/11469702.

[261] Barry Becker et al. *Adult*. UCI Machine Learning Repository. 1996. DOI: doi.org/10.24432/C5XW20.

[262] Ashwin Srinivasan. *Statlog (Landsat Satellite)*. UCI Machine Learning Repository. 1993. DOI: 10.24432/C55887.

[263] Catherine L Blake. "UCI repository of machine learning databases". In: *http://www.ics. uci. edu/~ mlearn/MLRepository. html* (1998).

[264] Kevin S Woods et al. "Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography". In: *International Journal of Pattern Recognition and Artificial Intelligence* 7.06 (1993), pp. 1417–1436. DOI: 10.1142/S0218001493000698.

[265] P Alinat. "Periodic progress report 4, ROARS project ESPRIT II-Number 5516". In: *Technical Thomson Report TS ASM 93/S/EGS/NC* 79 (1993).

[266] Joaquin Vanschoren et al. "OpenML: networked science in machine learning". In: *ACM SIGKDD Explorations Newsletter* 15.2 (2014), pp. 49–60.

[267] Ryan Elwell et al. "Incremental learning of concept drift in nonstationary environments". In: *IEEE transactions on neural networks* 22.10 (2011), pp. 1517–1531.

[268] Alexey Tsymbal et al. "Dynamic integration of classifiers for handling concept drift". In: *Information fusion* 9.1 (2008), pp. 56–68.

[269] Nitesh V Chawla et al. "Special issue on learning from imbalanced data sets". In: *ACM SIGKDD explorations newsletter* 6.1 (2004), pp. 1–6. DOI: 10.1145/1007730.1007733.

[270] Vicente García et al. "Theoretical analysis of a performance measure for imbalanced data". In: *2010 20th International Conference on Pattern Recognition*. IEEE. 2010, pp. 617–620.

[271] Jason Van Hulse et al. "Experimental perspectives on learning from imbalanced data". In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 935–942.

[272] Yuchun Tang et al. "SVMs Modeling for Highly Imbalanced Classification". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.1 (2009), pp. 281–288. DOI: 10.1109/TSMCB.2008.2002909.

[273] Tatjana Eitrich et al. "Efficient optimization of support vector machine learning parameters for unbalanced datasets". In: *Journal of computational and applied mathematics* 196.2 (2006), pp. 425–436. DOI: 10.1016/j.cam.2005.09.009.

[274] Maher Maalouf et al. "Weighted logistic regression for large-scale imbalanced and rare events data". In: *Knowledge-Based Systems* 59 (2014), pp. 142–148. DOI: 10.1016/j.knosys.2014.01.012.

[275] Paula Branco et al. "SMOGN: a pre-processing approach for imbalanced regression". In: *First international workshop on learning with imbalanced domains: Theory and applications*. PMLR. 2017, pp. 36–50.

[276] Xin Yue Song et al. "DistSMOGN: Distributed SMOGN for Imbalanced Regression Problems". In: *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR. 2022, pp. 38–52.

[277] Ni Wayan Surya Wardhani et al. "Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data". In: *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. 2019, pp. 14–18. DOI: 10.1109/IC3INA48034.2019.8949568.

[278] Philipp Thölke et al. "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data". In: *NeuroImage* 277 (2023), p. 120253. ISSN: 1053-8119. URL: https://www.sciencedirect.com/science/article/pii/S1053811923004044.

[279] Haibo He et al. "Learning from Imbalanced Data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[280] Amalia Luque et al. "The impact of class imbalance in classification performance metrics based on the binary confusion matrix". In: *Pattern Recognition* 91 (2019), pp. 216–231. DOI: 10.1016/j.patcog.2019.02.023.

[281] Roghayeh Soleymani et al. "F-measure curves: A tool to visualize classifier performance under imbalance". In: *Pattern Recognition* 100 (2020), p. 107146. DOI: 10.1016/j.patcog.2019.107146.

[282] Adane Nega Tarekegn et al. "A review of methods for imbalanced multi-label classification". In: *Pattern Recognition* 118 (2021), p. 107965. DOI: 10.1016/j.patcog.2021.107965.

[283] László A. Jeni et al. "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics". In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 245–251. DOI: 10.1109/ACII.2013.47.

[284] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[285] Jean-Gabriel Gaudreault et al. "Empirical analysis of performance assessment for imbalanced classification". In: *Machine Learning* (2024), pp. 1–43. DOI: 10.1007/s10994-023-06497-5.

[286] Jilun Tian et al. "A novel data augmentation approach to fault diagnosis with class-imbalance problem". In: *Reliability Engineering & System Safety* 243 (2024), p. 109832. DOI: 10.1016/j.ress.2023.109832.

