Human-Centred Machine Learning for Health and Cognitive Modelling

Tesis doctoral presentada por Amaia Pikatza Huerga

dentro del Programa de Doctorado en Ingeniería para la Sociedad de la Información y Desarrollo Sostenible

Dirigida por el Dr. Aitor Almeida y el Dr. Unai Zulaika

Deusto

Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Human-Centred Machine Learning for Health and Cognitive Modelling

Tesis doctoral presentada por Amaia Pikatza Huerga

dentro del Programa de Doctorado en Ingeniería para la Sociedad de la Información y Desarrollo Sostenible

Dirigida por el Dr. Aitor Almeida y el Dr. Unai Zulaika

El doctorando

PIKATZA HUERGA AMAIA - 45917543K

Firmado digitalmente por PIKATZA HUERGA AMAIA - 45917543K
Fecha: 2026.02.23 09:58:59 +01'00'

Los directores

ALMEIDA ESCONDRILLAS, AITOR (AUTENTICACIÓN)

Digitally signed by ALMEIDA ESCONDRILLAS, AITOR (AUTENTICACIÓN)
Date: 2026.02.23 10:09:51 +01'00'

UNAI ZULAIKA ZURIMENDI

Firmado digitalmente por UNAI ZULAIKA ZURIMENDI
Fecha: 2026.02.23 10:48:25 +01'00'

Bilbao, febrero de 2026

Human-Centred Machine Learning for Health and Cognitive Modelling

Author: Amaia Pikatza Huerga
Directors: Dr. Aitor Almeida and Dr. Unai Zulaika

Printed in Bilbao
First edition, February 2026

Para titxu.

Abstract

Machine learning is increasingly transforming research in healthcare, mental-health, and cognitive science by enabling predictive and adaptive systems that support human decision-making. Yet the most accurate models often remain opaque, limiting their interpretability and trustworthiness in domains where transparency is essential. This doctoral research proposes a unified methodological framework for explainable and multimodal machine learning, validated across four representative contexts: prediction of hospital readmission in patients with heart failure (ReIC), prediction of readmission in patients with multiple chronic conditions (RePluris), prediction of eating disorder risk and recovery, and multimodal assessment of creativity from drawings and textual titles.

The framework combines data preprocessing, imbalance management, multimodal feature integration, and embedded explainability through SHapley Additive exPlanations (SHAP) and attention-based analyses. Each case study adapts this structure to its domain while maintaining methodological coherence. In cardiovascular and multimorbidity settings, ensemble models integrating clinical, functional, and psychosocial indicators achieved up to 30% improvement in discrimination compared with traditional Cox and logistic regression baselines. In addition, they highlighted the prognostic value of frailty, anxiety, and depression. In the eating-disorder study, resilience and quality-of-life measures emerged as strong determinants of one-year outcomes, confirming that self-perception and adaptability are central to recovery prediction. The creativity assessment demonstrated that combining image and text embeddings provides complementary information: textual inputs enhanced sensitivity to originality, whereas visual features better captured elaboration and complexity.

Across all four domains, the integration of explainability within the modelling process produced systems that balance predictive precision with interpretability, enabling domain experts to trace and validate decision mechanisms. This thesis advances the methodological foundation of human-centred artificial intelligence by demonstrating that explicit incorporation of explainability enhances not only trust but also

model performance. The resulting approach contributes to the development of transparent, adaptive, and ethically grounded predictive systems applicable to complex clinical, psychological, and cognitive contexts.

Resumen

El aprendizaje automático está transformando cada vez más la investigación en el ámbito de la asistencia sanitaria, la salud mental y las ciencias cognitivas, al permitir el desarrollo de sistemas predictivos y adaptativos que respaldan la toma de decisiones humanas. Sin embargo, los modelos más precisos suelen seguir siendo opacos, lo que limita su interpretabilidad y fiabilidad en ámbitos en los que la transparencia es esencial. Esta investigación doctoral propone un marco metodológico unificado para el aprendizaje automático explicable y multimodal, validado en cuatro contextos representativos: predicción del reingreso hospitalario en pacientes con insuficiencia cardíaca (ReIC), predicción del reingreso en pacientes con múltiples enfermedades crónicas (RePluris), predicción del riesgo y la recuperación de trastornos alimentarios, y evaluación multimodal de la creatividad a partir de dibujos y títulos textuales.

El marco metodológico combina el preprocesamiento de datos, la gestión del desequilibrio, la integración de características multimodales y la explicabilidad integrada a través de SHapley Additive exPlanations (SHAP) y análisis basados en la atención. Cada caso de uso adapta esta estructura a su ámbito, manteniendo al mismo tiempo la coherencia metodológica. En entornos cardiovasculares y de multimorbilidad, los modelos que integran indicadores clínicos, funcionales y psicosociales lograron una mejora de hasta el 30 % en la discriminación en comparación con las bases de referencia tradicionales de Cox y regresión logística. Además, destacó el valor pronóstico de la fragilidad, la ansiedad y la depresión. En el estudio sobre los trastornos alimentarios, las medidas de resiliencia y calidad de vida se revelaron como fuertes determinantes de los resultados al cabo de un año, lo que confirma que la autopercepción y la adaptabilidad son fundamentales para predecir la recuperación. La evaluación de la creatividad demostró que la combinación de incrustaciones de imágenes y texto proporciona información complementaria: las entradas textuales mejoraron la sensibilidad a la originalidad, mientras que las características visuales captaron mejor la elaboración y la complejidad.

En los cuatro ámbitos, la integración de la explicabilidad en el proceso de modelización dio lugar a sistemas que equilibran la precisión

predictiva con la interpretabilidad, lo que permite a los expertos en la materia rastrear y validar los mecanismos de decisión. Esta tesis avanza en la base metodológica de la inteligencia artificial centrada en el ser humano al demostrar que la incorporación explícita de la explicabilidad mejora no solo la confianza, sino también el rendimiento del modelo. El enfoque resultante contribuye al desarrollo de sistemas predictivos transparentes, adaptables y con base ética, aplicables a contextos clínicos, psicológicos y cognitivos complejos.

Laburpena

Ikasketa automatikoa osasun fisikoaren, mentalaren eta zientzia kognitiboen esparruko ikerketa aldatzen ari da, giza erabakien laguntzaile diren sistema prediktiboak eta moldakorrak garatzea ahalbidetzen baitu. Hala ere, eredu zehatzenek opakuak izaten jarraitzen dute, soluzioen fidagarritasuna mugatuz, gardentasuna funtsezkoa den eremuetan. Doktorego ikerketa honek ikasketa automatiko azalgarri eta multimodalerako marko metodologiko bateratu bat proposatu, eta lau eremutan balioztatzen du: bihotz-gutxiegitasuna (ReIC) duten pazienteen berrospitaleratzearen iragarpena, patologia anitz (RePluris) dituzten pazienteen berrospitaleratzearen iragarpena, elikadura-nahasmenduen arrisku eta berreskurapenaren iragarpena, eta sormenaren ebaluazio multimodala testu eta irudietatik abiatuta.

Markoak datuen aurreprozesamendua, desorekaren kudeaketa, ezaugarri multimodalen integrazioa eta azalgarritasun integratua konbinatzen ditu. Kasu-azterketa bakoitzak egitura hori bere eremura egokitzen du, koherentzia metodologikoa mantenduz. Patologia aniztasuneko eta bihotz-gutxiegitasuneko eremuetan, adierazle klinikoak, funtzionalak eta psikosozialak erabiltzen ditu ereduak. Horrela, %30erainoko hobekuntza lortu zen diskriminazioan, Cox-en erreferentzia-oinarri tradizionalekin eta erregresio logistikoarekin alderatuta. Horrez gain, hauskortasuna, antsietatea eta depresioa nabarmendu ziren iragarle nagusi bezala. Elikadura-nahasmenduei buruzko ikerketan, erresilientzia eta bizi-kalitatea agertu ziren gaitza sufritzeko arriskuaren eta berreskurapenaren iragarle nagusi bezala. Honek berresten du autopertzepzioa eta moldagarritasuna funtsezkoak direla errekuperazioa aurreikusteko. Sormenaren ebaluazioak erakutsi zuen irudiak eta testuak inkrustatzeak informazio osagarria ematen duela: testuek originaltasunarekiko sentsibilitatea hobetu zuten, eta irudiek, berriz, elaborazioa eta konplexutasuna.

Modelizazio-prozesuan azalgarritasuna sartzeak iragarpen eta interpretazio doitasuna orekatzea ahalbidetu du lau eremuetan. Horrek arloko adituei erabaki-mekanismoak arakatzeko eta baliozkotzeko aukera ematen die. Gizakia ardatz bezala hartuta, adimen artifizialaren oinarri metodologikoan aurrera egiten du tesi honek. Azalgarritasuna esplizituki txertatzeak fidagarritasuna eta ereduaren errendimendua

hobetzen ditu. Ikerketa honek eremu kliniko, psikologiko eta kognitibo konplexuetan aplikatu daitezkeen sistema iragarle etiko, garden eta moldagarriak garatzea ahalbidetzen du.

Acknowledgements

A Ama e Irati, por vuestro amor incondicional y por estar siempre conmigo. A Aita, que aunque ya no esté, sé que estaría orgulloso. A Gorka, por creer en mí siempre. Y a Asier por estar a mi lado en cada paso de este camino. Este trabajo es para vosotros, con todo mi cariño. Os quiero.

A mis directores, Aitor y Unai, gracias por vuestra guía, vuestro apoyo constante y por creer en mí incluso en los momentos difíciles.