[287] Fadi Thabtah et al. "Data imbalance in classification: Experimental evaluation". In: *Information Sciences* 513 (2020), pp. 429–441.

[288] Bin Liu et al. "Dealing with class imbalance in classifier chains via random undersampling". In: *Knowledge-Based Systems* 192 (2020), p. 105292. ISSN: 0950-7051. URL: https://www.sciencedirect.com/science/article/pii/S0950705119305830.

[289] Wonjae Lee et al. "Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning". In: *Big Data Research* 28 (2022), p. 100314. ISSN: 2214-5796. URL: https://www.sciencedirect.com/science/article/pii/S2214579622000089.

[290] Maram Alamri et al. "Survey of Credit Card Anomaly and Fraud Detection Using Sampling Techniques". In: *Electronics* 11.23 (2022), p. 4003. DOI: 10.3390/electronics11234003.

[291] Bing Zhu et al. "An empirical comparison of techniques for the class imbalance problem in churn prediction". In: *Information sciences* 408 (2017), pp. 84–99.

[292] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[293] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE. 2008, pp. 1322–1328.

[294] Duygu Selin Turan et al. "The incremental SMOTE: A new approach based on the incremental k-means algorithm for solving imbalanced data set problem". In: *Information Sciences* 711 (2025), p. 122103.

[295] Aimin Zhang et al. "SMOTE-RkNN: A hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors". In: *Information Sciences* 595 (2022), pp. 70–88.

[296] Baiyun Chen et al. "RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise". In: *Information Sciences* 553 (2021), pp. 397–428.

[297] Nicola Lunardon et al. "ROSE: a Package for Binary Imbalanced Learning". In: *The R Journal* 6.1 (2014), pp. 79–89. URL: https://doi.org/10.32614/RJ-2014-008.

[298] Pedro Domingos. "Metacost: A general method for making classifiers cost-sensitive". In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 1999, pp. 155–164.

[299] Kai Ming Ting. "An instance-weighting method to induce cost-sensitive trees". In: *IEEE Transactions on Knowledge and Data Engineering* 14.3 (2002), pp. 659–665. DOI: 10.1109/TKDE.2002.1000348.

[300] Jingjing Tang et al. "Multi-view cost-sensitive kernel learning for imbalanced classification problem". In: *Neurocomputing* 552 (2023), p. 126562. DOI: 10.1016/j.neucom.2023.126562.

[301] Jan Kremer et al. "Active learning with support vector machines". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.4 (2014), pp. 313–326. DOI: 10.1002/widm.1132.

[302] Ricky Maulana Fajri et al. "Fal-cur: fair active learning using uncertainty and representativeness on fair clustering". In: *Expert Systems with Applications* 242 (2024), p. 122842. DOI: 10.1016/j.eswa.2023.122842.

[303] Zahra Ferdowsi et al. "Online active learning with imbalanced classes". In: *2013 IEEE 13th International Conference on Data Mining*. IEEE. 2013, pp. 1043–1048.

[304] Zhong Chen et al. "An effective cost-sensitive sparse online learning framework for imbalanced streaming data classification and its application to online anomaly

detection". In: *Knowledge and Information Systems* 65.1 (2023), pp. 59–87. DOI: 10.1007/s10115-022-01745-x.

[305] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[306] Leo Breiman. "Bagging predictors". In: *Machine learning* 24 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[307] Shohei Hido et al. "Roughly balanced bagging for imbalanced data". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 2.5-6 (2009), pp. 412–426. DOI: 10.1002/sam.10061.

[308] Fithria Siti Hanifah et al. "Smotebagging algorithm for imbalanced dataset in logistic regression analysis (case: Credit of bank x)". In: *Applied Mathematical Sciences* 9.138 (2015), pp. 6857–6865. DOI: 10.12988/ams.2015.58562.

[309] Nitesh V Chawla et al. "SMOTEBoost: Improving prediction of the minority class in boosting". In: *Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7*. Springer. 2003, pp. 107–119.

[310] Xu-Ying Liu et al. "Exploratory Undersampling for Class-Imbalance Learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2009), pp. 539–550. DOI: 10.1109/TSMCB.2008.2007853.

[311] Shengnan Shi et al. "A hybrid imbalanced classification model based on data density". In: *Information Sciences* 624 (2023), pp. 50–67. DOI: 10.1016/j.ins.2022.12.046.

[312] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[313] Joy Paul Guilford. "The minimal phi coefficient and the maximal phi". In: *Educational and psychological measurement* 25.1 (1965), pp. 3–8. DOI: 10.1177/001316446502500101.

[314] Martin Binder et al. "Multi-objective hyperparameter tuning and feature selection using filter ensembles". In: *Proceedings of the 2020 genetic and evolutionary computation conference*. 2020, pp. 471–479.

[315] André Pfob et al. "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison". In: *BMC medical research methodology* 22.1 (2022), pp. 1–15. DOI: 10.1186/s12874-022-01758-8.

[316] Max Kuhn. "Building predictive models in R using the caret package". In: *Journal of statistical software* 28 (2008), pp. 1–26. DOI: 10.18637/jss.v028.i05.

[317] Jeffrey C Schlimmer et al. "Incremental learning from noisy data". In: *Machine learning* 1.3 (1986), pp. 317–354.

[318] Damian A Tamburri. "Sustainable mlops: Trends and challenges". In: *2020 22nd international symposium on symbolic and numeric algorithms for scientific computing (SYNASC)*. IEEE. 2020, pp. 17–23.

[319] Dominik Kreuzberger et al. "Machine learning operations (mlops): Overview, definition, and architecture". In: *IEEE access* 11 (2023), pp. 31866–31879.

[320] Haixun Wang et al. "Mining concept drift from data streams using ensemble classifiers". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 226–235.

[321] T Ryan Hoens et al. "Learning from streaming data with concept drift and imbalance: an overview". In: *Information Sciences* 181.10 (2011), pp. 1925–1944.

[322] Peng Cao et al. "SERA: Selectively recursive approach for incremental learning from imbalanced data streams". In: *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE. 2013, pp. 3487–3492.

[323] Gregory Ditzler et al. "Learn++. CDS: A classifier ensemble for incremental learning in nonstationary environments with unbalanced class distributions". In: *IEEE Transactions on Neural Networks and Learning Systems* 24.4 (2013), pp. 637–651.

[324] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[325] J Zico Kolter et al. "Dynamic weighted majority: An ensemble method for drifting concepts". In: *The Journal of Machine Learning Research* 8 (2007), pp. 2755–2790.

[326] Elham Nasarian et al. "Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework". In: *Information Fusion* (2024), p. 102412. URL: https://doi.org/10.1016/j.inffus.2024.102412.

[327] Abdolreza Nazemi et al. "Interpretable machine learning for creditor recovery rates". In: *Journal of Banking & Finance* 164 (2024), p. 107187. DOI: 10.1016/j.jbankfin.2024.107187.

[328] Tengfei Yang et al. "Interpretable probabilistic bayesian neural networks for cybersecurity intrusion detection". In: *IEEE Access* (2024). DOI: 10.1109/ACCESS.2024.3421330.

[329] Ana Victoria Ponce-Bobadilla et al. "Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development". In: *Clinical and Translational Science* 17.11 (2024), e70056. DOI: 10.1111/cts.70056.

[330] Wen-Dong Jiang et al. "RealExp: Decoupling correlation bias in Shapley values for faithful model interpretations". In: *Information Processing & Management* 62.4 (2025), p. 104153.

[331] Ahmed M Salih et al. "A perspective on explainable artificial intelligence methods: SHAP and LIME". In: *Advanced Intelligent Systems* 7.1 (2025), p. 2400304.

[332] Raphael Baudeu et al. "Are SHAP Values Biased Towards High-Entropy Features?" In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2022, pp. 418–433. DOI: 10.1007/978-3-031-23618-1_28.

[333] Yujia Chen et al. "Interpretable machine learning for imbalanced credit scoring datasets". In: *European Journal of Operational Research* 312.1 (2024), pp. 357–372. URL: https://doi.org/10.1016/j.ejor.2023.06.036.

[334] Firuz Kamalov et al. "Feature selection in imbalanced data". In: *Annals of Data Science* 10.6 (2023), pp. 1527–1541. DOI: 10.1007/s40745-021-00366-5.

[335] Fatih Gurcan et al. "Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis". In: *Cancers* 16.19 (2024), p. 3417. DOI: 10.3390/cancers16193417.

[336] Imane Araf et al. "Cost-sensitive learning for imbalanced medical data: a review". In: *Artificial Intelligence Review* 57.4 (2024), p. 80. DOI: 10.1007/s10462-023-10652-8.

[337] Mingxuan Liu et al. "Balanced background and explanation data are needed in explaining deep learning models with SHAP: An empirical study on clinical decision making". In: *arXiv preprint arXiv:2206.04050* (2022). URL: https://arxiv.org/abs/2206.04050.

[338] Matteo Ballegeer et al. "Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring". In: *European Journal of Operational Research* 326.3 (2025), pp. 630–640. URL: https://www.sciencedirect.com/science/article/abs/pii/S0377221725004230.