Y al Istituto Nazionale dei Tumori de Milán, gracias por abrirme sus puertas y acogerme con tanta calidez durante mi estancia.

Eskerrik asko,

Amaia

Bilbao, febrero 2026

Contents

List of Figures **xvii**

List of Tables **xix**

Acronyms **xxi**

1 Introduction **1**
1.1 Context and Motivation 3
1.1.1 Importance of Interpretability in Physical and Mental Health 4
1.1.2 Extending Interpretability to Creativity and Human Capabilities 5
1.1.3 Towards a Unified Framework: Predictive Precision and Interpretability 5
1.1.4 Use Cases Underpinning the Thesis 6
1.2 Hypothesis, Objectives and Scope 7
1.3 Methodological Overview 8
1.3.1 Multimodal Data Integration and Preprocessing 8
1.3.2 Model Construction 9
1.3.3 Explainability Embedding 9
1.3.4 Comprehensive Evaluation 9
1.3.5 Iterative Refinement and Domain Validation 10
1.3.6 The Iterative and Interconnected Pipeline 10
1.4 Main Contributions 11
1.5 Thesis Outline 13

2 Related Work **15**
2.1 Machine Learning in Healthcare 16
2.1.1 Predictive Modelling in Chronic and Complex Conditions 17
2.1.2 Challenges and Research Gap in Predictive Modelling for Readmissions 19
2.2 Machine Learning in Psychiatry and Psychology 20

2.2.1 Prediction of Treatment Outcomes and Recovery Trajectories in Eating Disorders . . . 22
2.2.2 Questionnaires as Predictors in Eating Disorders . . . 24
2.2.3 Challenges and Research Gap in Questionnaire-Based Prediction of Eating Disorders . . . 25
2.3 Multimodal Learning for Human-Centered Assessment . . . 27
2.3.1 Creativity Assessment: From Subjective Ratings to ML Automation . . . 29
2.3.2 Challenges and Research Gap in Multimodal Creativity Assessment . . . 30
2.4 Explainability in Clinical Machine Learning . . . 32
2.4.1 Post-Hoc and User-Centred Explanation Methods . . . 33
2.4.2 Challenges and Research Gap in Clinical Explainability . . . 35
2.5 Summary of Research Gaps . . . 36

3 Predicting 30-Day Hospital Readmission in Patients with HF 39
3.1 ReIC Dataset . . . 40
3.1.1 Eligibility and Study Flow . . . 40
3.1.2 Baseline Characteristics . . . 41
3.1.3 Patient-Reported Outcome Measures . . . 43
3.1.4 Follow-Up at One Month . . . 43
3.1.5 Outcome Definition . . . 44
3.2 Problem Definition . . . 44
3.2.1 Formal Statement . . . 44
3.2.2 Data-Generating and Temporal Structure . . . 45
3.2.3 Clinical and Operational Constraints . . . 46
3.2.4 Specific Methodological Challenges . . . 46
3.2.5 Scope and Exclusions . . . 47
3.2.6 Success Criteria and Evaluation Plan . . . 47
3.3 Solution Design . . . 47
3.4 Explainability Framework . . . 50
3.5 Summary and Conclusions . . . 52

4 Predicting 30-Day Hospital Readmission in Patients with MCC 55
4.1 RePluris Dataset . . . 56
4.1.1 Eligibility and Study Flow . . . 56
4.1.2 Baseline Characteristics . . . 57
4.1.3 Patient-Reported Outcome Measures . . . 59
4.1.4 Follow-Up at One Month . . . 60
4.1.5 Outcome Definition . . . 60
4.2 Problem Definition . . . 60
4.2.1 Formal Statement . . . 60

4.2.2 Data-Generating and Temporal Structure 61
4.2.3 Clinical and Operational Constraints 61
4.2.4 Specific Methodological Challenges 62
4.2.5 Scope and Exclusions 63
4.2.6 Success Criteria and Evaluation Plan 63
4.3 Solution Design 63
4.4 Explainability Framework 64
4.5 Summary and Conclusions 66

5 Eating Disorder Risk and Recovery Prediction 69
5.1 ED Dataset 70
5.1.1 Study Population 70
5.1.2 Data Collection and Instruments 72
5.2 Problem Definition 73
5.2.1 Formal Statement 73
5.2.2 Data-Generating and Temporal Structure 73
5.2.3 Clinical and Operational Constraints 73
5.2.4 Specific Methodological Challenges 74
5.2.5 Scope and Exclusions 74
5.2.6 Success Criteria and Evaluation Plan 75
5.3 Solution Design 75
5.4 Explainability Framework 77
5.5 Summary and Conclusions 78

6 Multimodal Prediction of Human Creativity 81
6.1 Creativity Dataset 82
6.1.1 Participants 82
6.1.2 Data Collection Procedure 83
6.1.3 Dataset Content and Annotations 83
6.2 Problem Definition 85
6.2.1 Formal Statement 85
6.2.2 Data-Generating and Temporal Structure 86
6.2.3 Clinical and Operational Constraints 86
6.2.4 Specific Methodological Challenges 87
6.2.5 Scope and Exclusions 87
6.2.6 Success Criteria and Evaluation Plan 87
6.3 Solution Design 88
6.4 Explainability Framework 91
6.5 Summary and Conclusions 92

7 Evaluation **95**
7.1 Prediction of 30-Day Readmission in HF Patients 96
7.1.1 Handling Missing Values 98
7.1.2 Class Imbalance Handling 100
7.1.3 Bagging Ensemble Strategies 102
7.1.4 Feature Selection Guided by SHAP and Clinicians 103
7.2 Prediction of 30-Day Readmission in MCC Patients 111
7.2.1 Results in readmissions within 30 days in MCC 113
7.2.2 Calibration and Validation 116
7.3 Prediction of ED Risk and Recovery in 1 Year 119
7.3.1 One-Year Risk Prediction in Eating Disorders 121
7.3.1.1 Importance of characteristics in the RED questionnaire for risk 125
7.3.2 One-Year Recovery Prediction in Eating Disorders 126
7.3.2.1 Importance of questionnaires for recovery . . . 127
7.4 Automatic Assessment of Human Creativity Using Embeddings . 129
7.4.1 Assessing Originality 131
7.4.2 Assessing Flexibility 134
7.4.3 Assessing Elaboration 136
7.4.4 Assessing Title Creativity 138
7.4.5 Synthesis of Findings Across Creativity Dimensions . . . 140
7.5 Cross-Case Insights 141
7.6 Limitations 142
7.7 Summary 143

8 Conclusions and Future Work **145**
8.1 Summary of Work and Conclusions 146
8.2 Contributions 148
8.2.1 Application-Driven Contributions 148
8.2.2 Methodological Contributions 150
8.3 Hypothesis and Objectives Validation 151
8.4 Relevant Publications 154
8.4.1 International JCR Journals 154
8.4.2 International Conferences 155
8.5 Future Work 155
8.5.1 Data and Validation 155
8.5.2 Methodological Advances 156
8.5.3 Translation into Practice 156
8.5.4 Closing Perspective 156
8.6 Final Remarks 156

Bibliography **159**

A Feature importance ranking for ReIC **171**

B Descriptions of the variables used in the complete MCC model **175**

List of Figures

1.1 Iterative methodological pipeline for explainable multimodal ML. 10

3.1 Flowchart of patient selection and inclusion in the ReIC dataset. . 41
3.2 Bagging and ensemble strategy for mitigating the bias toward majority class prediction. 49
3.3 Cyclic explainability pipeline showing the integration of SHAP with clinician-driven feature refinement and model selection. . . . 51

4.1 Flowchart of patient selection and inclusion in the RePluris dataset. 57
4.2 Explainability framework for the RePluris case study integrating SHAP analyses for feature selection and model validation. 65

5.1 Flowchart of participant selection for eating disorders use case. . . 71
5.2 Predictive pipeline for eating disorder outcomes. 77
5.3 Calculation of feature importance using permutation. 79

6.1 Comparison of baseline (a) and completed (b) drawings — "ZONA MONTAÑOSA" (mountainous area). 84
6.2 Multimodal architecture combining text and image branches through feature fusion for creativity prediction. 89
6.3 CLIP model (Radford et al., 2021). 90
6.4 CLIP strategy for predicting creativity dimensions. 90

7.1 The 20 most important features of the best model to date. Go to Appendix A for the complete list of 50 variables that clinicians reviewed and refined. 104
7.2 ROC curves of models trained with the data subset compared with CoX and Logistic Regression. 107
7.3 Calibration of the final model for ReIC. 108
7.4 SHAP values of the final model for ReIC. 109

7.5 Relative importance of variables in the complete MCC model. Detailed descriptions of all variables included in this ranking are provided in Appendix B. 115
7.6 Precision–recall (a) and calibration (b) curves for the Decision Tree model trained on the PROFUND + Alcoholism subset. 116
7.7 Precision–recall (a) and calibration (b) curves for the XGBoost model trained on the PROFUND subset. 117
7.8 PROFUND components' contribution to predicting readmission in MCC. 119
7.9 Permutation importance for the Naïve Bayes classifier trained on the full RED questionnaire. 125
7.10 Permutation importance for the Support Vector Regression model trained on aggregated questionnaire scores. 128

8.1 The relation of the objectives defined, the contributions and their location in this dissertation. 153

List of Tables

3.1 Baseline sociodemographic and clinical characteristics. 41
3.2 Patient-reported outcomes at baseline. 44

4.1 Baseline sociodemographic, clinical, functional, and social characteristics of the RePluris dataset according to 30-day readmission status[a]. 58
4.2 Patient-reported outcomes at baseline reported as (mean ± SD). . 59

5.1 Sociodemographic and diagnostic characteristics of the eating disorder dataset at baseline and follow-up. 72

6.1 Sociodemographic characteristics of participants of creativity dataset. 82
6.2 Summary of target variables. 85

7.1 Evaluation protocol for 30-day readmission modelling in the ReIC cohort. 97
7.2 Performance without explicit handling of missing values. 99
7.3 Performance of classifiers under different imputation strategies. . . 99
7.4 Summary of model performance under different imbalance handling strategies. Values are reported as mean ± standard deviation. . 101
7.5 Performance of ensemble models and baselines. Data are presented as mean (SD) 103
7.6 Final feature set selected based on SHAP importance and clinician guidance. 105
7.7 Evaluation protocol for 30-day readmission modelling in the RePluris cohort. 112
7.8 Best-performing configurations for prediction of 30-day readmission in MCC. 114
7.9 Performance metrics for the best configurations in MCC (Part I). Values are mean (95% CI). 114
7.9 Performance metrics for the best configurations in MCC (Part II). Values are mean (95% CI). 115

7.10 Performance comparison between the PROFUND-based model (XGBoost) and the LACE index in the test and external cohorts. Percentage change (%Δ) is relative to LACE. 118
7.11 Evaluation protocol for one-year risk and recovery modelling in eating disorders. 120
7.12 Predictive performance of ML models for 1-year ED risk across different input sets. Outcome defined as EAT-26 > 20 at follow-up. 122
7.13 R^2 scores predicting recovery level in eating disorder 127
7.14 Evaluation protocol for multimodal creativity modelling (originality, flexibility, elaboration, title creativity). 130
7.15 Performance in predicting originality (O) using embedding–CNN combinations and traditional classifiers. 131
7.16 Performance in predicting flexibility (FLE) using (a) embedding–CNN combinations and (b) CLIP embeddings with traditional classifiers. 134
7.17 Performance in predicting elaboration (E) using (a) embedding–CNN regressors and (b) CLIP embeddings with traditional regressors. . 136
7.18 Performance in predicting title creativity (T) using (a) embedding–CNN configurations and (b) CLIP embeddings with traditional classifiers. 138

Acronyms

ADASYN Adaptive Synthetic Sampling

ADHF Acute Decompensated Heart Failure

AI Artificial Intelligence

AN Anorexia Nervosa

ARMS Adherence to Refill and Medications Scale

AUC Area Under the Curve

AUPRC Area Under the Precision-Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

BED Binge Eating Disorder

BERT Bidirectional Encoder Representations from Transformers

BMI Body Mass Index

BN Bulimia Nervosa

CHF Congestive Heart Failure

CI Confidence Interval

CLIP Contrastive Language–Image Pre-training

CNN Convolutional Neural Network

COPD Chronic Obstructive Pulmonary Disease

COSMIN COnsensus-based Standards for the selection of health Measurement INstruments

CP Current Patients

DT Decision Tree

EAT Eating Attitudes Test

ED Eating Disorder

EDE Eating Disorder Examination

EDNOS Eating Disorder Not Otherwise Specified

EEG Electroencephalography

EHR Electronic Health Record

FLE Flexibility

FP Former Patient

GB Gradient Boosting

GP General Population

HAD Hospital Anxiety and Depression

HF Heart Failure

ICU Intensive Care Unit

IQR Interquartile range

KNN K Nearest Neighbors

LACE Length of stay, Acuity of admission, Comorbidities and Emergency department visits

LASSO Least Absolute Shrinkage Selection Operator

LGBM Light Gradient-Boosting Machine

LGMM Latent Growth Mixture Modelling

LIME Local Interpretable Model-Agnostic Explanations

LR Logistic Regression

MAE Mean Absolute Error

MCC Multiple Chronic Conditions

ML Machine Learning

MLHFQ Minnesota Living with Heart Failure Questionnaire

MLP Multi-layer Perceptron

MMLA Multimodal Learning Analytics

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

NB Näive Bayes

NN Neural Netwok

PCA Principal Component Analysis

PR Precision-Recall

PROM Patient-Reported Outcome Measure

QEWP Questionnaire on Eating and Weight Patterns

RANSAC Random Sample Consensus

RED Resilence in Eating Disorders

RF Random Forest

RMSE Root Mean Squared Error

ROC Receiver Operating Characteristic

RS Resilence Scale

SBP Systolic Blood Pressure

SD Standard Deviation

SEIQoL Schedule for the Evaluation of Individual Quality of Life

SHAP SHapley Additive ex-Planations

SMOTE Synthetic Minority Over-sampling Technique

SMOTEENN Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors

SPSS Statistical Package for the Social Sciences

SVC Support Vector Classification

SVM Support Vector Machine

SVR Support Vector Regression

TCT-DP Test of Creative Thinking – Drawing Production

TFI Tilburg Frailty Indicator

UD University of Deusto

VAS visual analogue scale

WHOQOL World Health Organization Quality of Life

XAI Explainable AI

XR Extended Reality

Nothing in life is to be feared; it is only to be understood.

MARIE CURIE (1867 A.D. - 1934 A.D.)

CHAPTER

1

Introduction

ARTIFICIAL Intelligence (AI) and Machine Learning (ML) have become central to the transformation of healthcare. Advancements in computer vision, natural language processing and deep learning have given rise to intelligent systems capable of interpreting medical images, analysing electronic health records (EHRs), modelling physiological signals and assisting with decision making (Fahim et al., 2025). These algorithms can recognise complex patterns in high-dimensional data and generate predictions or recommendations that support clinicians. In physical healthcare, AI-driven analysis of imaging modalities (e.g., radiology, dermatology and cardiology) aids early disease detection and improves diagnostic accuracy (Fahim et al., 2025). ML techniques also underpin precision medicine by integrating diverse sources, including genomic profiles, immunological data and patient histories, to identify high-risk individuals, predict disease progression and optimise therapeutic strategies (Chen et al., 2025). Through wearable sensors and remote monitoring, AI enables real-time management of chronic diseases, allowing personalised interventions that can improve quality of life and reduce hospitalisation (Pan et al., 2025). Moreover, telemedicine platforms equipped with natural language processing facilitate remote consultations and triage, extending access to care and reducing burden on overstretched clinical services (Li et al., 2024).

While physical health care has seen rapid adoption of AI tools, mental health applications are emerging as a similarly transformative domain. Psychiatric disorders represent a substantial global burden with significant morbidity and mortality (Dykxhoorn y Kirkbride, 2018). ML models can exploit multimodal data, from brain imaging and speech patterns to sensor-derived behavioural signals and social

media, to detect early signs of depression, anxiety and other mental illnesses (Narimani y Naeim, 2025). Predictive analytics can forecast relapse risk and treatment response by analysing longitudinal records and physiological indicators (Narimani y Naeim, 2025). Such approaches enable timely interventions and personalised treatment plans, and they offer the potential to triage resources in under-resourced mental health systems (Lee et al., 2021). AI-driven conversational agents and digital therapeutics may also provide continuous psychological support, complementing face-to-face therapy. However, mental health data are particularly sensitive; ensuring privacy, minimising algorithmic bias and maintaining the therapeutic alliance demand careful oversight (Narimani y Naeim, 2025).

A unifying theme across these applications is the promise of personalised, proactive and cost-effective healthcare. By synthesising diverse information sources, AI systems can deliver individualised recommendations that surpass one-size-fits-all approaches (Chen et al., 2025). In chronic disease management, integrated AI platforms support patients in daily self-management tasks, improve glycaemic control and pain management, and promote psychosocial well-being (Hwang et al., 2025). The technology's potential extends to healthcare equity; mobile diagnostics and inexpensive wearables can reduce geographical and socioeconomic disparities (Fahim et al., 2025). Nevertheless, deploying AI at scale reveals substantial challenges. Many models rely on large datasets that may be biased or unrepresentative, risking exacerbation of existing disparities (Narimani y Naeim, 2025). Data quality, interoperability and portability issues can limit generalisability across healthcare settings (Dong et al., 2025). There is a shortage of research on how to integrate AI into the behavioural and emotional aspects of chronic condition self-management (Hwang et al., 2025), and existing work often neglects the needs and workflows of end users (Liu et al., 2025).

Moreover, trust is foundational in clinical practice; clinicians and patients must understand and believe in the recommendations generated by AI systems. However, many state-of-the-art models are "black boxes" whose internal logic remains opaque even to developers. Ethical and legal frameworks increasingly emphasise the need for transparency and interpretability (Hildt, 2025). Explainable AI (XAI) seeks to provide insight into how a model arrives at its outputs and why certain features influence its predictions. Authors have argued that explainability can help mitigate potential negative impacts on doctor–patient relationships, autonomy and informed consent (Hildt, 2025). Explanations that are clear and concise may increase clinicians' willingness to adopt AI and facilitate shared decision making (Rosenbacke et al., 2024). When AI tools integrate seamlessly into clinical workflows and offer patient-specific reasoning, they can support more efficient care and allow providers to concentrate on human-centred tasks (Liu et al., 2025). Conversely, black-box systems may erode trust; in real-world deployments of AI-assisted emergency call

triage, the absence of explanations led dispatchers to question the system's alerts and reduced compliance (Hildt, 2025). Similar concerns arise in mental health, where algorithmic opacity can undermine therapeutic alliances and hinder adoption (Narimani y Naeim, 2025).