[339] Marco Tulio Ribeiro et al. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[340] Jeremy Goldwasser et al. "Statistical Significance of Feature Importance Rankings". In: *arXiv preprint arXiv:2401.15800* (2024). URL: https://arxiv.org/abs/2401.15800.

[341] Xuanxiang Huang et al. "The Inadequacy of Shapley Values for Explainability". In: *arXiv preprint arXiv:2302.08160* (2023). URL: https://arxiv.org/abs/2302.08160.

[342] Olivier Letoffe et al. "From SHAP Scores to Feature Importance Scores". In: *arXiv preprint arXiv:2405.11766* (2024). URL: https://arxiv.org/abs/2405.11766.

[343] Lloyd S. Shapley. "A value for n-person games". In: *Contributions to the Theory of Games*. Vol. 2.28. Princeton University Press, 1953, pp. 307–317. DOI: 10.1515/9781400829156-012.

[344] Scott M Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[345] Avanti Shrikumar et al. "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMlR. 2017, pp. 3145–3153.

[346] Min Zhu et al. "Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data". In: *IEEE Access* 6 (2018), pp. 4641–4652. DOI: 10.1109/ACCESS.2018.2789428.

[347] Gissel Velarde et al. "Tree boosting methods for balanced and imbalanced classification and their robustness over time in risk assessment". In: *Intelligent Systems with Applications* 22 (2024), p. 200354.

[348] Gissel Velarde et al. "Evaluating XGBoost for balanced and imbalanced data: application to fraud detection". In: *arXiv preprint arXiv:2303.15218* (2023).

[349] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice hall PTR, 1994.

[350] Trevor Hastie. *The elements of statistical learning: data mining, inference, and prediction*. 2009.

[351] S Rasoul Safavian et al. "A survey of decision tree classifier methodology". In: *IEEE transactions on systems, man, and cybernetics* 21.3 (1991), pp. 660–674.

[352] Fei Tony Liu et al. "Isolation forest". In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.

[353] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[354] Indrė Žliobaitė et al. "An Overview of Concept Drift Applications". In: *Big Data Analysis: New Algorithms for a New Society*. Ed. by Nathalie Japkowicz et al. Cham: Springer International Publishing, 2016, pp. 91–114. ISBN: 978-3-319-26989-4. URL: https://doi.org/10.1007/978-3-319-26989-4_4.

[355] Jakub Jakubowski et al. "Roll Wear Prediction in Strip Cold Rolling with Physics-Informed Autoencoder and Counterfactual Explanations". In: *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. 2022, pp. 1–10. DOI: 10.1109/DSAA54385.2022.10032357.

[356] Ananya Jana et al. *Global and Local Interpretation of black-box Machine Learning models to determine prognostic factors from early COVID-19 data*. 2021. arXiv: 2109.05087 [cs.LG]. URL: https://arxiv.org/abs/2109.05087.

[357] URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[358] Alejandro Cisterna-García et al. "A predictive model for hospitalization and survival to COVID-19 in a retrospective population-based study". In: *Scientific Reports* 12.1 (2022), p. 18126. URL: https://doi.org/10.1038/s41598-022-22547-9.

[359] Fabrizio Pecoraro et al. "The efficiency in the ordinary hospital bed management: A comparative analysis in four European countries before the COVID-19 outbreak". In: *Plos one* 16.3 (2021), e0248867. URL: https://doi.org/10.1371/journal.pone.0248867.

[360] A González-Castro et al. "First and second wave of coronavirus-19 disease: A comparative study in patients hospitalized in an ICU of a third-level university hospital". In: *Medicina intensiva* 46.3 (2021), p. 166. URL: https://doi.org/10.1016/j.medine.2021.12.008.

[361] Guihong Fan et al. "Decreased case fatality rate of COVID-19 in the second wave: a study in 53 countries or regions". In: *Transboundary and emerging diseases* 68.2 (2021), pp. 213–215. URL: https://doi.org/10.1111/tbed.13819.

[362] Marta Massó-Muratel et al. "Predictive factors for hospitalization in a cohort of primary healthcare patients with suspected COVID-19". In: *Atención Primaria Práctica* 5.1 (2023), p. 100168. URL: https://doi.org/10.1016/j.appr.2023.100168.

[363] Hanife Tuğçe Çağlar et al. "Delta variant effect on the clinical course of adolescent COVID-19 patients". In: *Pediatric Pulmonology* 58.1 (2023), pp. 115–121. URL: https://doi.org/10.1002/ppul.26166.

[364] Yoanna Slabakova et al. "SARS-CoV-2 variant-specific gastrointestinal symptoms of COVID-19: 2023 update". In: *Gastroenterology Insights* 14.4 (2023), pp. 431–445. URL: https://doi.org/10.3390/gastroent14040032.

This page intentionally left blank.