Explainability is not a universal remedy, and there is debate about the appropriate form and necessity of explanations. Some authors note that while explainability may increase trust and aid shared decision making, it is unclear whether it always improves clinical outcomes (Hildt, 2025). Overly simplistic or misleading post-hoc explanations can provide false reassurance, whereas complex explanations may overwhelm users (Rosenbacke et al., 2024). Balanced trust is required: clinicians should neither follow AI recommendations uncritically nor disregard them due to scepticism. Nevertheless, the ethical argument for explainability remains compelling. Respect for patient autonomy demands that individuals understand the basis for decisions affecting their health and can make informed choices (Hildt, 2025). Regulatory guidelines emphasise that explainability, along with fairness, accountability and respect for privacy, is a prerequisite for trustworthy AI (Hildt, 2025). As AI continues to permeate healthcare, developing methods that provide faithful, understandable and context-appropriate explanations is an active research area (Vani et al., 2025).

This introduction has sketched the transformative potential of AI and ML across physical and mental health domains and highlighted the ethical imperative of explainability for fostering trust and responsible adoption. The remainder of Chapter 1 deepens this foundation. Section 1.1 presents background context on AI technologies, healthcare systems and the broader societal drivers motivating this research. Section 1.2 formulates the research hypothesis and objectives, articulating the overarching questions addressed in this thesis. Section 1.3 outlines the methodological framework used to investigate these questions, including data sources, modelling approaches and evaluation strategies. Section 1.4 delineates the principal contributions of the thesis, situating them within the existing literature. Finally, Section 1.5 describes the structure of the dissertation. Together, these sections provide a comprehensive roadmap for the thesis while maintaining focus on the overarching goals of improving healthcare through AI and ensuring that the resulting systems are transparent, trustworthy and ethically grounded.

1.1 Context and Motivation

Machine Learning (ML) techniques have become a central engine for personalised and precision medicine (MacEachern y Forkert, 2021). In clinical settings, algorithms can analyse large volumes of heterogeneous data (genetic, imaging and contextual) to support diagnosis and prognosis. Yet many of the most accurate

ML models are opaque: deep neural networks and other non-linear models often act as "black boxes" whose internal workings remain hidden to clinicians (Vani et al., 2025). Evidence from personalised health monitoring research shows that black-box models deliver high predictive accuracy but rarely provide patient-specific explanations; explanations are either generic or post-hoc and often do not generalise across populations (Vani et al., 2025). For clinical adoption, predictions must be accompanied by transparent reasoning so that practitioners can trust and validate automated recommendations. This tension between predictive performance and interpretability is visualised in the complexity–interpretability trade-off: while logistic regression models offer high interpretability, heavily-parametrised neural networks can be difficult to explain and sometimes do not improve accuracy (Sidey-Gibbons y Sidey-Gibbons, 2019).

The methodological stance adopted in this dissertation explicitly embraces this complexity–interpretability tension rather than attempting to resolve it through a single modelling paradigm. The heterogeneity of algorithms employed across the different case studies is not arbitrary but reflects a context-dependent design strategy. For each use case, model families were selected according to the structure and dimensionality of the data, the clinical or psychological nature of the prediction task, and the required balance between predictive performance and interpretability. Consequently, both classical statistical learning methods and deep learning architectures were considered within a unified evaluation framework, allowing systematic comparison rather than methodological fragmentation.

1.1.1 Importance of Interpretability in Physical and Mental Health

In physical health, ML is used to identify individuals at risk of conditions such as heart failure or complications from chronic diseases. Clinicians need to understand the factors driving a model's prediction to justify interventions and maintain accountability. Practical guides to ML in medicine observe that regularised logistic models (e.g., LASSO-regularised linear regression) can perform competitively while preserving the ability to identify features guiding the prediction, whereas the archetypal "black-box" neural network may not yield better accuracy (Sidey-Gibbons y Sidey-Gibbons, 2019). Even when more complex models are considered, there is an urgent need for explainable AI (XAI) techniques that deliver patient-specific explanations; research emphasises closing the gap between algorithmic accuracy and interpretability by integrating methods like attention mechanisms and SHAP values to derive both global and local insights (Vani et al., 2025).

In mental health, prediction tasks are complicated by symptom heterogeneity, contextual factors and sparse data. Models optimised solely for accuracy risk, obscur-

ing biases and might perpetuate inequities. Scholars argue that, compared with pragmatic models aimed at high predictive performance, explanatory models are necessary to ensure patient safety and gain clinical trust (Lee et al., 2021). There is a bias-variance trade-off: increasing model complexity may reduce bias but can obscure interpretability; hence, future AI systems must strive for transparency to capture patient complexity responsibly (Joyce et al., 2023; Abdullah et al., 2021). This is particularly relevant for mental-health conditions such as eating disorders, where early detection and personalised recovery plans require models that clinicians and patients can scrutinise and discuss.

1.1.2 Extending Interpretability to Creativity and Human Capabilities

Beyond illness, human well-being includes capacities such as creativity, which is linked to cognitive flexibility, problem solving and resilience. Traditional assessments of creativity, often used in psychology and education, rely on subjective judgments of fluency, originality or elaboration and are time-consuming and inconsistent (Zhang et al., 2025). Recent work demonstrates that multimodal ML models combining visual features with semantic information can automate scoring of divergent-thinking tasks, achieving high correlation with human raters while reducing subjectivity and labour (Zhang et al., 2025). These models can simultaneously evaluate novelty, fluency and flexibility, showing that integrating different data modalities enriches the representation of creative behaviour (Zhang et al., 2025). Similarly, interpretable Machine Learning methods applied to art evaluation reveal that attributes such as emotional expressiveness, valence and symbolism are non-linearly related to creativity judgments; sudden changes in these relationships around rating midpoints suggest aesthetic thresholds (Spee et al., 2024). Such findings highlight the importance of modelling complex, non-linear interactions while still providing interpretable insights.

Creativity assessment is relevant to mental health because creative thinking exercises can serve as markers of cognitive health and resilience, especially in populations recovering from illness or trauma. Including creativity in an ML framework, therefore, broadens the thesis beyond pathology to embrace positive psychological constructs, underscoring that trustworthy AI should support both risk prediction and the promotion of human capabilities.

1.1.3 Towards a Unified Framework: Predictive Precision and Interpretability

The need to balance predictive precision with interpretability is an inherent part of any human assessment. While flexible models such as neural networks can

capture complex non-linear relationships, their opaque nature raises ethical and legal concerns; simpler models like logistic regression offer clearer explanations but may miss important interactions (Sidey-Gibbons y Sidey-Gibbons, 2019). Research in mental health warns that deploying black-box models without adequate transparency can compromise patient safety and erode trust (Lee et al., 2021). Therefore, the thesis advocates for interpretable ML approaches that harness multimodal data (e.g., clinical records, images and text) and generate context-sensitive explanations. Techniques such as attention mechanisms, feature importance scores, and other XAI tools can illuminate why a model reached its conclusion (Vani et al., 2025), empowering clinicians and users to evaluate predictions, detect biases and make informed decisions.

1.1.4 Use Cases Underpinning the Thesis

This dissertation operationalises the above insights through four representative use cases spanning physical health, mental health and human capabilities.

1. Predicting hospital readmissions for heart-failure patients and individuals with multiple chronic conditions. Readmissions are costly and often preventable; ML models can synthesise electronic health records, vital signs and contextual factors to identify high-risk patients. Incorporating interpretable features enables clinicians to design targeted interventions and improve care coordination.

2. Modelling eating-disorder risk and recovery trajectories. Eating disorders involve multifaceted behavioural, psychological and social factors. Interpretable ML can help clinicians recognise early warning signs, personalise treatment plans and monitor recovery, while ensuring that predictions are comprehensible to patients and caregivers (Lee et al., 2021).

3. Evaluating creative capacities through multimodal integration. Automated, interpretable assessment of divergent-thinking tasks shows how ML can quantify abstract human traits and reveal the underlying factors driving creativity (Zhang et al., 2025). This case illustrates the broader applicability of interpretable models to cognitive assessment and underscores the value of modelling positive psychological constructs.

Together, these cases demonstrate that explainable and multimodal machine learning can address complex, human-centred prediction tasks across domains. They form the basis for the central hypothesis of this dissertation, which is presented in the next section.

1.2 Hypothesis, Objectives and Scope

Based on the current state of machine learning in healthcare, mental health, and creativity research, the hypothesis of this dissertation is:

> **Hypothesis.** *Machine learning models that incorporate explicit explainability strategies improve decision-making in complex human-centred contexts by providing both accurate predictions and understandable basis.*

To be able to validate this hypothesis, the general goal of this dissertation is:

> **Goal.** *To design and implement explainable and multimodal machine learning systems that predict relevant events in health and behavioural domains, evaluate their predictive performance, and demonstrate their interpretability and applicability in real-world contexts.*

This general goal can be achieved by addressing the following more specific and measurable objectives:

O1. To develop machine learning models for representative use cases: predicting hospital readmissions in patients with heart failure, predicting hospital readmissions in patients with multiple chronic conditions, assessing risk and recovery in eating disorders, and evaluating creativity through multimodal (image and text) analysis.

O2. To analyse the performance of the developed models by comparing them with traditional approaches (e.g., logistic regression, Cox regression) and benchmarking their predictive power, calibration, and robustness to issues such as class imbalance and missing data.

O3. To integrate explainability techniques, such as SHAP values and feature attribution methods, into the models to provide human-understandable explanations of predictions, thereby supporting domain experts in interpreting outcomes and validating their clinical or psychological plausibility.

O4. To assess the applicability of the models within their respective fields together with clinicians, psychologists, and researchers.

The resulting systems should also fulfil the following requirements:

R1. Domain independence: the modelling framework should be applicable to heterogeneous contexts (clinical, psychological, and cognitive) without requiring fundamental changes in methodology.

R2. User relevance: predictions and explanations should reflect individual variability, thereby allowing for personalised assessment and intervention.

R3. Transparency and accountability: explanations must be sufficiently clear to support decision-making, reduce the risk of algorithmic opacity, and align with ethical standards of responsible AI.

The work presented in this dissertation does not deal with the following conditions:

C1. Unknown event detection: this work does not propose methods to discover entirely novel outcomes or behaviours; instead, it focuses on improving prediction and explanation of predefined targets.

C2. Multi-user joint modelling: the research assumes individual-level modelling and does not explicitly consider simultaneous predictions across multiple interacting users.

C3. Causal inference: while associations and predictive factors are identified, establishing causal relationships remains outside the scope of this dissertation.

1.3 Methodological Overview

The research presented in this dissertation follows a methodological pipeline inspired by iterative principles of MLOps, the framework for operationalising machine learning in production environments (Kreuzberger et al. (2022)). However, whereas industrial MLOps emphasises deployment, scalability, and monitoring, the present framework adapts these principles to the scientific and human-centred context. Here, the focus lies on ensuring interpretability, domain-specific applicability, and continuous refinement through expert feedback.

The methodology is structured around five main stages, connected by a cyclical backbone but also enriched with cross-links and feedback loops. This design reflects the reality of research, where insights gained in one stage often prompt revisiting earlier steps.

1.3.1 Multimodal Data Integration and Preprocessing

The process begins with the collection and preprocessing of multimodal data, including structured clinical records, psychometric questionnaires, and creative artefacts (images and text). Preprocessing entails: handling missing values, addressing class imbalance, feature scaling, and harmonising heterogeneous formats into a common representation.

Unlike classical pipelines, preprocessing is not a single-pass step but may be revisited based on later evaluation results.

1.3.2 Model Construction

This stage involves developing machine learning models with two complementary strategies:

- Inherently interpretable models to provide transparency by design,
- Complex predictive frameworks to maximise accuracy on heterogeneous data.

Model construction is highly iterative: design choices are frequently refined in response to preprocessing outcomes or interpretability analyses.

1.3.3 Explainability Embedding

Explainability constitutes a methodological pillar rather than a post-hoc add-on. The goal is to provide explanations that are not only mathematically consistent but also domain-understandable by clinicians, psychologists, and creativity researchers.

This stage also serves as a feedback signal: insights from explanation analysis may highlight spurious correlations or irrelevant features, triggering model redesign or further data preprocessing.

1.3.4 Comprehensive Evaluation

Evaluation extends beyond traditional performance metrics. While predictive quality is measured, equal importance is placed on:

- Calibration of probabilistic outputs,
- Robustness to imbalance and noise,
- Interpretability quality, assessed with domain expert input.

In analogy with MLOps monitoring, this stage functions as an early warning system, identifying deficiencies that necessitate iteration to earlier steps.

1.3.5 Iterative Refinement and Domain Validation

The final stage emphasises cyclical improvement, grounded in expert feedback and domain validation. Instead of deployment monitoring (as in MLOps), the loop integrates clinical, psychological, and cognitive experts into the evaluation process, ensuring that outputs are contextually relevant and ethically sound. This step may trigger new iterations starting from data preprocessing, model design, or explanation embedding, depending on the insights obtained.

1.3.6 The Iterative and Interconnected Pipeline

The methodological framework is thus not a strict cycle but a network of interactions, where evaluation, explainability, and domain validation feed back into earlier stages. This mirrors the philosophy of MLOps while adapting it to the research context, thereby fostering reproducibility, interpretability, and trustworthiness.

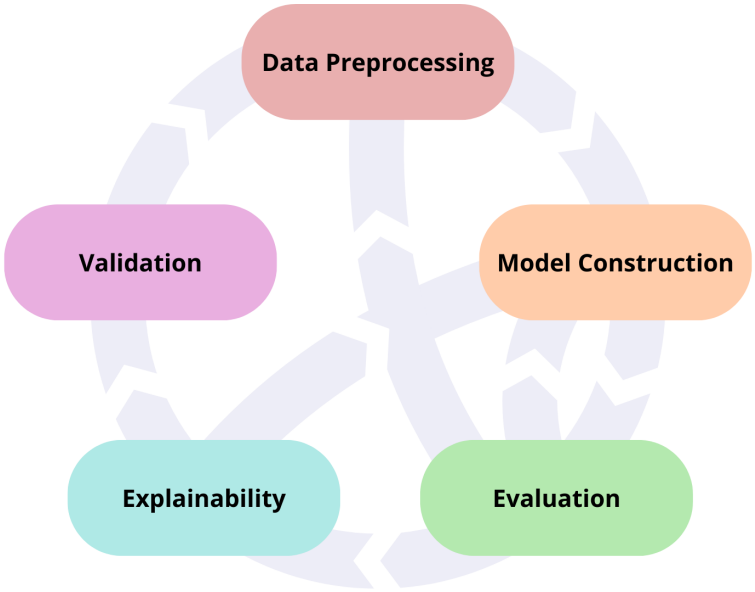


Figure 1.1: Iterative methodological pipeline for explainable multimodal ML.

The graph illustrates the cyclical backbone of the methodology and its feedback loops, showing the dynamic interactions between preprocessing, model construction, explainability, evaluation, and refinement. The figure highlights how evaluation and explainability are not endpoints but central elements that drive iteration (see Figure 1.1).

1.4 Main Contributions

The contributions of this dissertation are twofold: (i) application-driven contributions, arising from the three case studies in healthcare, mental health, and creativity; and (ii) methodological and technical contributions, which address cross-cutting challenges such as multimodal integration, class imbalance, and explainability. Together, they demonstrate how explainable and multimodal ML can be designed and validated in human-centred contexts.

The first contribution emerges from the domain of cardiovascular health, where predicting hospital readmissions for patients with heart failure remains a persistent clinical challenge. By integrating patient-reported outcomes with conventional clinical variables, the research demonstrates that ensemble models not only outperform traditional statistical baselines but also provide interpretable risk stratification. This work shows that psychosocial dimensions such as frailty, anxiety, and depression are central to reliable prediction, thereby highlighting the value of blending biomedical and self-reported data in decision support systems for early intervention.

The second contribution focuses on the treatment of patients with multiple chronic conditions (MCC) through the RePluris study. This population is characterised by multimorbidity, frailty and vulnerability to rehospitalisation. The thesis demonstrates that predictive models combining clinical, functional, and social data, enriched with outcomes reported by patients and caregivers, can provide meaningful information about the risk of rehospitalisation. In particular, the integration of social determinants of health and functional indices into explainable machine learning frameworks improves the ability to generate clinically applicable and ethically grounded predictions for a highly complex patient group, surpassing the predictive power of validated instruments such as LACE (Length of stay, Acuity of admission, Comorbidity, Emergency department visits) index.

The third contribution is situated within the field of mental health, specifically the prediction of risk and recovery trajectories in eating disorders. Through the exclusive use of psychometric questionnaires, the study demonstrates that resilience, self-perception, and quality of life are strong determinants of both onset risk and recovery. Beyond confirming the predictive value of these factors, the findings reveal that questionnaires, traditionally regarded as diagnostic or monitoring tools, can also serve as powerful predictors. This contribution advances the understanding of recovery processes and supports the design of personalised monitoring strategies grounded in interpretable assessments.

The fourth contribution focuses on the computational evaluation of human creativity. Combining visual artefacts with textual descriptions is shown to enhance

assessment. Textual information, in particular, adds substantial value for evaluating originality, flexibility, and elaboration. This finding challenges the dominance of image-only approaches and demonstrates that multimodal integration is more effective in capturing the abstract and multidimensional nature of creative performance. It thus contributes to the broader effort of making the assessment of creativity more objective, nuanced, and scalable.

Beyond the specific case studies, this dissertation delivers several methodological advances:

- Multimodal integration framework capable of combining structured clinical data with psychometric responses or perceptual inputs, such as images and text.
- A methodology that systematically manage of real-world data issues, including imbalance, missingness, and noise, applied consistently across heterogeneous domains.
- Embedded explainability techniques, employing feature attribution methods to generate domain-meaningful interpretations that bridge algorithmic outputs with expert reasoning.
- Iterative research pipeline adapted to scientific contexts to ensure methodological robustness, reproducibility, and expert-driven validation.

An additional cross-cutting contribution lies in the principled comparison of heterogeneous modelling strategies across domains. Rather than committing to a single algorithmic family, the dissertation systematically contrasts traditional statistical models, ensemble methods, and deep neural architectures within each use case. This comparative approach enables the evaluation of trade-offs between model complexity, generalisability, calibration, and explainability. By embedding these comparisons into the experimental design, the work advances a methodological framework in which algorithm selection becomes a reasoned, data- and context-driven decision rather than a purely performance-oriented choice.

Beyond the individual case studies, the overarching contribution of this dissertation lies in the articulation of a coherent human-centred machine learning framework. Rather than presenting isolated applications, the thesis proposes a unifying methodological strategy that integrates multimodal data processing, context-aware algorithm selection, embedded explainability, and iterative expert validation. This framework demonstrates that predictive accuracy and interpretability need not be opposing objectives, but can be jointly optimised through principled design choices. The general contribution therefore resides not only in domain-specific findings, but in establishing transferable methodological principles for explainable machine

learning in heterogeneous human-centred domains.

This dissertation advances the state of the art by showing that it is possible to design models that are both accurate and interpretable across diverse human-centred domains. It provides domain-specific insights into heart failure, patients with multiple chronic conditions, eating disorders, and creativity assessment, while contributing generalisable methodological principles that bridge the gap between cutting-edge ML techniques and their responsible application in practice.

The remainder of this dissertation elaborates on these contributions through four case studies, following the structure outlined below.

1.5 Thesis Outline

This dissertation is structured in eight chapters.
Chapter 1 introduces the research. It presents the general context and motivation, formulates the central hypothesis, specifies the objectives and scope, and outlines the methodological strategy. The chapter also summarises the scientific and technical contributions and provides the structure of the document.

Chapter 2 reviews the related work. It examines the role of machine learning in healthcare and psychology, explores advances in explainable AI (XAI) within clinical contexts, and analyses multimodal approaches for human-centred assessment. The chapter concludes with a synthesis of open research gaps that motivate the proposed work.

Chapter 3 presents the first case study: predicting hospital readmissions in patients with heart failure. It describes the dataset and problem formulation, the modelling strategies employed, their evaluation, and the interpretability analyses conducted. The chapter concludes with a discussion of the clinical implications of the findings.

Chapter 4 presents the second case study: predicting hospital readmissions in patients with multiple chronic conditions. It introduces the RePluris dataset, formalises the problem, and evaluates different modelling approaches including survival and classification frameworks. Special emphasis is placed on the integration of social determinants of health and patient-reported outcomes. The chapter concludes with insights on how explainable models can inform the management of multimorbidity.

Chapter 5 addresses the third case study: predicting eating disorder risk and recovery. It introduces the dataset of questionnaires, describes the classification and regression models used, and analyses predictive performance. It also evaluates the

contribution of specific questionnaires, highlighting resilience and quality-of-life measures as key recovery factors.

Chapter 6 focuses on the fourth case study: assessing human creativity through multimodal modelling. It describes the integration of visual and textual data, the design of unimodal and multimodal architectures, and the results obtained for originality, elaboration, flexibility, and title creativity. The chapter further discusses the added value of textual descriptions for capturing abstract creative features.

Chapter 7 evaluates the outcomes of all four case studies. It provides a comparative analysis of model performance, synthesises cross-case insights, and discusses shared methodological patterns. The chapter also reflects on ethical and practical implications, identifies limitations, and outlines lessons learned from the evaluation.

Chapter 8 concludes the dissertation. It summarises the findings, revisits the central hypothesis and objectives, and highlights the broader contributions. Finally, it outlines future lines of research, including the extension of multimodal explainability methods, the integration of causal inference, and the exploration of deployment scenarios aligned with responsible AI practices.

The most dangerous phrase in the language is, 'We've always done it this way.'

GRACE HOPPER (1906 A.D. – 1992 A.D.)

CHAPTER

2

Related Work

ML in human-centred contexts has grown into a diverse research area encompassing medicine, psychology, and computational models of cognition. The objective of this chapter is to provide a high-level overview of the current state of research relevant to this dissertation, followed by detailed examinations of approaches in healthcare, psychology, explainable clinical applications, and multimodal human-centred assessment. This review establishes the scientific background that motivates the methodological choices of the thesis and identifies gaps that guide its contributions.

First, Section 2.1 discusses ML in Healthcare, highlighting predictive modelling for clinical decision support, with emphasis on chronic conditions such as heart failure where traditional models have limited accuracy. Section 2.2 focuses on ML in Psychology, covering predictive approaches to mental health disorders, with a particular emphasis on eating disorder risk and recovery, where questionnaires have been central. Section 2.4 surveys the field of Explainable AI (XAI) in Clinical Applications, presenting methods that render models interpretable for clinicians and patients, including feature importance analyses and model-agnostic tools. Section 2.3 turns to Multimodal Learning for Human-Centred Assessment, reviewing the integration of text, images, and structured data to predict abstract constructs such as creativity, resilience, and well-being. Section 2.5 synthesises the findings in a Summary of Research Gaps, highlighting the limitations of current approaches and positioning this dissertation as a response to those gaps.

2.1 Machine Learning in Healthcare

The application of machine learning (ML) within healthcare represents one of the most transformative developments in modern medical science. In the realm of diagnostic imaging, algorithms leveraging convolutional neural networks (CNNs) have demonstrated exceptional accuracy in interpreting radiological data. For instance, in the domain of diagnostic radiology, convolutional neural networks and other deep learning architectures have enabled automatic interpretation of radiological images with enhanced speed and sensitivity, fostering integration into clinical practice and driving adoption in FDA-approved diagnostic devices. Review studies underscore the profound influence of ML on radiology, highlighting its capacity to augment clinician performance through automated pattern recognition and feature extraction (Najjar, 2023).

Parallel to imaging, ML has significantly impacted drug discovery and development. Comprehensive reviews detail applications ranging from target identification to toxicity prediction, emphasising that ML techniques have been deployed at every stage of the drug development pipeline to accelerate timelines and reduce costs (Dara et al., 2021; Dhudum et al., 2024; Kant et al., 2025; Ocana et al., 2025). Structural deep learning approaches, including generative models and graph-based architectures, are particularly promising for de novo molecular design and activity prediction (ozcelik et al., 2022). Recent systematic evaluations affirm that deep learning methods offer marked improvements in both throughput and predictive accuracy for molecular properties such as binding affinity and pharmacokinetic parameters (Schapin et al., 2023).

Beyond experimental science, the deployment of ML in managing real-world clinical data and electronic health records has received substantial attention. A 2025 systematic review reveals that disease prediction and management models frequently utilise supervised and unsupervised approaches trained on observational data, highlighting their utility in informing clinical decision-making (Alhumaidi et al., 2025). Further reviews contextualise the broader role of AI, including ML, in improving diagnostic workflows, treatment recommendations, and patient engagement, while emphasising ethical and legal challenges intrinsic to clinical adoption (Alowais et al., 2023).

Methodologically, the landscape of ML in healthcare spans interpretable statistical models to high-capacity learning architectures. Reviews of deep learning applications emphasise that CNNs, RNNs, autoencoders, and Restricted Boltzmann Machines are now routinely employed across imaging, genomics, electronic health records, and wearable sensor data domains, but also note the need for improved interpretability to ensure acceptance by domain experts (Miotto et al., 2017). Mul-

timodal ML techniques, which integrate heterogeneous data sources, are likewise gaining traction for enhancing predictive accuracy in complex clinical scenarios (Krones et al., 2025). Drug discovery methodologies increasingly incorporate graph neural networks and generative frameworks to model molecular structures, whereas structure-based methods contribute to affinity prediction and rational design (ozcelik et al., 2022).

Briefly, machine learning underpins a broad spectrum of innovations in healthcare, including diagnostic imaging, risk stratification, drug discovery, administrative optimisation, training of privacy-preserving models, and multimodal data fusion. From a methodological standpoint, its arsenal ranges from traditional interpretable models to advanced deep learning architectures, each carefully selected to address specific forms of clinical data and complexity. This scientific and methodological foundation lays the groundwork for one of the most impactful applications of machine learning in medicine: outcome prediction. The following section will explore in detail the theoretical foundations, algorithmic strategies, and empirical evaluations of models designed to predict clinical outcomes such as mortality, readmission, disease progression, and quality of life indicators.

2.1.1 Predictive Modelling in Chronic and Complex Conditions

Predictive modelling has emerged as a cornerstone of machine learning (ML) applications in healthcare, offering the ability to anticipate disease onset, progression, and patient outcomes. In oncology, for example, ML algorithms have been successfully deployed to predict the development and trajectory of skin, breast, and lung cancers, while in neurology they have been applied to the modelling of Parkinson's disease progression (Banapuram et al., 2024; Carvalho y Cruz, 2020). Similar approaches have also been adopted in cardiovascular medicine, where risk prediction is an essential component of clinical management. These methods typically integrate individual patient data with historical clinical records, thereby improving both the timeliness and accuracy of diagnosis compared with conventional methods.

A notable area of advancement lies in multimodal data fusion, which seeks to integrate heterogeneous datasets such as imaging, genomic profiles, electronic health records (EHRs), and clinical notes. Studies in neurology and oncology consistently demonstrate that multimodal integration enhances predictive accuracy by over six percent compared with unimodal models (Banapuram et al., 2024). Early fusion methods, where features from different modalities are combined at the input stage, are the most frequently applied, especially with imaging and EHR data. However, interest is growing in other combinations such as genomics with imaging, or time-series physiological signals with EHRs, which hold promise for improving clinical

insight in complex chronic conditions (Kline et al., 2022).

The digitisation of health records has been a major catalyst for predictive analytics, enabling the deployment of ML models on large, routinely collected datasets. Nevertheless, EHR data are inherently complex due to heterogeneity in format, variable diagnostic standards, and frequent missing values. To address these challenges, research has focused on computable phenotyping and modelling diagnostic decision-making processes. Notably, initiatives such as those at the Massachusetts Institute of Technology are working toward the development of intelligent health records in which ML capabilities are embedded at the system level, supporting both diagnostic assistance and treatment recommendations (Kline et al., 2022). While predictive modelling is now widely applied across chronic diseases, its role in heart failure (HF) is particularly critical, given the high burden of morbidity, mortality, and hospital readmissions associated with this condition.

Within HF, predictive modelling has evolved significantly with the introduction of advanced ML approaches. Logistic regression, random forests, support vector machines (SVMs), neural networks, and gradient boosting methods such as XGBoost are among the most frequently employed algorithms (Hidayaturrohman y Hanada, 2024). These models have demonstrated value in identifying patients with HF and in estimating individualised risks of mortality and readmission (Guo et al., 2020). Reported performance varies across studies, with mortality prediction models achieving area under the receiver operating characteristic curve (AUC) values ranging from 0.48 to 0.92, while readmission prediction models report AUCs between 0.47 and 0.84 (Mpanya et al., 2021). A recent meta-analysis found that neural networks attained the highest overall AUC for mortality prediction (0.808), whereas SVMs performed best in predicting readmission (AUC 0.733) (Hajishah et al., 2025).

The predictive capacity of these models depends heavily on the data sources and features available. Most studies rely on structured EHR data, including demographic variables such as age and sex, lifestyle characteristics, and comorbidities. Routine clinical measures such as weight, blood pressure, and heart rate are commonly incorporated. Some studies extend this by including unstructured data such as physician notes, though their use remains limited. Preprocessing methods, including imputation for missing values and feature selection, are widely applied. Statistical imputation methods such as mean substitution remain common, though ML-based imputation techniques have demonstrated superior performance. Feature selection strategies are also widely used, although most studies do not perform systematic comparisons between alternative methods (Hidayaturrohman y Hanada, 2024). While these approaches demonstrate promise, their clinical applicability in

HF remains constrained by persistent methodological and translational challenges, as discussed in the following subsection.

2.1.2 Challenges and Research Gap in Predictive Modelling for Readmissions

Despite considerable progress in applying machine learning to outcome prediction in chronic conditions, important limitations continue to restrict the translation of predictive models for hospital readmissions into routine clinical practice. These challenges apply both to focused conditions such as heart failure (HF) and to broader multimorbidity contexts such as multiple chronic conditions (MCC).

First, external validation is largely absent. Many studies, particularly in HF, are developed on single-site, retrospective cohorts, which prevents generalisability and hinders transferability across populations (Błaziak et al., 2022). A systematic review reported that over one hundred HF studies were excluded due to insufficient validation (Mpanya et al., 2021). Similarly, research on MCC cohorts remains scarce and typically limited to localised healthcare systems, further amplifying demographic and geographic biases. The absence of representative studies in regions such as Africa or the Middle East underscores persistent equity and fairness concerns.

Second, data quality and heterogeneity represent critical barriers. Electronic health records (EHRs) are often incomplete, inconsistently labelled, and heterogeneous in structure, creating challenges for preprocessing and harmonisation (Ghassemi et al., 2018). These limitations are especially acute in MCC, where fragmented care and the diversity of comorbidities exacerbate missingness and noise. Although imputation and feature selection strategies are widely applied, most studies do not systematically compare alternatives, leaving methodological uncertainty and undermining reproducibility (Hidayaturrohman y Hanada, 2024).

Third, interpretability remains limited. While decision trees and other transparent approaches offer intuitive insights, higher-performing methods such as neural networks or gradient boosting remain opaque to clinicians (Hidayaturrohman y Hanada, 2024). Even when predictive performance is strong, the lack of transparent explanations reduces clinical trust and impedes integration into decision-making. Few HF or MCC prediction studies have systematically applied explainability frameworks such as SHAP or counterfactual reasoning to provide human-understandable insights aligned with clinical reasoning.

Fourth, the scope of current models is narrow. In HF, most studies focus on short-term outcomes such as 30-day readmissions, while long-term prognostic models are rare (Zhuang et al., 2021). In MCC, models often exclude social and functional

determinants of health, despite their known impact on readmission risk. Multimodal integration remains underexplored: most work relies solely on structured EHR variables, with limited incorporation of unstructured notes, imaging, patient-reported outcomes, or caregiver-reported measures. This contrasts with oncology and neurology, where multimodal fusion has consistently improved predictive accuracy (Banapuram et al., 2024).

Taken together, these limitations define a clear research gap: there is currently no established framework that integrates heterogeneous clinical, functional, and social data into interpretable predictive models for hospital readmissions in either HF or MCC. Addressing this gap is essential for the development of clinically valid, trustworthy, and generalisable tools that can support outcome prediction and personalised care in chronic and complex conditions.

In this dissertation, the ReIC case study responds to this gap by developing and validating interpretable ensemble models for 30-day readmissions in patients with heart failure, incorporating both clinical and psychosocial variables to demonstrate the added value of explainable ML in cardiovascular care. Likewise, the RePluris case study addresses the same gap in the context of multiple chronic conditions by integrating clinical, functional, and social determinants of health into predictive frameworks that extend beyond traditional scores such as LACE, thereby advancing the state of the art in multimorbidity management.

Beyond physical health conditions, similar predictive modelling challenges arise in psychiatry, where psychological disorders present additional complexity due to subjective measures and variability in recovery trajectories.

2.2 Machine Learning in Psychiatry and Psychology

In recent years, machine learning has gained traction in psychology and psychiatry as a powerful set of tools for detection, prediction, and treatment optimisation. By leveraging large and often heterogeneous datasets, ML is beginning to provide insights into complex mental health conditions, including eating disorders, depression, and substance use, while also supporting the development of digital mental health interventions (Tuena et al., 2022).

One of the most prominent applications is in detection and diagnosis. Supervised learning algorithms have been trained to predict eating disorder status from survey responses, social media data, or neuroimaging, often achieving promising levels of accuracy (Aryal et al., 2020; Fardouly et al., 2022). More broadly, ML has been employed to identify symptoms and risk factors across a wide spectrum of psychiatric conditions, to model disease progression, and to support the differentiation of

closely related diagnoses (Thieme et al., 2020). These models frequently combine psychometric tests, neuroimaging, clinical histories, and, in some cases, genetic or physiological markers, demonstrating the value of multimodal integration in improving diagnostic precision (Chandler et al., 2020).

ML is also widely used in predicting treatment outcomes, where the goal is not only to anticipate whether a patient will respond to a given intervention but also to determine which treatment is likely to be most effective for a particular individual (Aafjes-van Doorn et al., 2020). In the context of pharmacological therapies, most work has centred on antidepressants prescribed during acute episodes of depression, with reported accuracy rates of 60–65%. In psychotherapy research, both prognostic models that estimate the likelihood of recovery with a specific therapy and prescriptive models that identify the most suitable treatment among alternatives have been developed. These advances signal a gradual shift towards personalised treatment planning in psychiatry (Chekroud et al., 2021).

Beyond diagnostic applications, ML also shapes intervention delivery itself, both in guiding treatment selection and in powering novel digital platforms. Chatbots, for example, are being used to deliver prevention programmes and therapeutic content. In eating disorders, a chatbot-based intervention reduced weight and shape concerns in women at high risk, with improvements persisting at six-month follow-up (Fardouly et al., 2022). Internet-delivered cognitive behavioural therapy (iCBT) has been enhanced by ML techniques, offering scalable and accessible treatment solutions that reduce barriers to care. Such developments suggest that algorithmic support may play a growing role in the dissemination of psychological interventions (Chekroud et al., 2021).

Applications extend to substance use disorders, where ML models detect current abuse, predict future risk, and forecast treatment success. These studies draw on physiological measurements, longitudinal surveys, medical records, and social media, illustrating the adaptability of ML across contexts. Similarly, in neuropsychological assessment, machine learning methods are applied to streams of behavioural and cognitive data, such as speech and movement, to automatically infer aspects of cognitive function, offering an efficient complement to traditional assessment methods (Barenholtz et al., 2020).

The success of these approaches rests heavily on the availability of diverse data sources. Electronic health records, genetics, neuroimaging, electrophysiology, and cognitive testing are all increasingly incorporated into ML pipelines. Social media, in particular, has become a notable source for eating disorder research, given the prominence of appearance-related images and online communities that can reveal early signs of risk. EHRs provide another valuable foundation, offering routinely

collected longitudinal data that can support predictive models directly integrated into clinical care (Chekroud et al., 2021).

Nevertheless, significant limitations and challenges remain. Most studies to date have not been externally validated, limiting confidence in their generalisability. Few prediction tools have advanced to implementation studies or clinical deployment. Moreover, the demographic bias of existing research is a major concern, as discussed further in the ED context. These issues are particularly salient for eating disorders, where heterogeneity and relapse complicate prognosis, as discussed next, with emphasis on recovery trajectories and the predictive use of psychometric questionnaires.

2.2.1 Prediction of Treatment Outcomes and Recovery Trajectories in Eating Disorders

Another area of growing application is the study of recovery trajectories in eating disorders (ED), where individual variability and high relapse rates make prognosis particularly challenging. Recent work demonstrates that predictive modelling can help stratify patients by expected response, identify risk factors for poor prognosis, and inform personalised treatment planning.

A broad range of predictors has been used in these studies. Baseline clinical characteristics, such as ED diagnosis, symptom severity, body mass index (BMI), binge eating frequency, and compensatory behaviours, consistently feature as strong predictors of outcome (Haynos et al., 2020; Espel-Huynh et al., 2021; Espel-Huynh et al., 2020). General psychopathology levels and comorbidities are also important (de Vos et al., 2023). Demographic and historical variables, including age, ethnicity, illness duration, treatment history, trauma exposure, and psychiatric hospitalisations, further enrich predictive capacity (Haynos et al., 2020; Gorrell et al., 2023). In addition, early indicators of treatment response, such as symptom change during the first 2–6 weeks, motivational factors, and hope for recovery, have emerged as robust prognostic signals. Psychological constructs like emotion regulation ability, treatment readiness, and insight add further explanatory value, underscoring the multifactorial nature of ED recovery (Espel-Huynh et al., 2021; de Vos et al., 2023; Espel-Huynh et al., 2020; Chapa et al., 2020).

The outcomes being predicted span multiple domains. Studies frequently identify distinct treatment response trajectories, typically clustering patients into rapid responders, gradual improvers, and static or non-improving groups (Espel-Huynh et al., 2020; Espel-Huynh et al., 2021; de Vos et al., 2023). Predictive targets also include specific symptom outcomes such as changes in ED psychopathology at discharge and follow-up, binge eating frequency, compensatory behaviour patterns,

weight restoration, and overall well-being (Haynos et al., 2020). Long-term clinical outcomes, including response at one- or two-year follow-ups, clinical impairment, and overall recovery status, have also been investigated, providing valuable insights into prognosis beyond the treatment episode (Haynos et al., 2020; de Vos et al., 2023).

Methodologically, most studies rely on longitudinal observational designs with repeated assessments across treatment and follow-up (Espel-Huynh et al., 2020; de Vos et al., 2023). Sample sizes vary widely, from several hundred cases in residential and outpatient cohorts to thousands in national or app-based datasets (Chapa et al., 2020). Analytical approaches include latent growth mixture modelling (LGMM), which has been widely used to identify latent classes of symptom trajectories, and machine learning methods such as support vector machines, k-nearest neighbours, and elastic net regression (Espel-Huynh et al., 2020; de Vos et al., 2023). These ML models are often benchmarked against traditional approaches such as logistic regression. Performance varies across cohorts and designs. In some studies, machine learning offers only marginal improvements over regression, whereas in others elastic net or related methods show clearer advantages (Haynos et al., 2020; Espel-Huynh et al., 2020). In all cases, model performance is evaluated using standard indices such as AUC, entropy, and posterior classification probabilities. Note that these figures derive from different cohorts and study designs; they should be interpreted as indicative rather than directly comparable.

The measurement strategies are consistent across studies, drawing on validated psychometric instruments. The Eating Disorder Examination Questionnaire (EDE-Q) remains the most widely used measure of ED psychopathology, complemented by tools such as the Mental Health Continuum–Short Form for well-being and the Progress Monitoring for Eating Disorders scale for weekly symptom tracking. These instruments are typically administered at baseline, during treatment (often weekly or monthly), and at discharge or follow-up. Missing data are handled using robust techniques such as maximum likelihood estimation, and statistical analyses are conducted using packages including Mplus, R (lcmm), and SPSS (de Vos et al., 2023). The ubiquity of these instruments and their standardised administration motivate the present dissertation's focus on psychometric data as a primary predictive source.

Overall, the literature highlights both the potential and limitations of predictive modelling in ED treatment. On the one hand, these models reliably identify clinically meaningful trajectory groups, demonstrate the predictive value of early symptom change and motivational factors, and offer performance levels that approach clinical utility. On the other hand, challenges remain in terms of generalisability, sample diversity, and the gap between methodological development and clinical implementation. Current studies often rely on predominantly female, white samples, which

restricts the external validity of findings. Moreover, while predictive accuracies are encouraging, they are not yet sufficient for routine decision-making. A central data source across these studies has been validated questionnaires, long established in ED research for screening and monitoring. However, their use as predictive tools raises specific challenges, as outlined below.

2.2.2 Questionnaires as Predictors in Eating Disorders

Validated questionnaires are a cornerstone of eating disorder (ED) research and clinical practice, providing efficient and standardised tools for screening, assessment, and symptom monitoring (Stankić et al., 2023); they are also the principal data source in many of the recovery-trajectory studies reviewed above. Instruments such as the Eating Attitudes Test (EAT), the Eating Disorder Examination Questionnaire (EDE-Q), and the Questionnaire on Eating and Weight Patterns (QEWP) are widely used to detect risk, identify cases, and quantify the severity of ED psychopathology. Their psychometric properties are well established, with reported sensitivities ranging from 0.07 to 0.88 and specificities from 0.0 to 1.0 depending on the instrument and study design (House et al., 2022). These ranges reflect heterogeneous instruments, thresholds, and sampling frames. Despite their ubiquity in assessment and screening, the use of these questionnaires as predictors in machine learning models for treatment outcomes remains relatively underdeveloped.

The primary applications of these instruments have historically focused on detection rather than prognosis. The EDE-Q and QEWP, for example, have been extensively validated as screening tools, and several studies employ multiple questionnaires in combination to provide a comprehensive picture of symptom severity and quality of life (House et al., 2022). These measures are effective for assessing the current state of patients and capturing symptom intensity, which is frequently used as a baseline reference for clinical monitoring. In many cases, research highlights the importance of using multiple questionnaires simultaneously, especially when exploring associations between symptom severity and quality of life in ED patients (Stankić et al., 2023).

A smaller but growing body of research has applied machine learning methods to questionnaire data for predictive purposes. Studies have attempted to forecast treatment-related outcomes such as uptake, adherence, dropout, and symptom-level change in digital interventions (Linardon et al., 2022). While these applications demonstrate the feasibility of predictive modelling with questionnaire data, their added value over traditional regression-based approaches remains inconsistent. In one study, machine learning offered no clear performance improvement over generalised linear models for most outcomes. In another, ML was employed to enhance the prediction of ED onset and differential diagnosis from risk factors captured via

self-report (Krug et al., 2021). These findings suggest that while questionnaires provide rich information, their predictive capacity may be limited when used in isolation.

Beyond symptom detection, questionnaires are also widely employed to assess quality of life in ED patients (Stankić et al., 2023). This line of research consistently shows that a combination of several instruments provides the most reliable basis for monitoring patient well-being and preventing clinical deterioration. Measures of quality of life are increasingly recognised as important outcomes in their own right and as potential predictors of long-term recovery. However, the predictive utility of these constructs in machine learning models has yet to be fully realised.

Overall, while validated questionnaires are indispensable for screening and assessment, their role in outcome prediction remains constrained. Studies indicate that a limited set of routinely measured baseline variables is often insufficient for machine learning to outperform traditional models, and responsiveness to digital interventions has proven particularly difficult to predict. These limitations highlight a gap between the extensive use of questionnaires in ED research and their integration into robust predictive frameworks. Future work will need to focus on combining questionnaire data with other modalities, such as physiological, behavioural, or ecological momentary assessments, to enhance their predictive power and unlock their full potential in supporting personalised treatment and recovery trajectories.

2.2.3 Challenges and Research Gap in Questionnaire-Based Prediction of Eating Disorders

Although validated questionnaires such as the EDE-Q and EAT are indispensable for screening and assessment, their use as predictive tools in machine learning applications for eating disorder (ED) outcomes remains constrained by several methodological, demographic, and clinical challenges.

First, predictive performance is limited. Baseline questionnaires alone are frequently insufficient for accurate forecasting of treatment outcomes. When multiple machine learning methods were applied to a set of 36 baseline variables, performance was poor for predicting engagement-related outcomes, with AUC values close to chance levels (0.48–0.52). Responsiveness to digital interventions has also proven difficult to predict, with highly variable response rates across samples. These findings suggest that questionnaire-based predictors do not capture the full complexity of recovery and treatment trajectories.

Second, data quality and sample size represent major constraints. Many studies rely on relatively small or homogeneous cohorts, reflecting the challenges of obtaining

large datasets in a population with relatively low prevalence. Small, homogeneous samples, missing data, and inconsistent measurement further compromise performance. Most existing models are trained on individuals seeking treatment, which reduces generalisability to undiagnosed or untreated populations and introduces selection bias.

Third, demographic and cultural biases undermine external validity. The vast majority of studies are conducted in samples of young white women, limiting generalisability to males, individuals with obesity, or culturally diverse populations. Language bias is also apparent, with nearly all studies conducted in English and very limited research in other linguistic and cultural settings. These limitations raise concerns about the equity and applicability of questionnaire-based predictive models.

Fourth, methodological gaps persist. Much of the research is cross-sectional, with limited longitudinal modelling of change over time. Even when longitudinal data are available, predictors are often measured only at baseline, whereas repeated assessment throughout treatment may provide richer prognostic information. Constructs such as motivation or emotion regulation, for example, may evolve dynamically across the intervention phase and could be stronger predictors if modelled longitudinally. Few studies systematically involve clinicians in model design, validation, or interpretation, reducing clinical credibility and hindering integration into practice.

Finally, clinical complexity remains a fundamental challenge. Distinguishing between behaviours such as overeating and binge eating is difficult, as DSM-5-based screening items do not always achieve satisfactory accuracy. High rates of comorbidity and diagnostic crossover exacerbate these issues: approximately one third of patients continue to meet ED criteria five or more years after initial treatment, and up to 40% experience shifts between diagnostic subtypes. Such heterogeneity reduces the reliability of questionnaires as sole predictors of outcomes.

Taken together, these limitations define a clear research gap: there is currently no systematic use of interpretable machine learning with psychometric-only data for predicting risk and recovery in eating disorders. This dissertation addresses that gap by developing interpretable models based on psychometric data and by evaluating their generalisability and clinical relevance.

2.3 Multimodal Learning for Human-Centered Assessment

Multimodal learning has emerged as a promising paradigm for assessing human cognition and behaviour, particularly in contexts where single data sources fail to capture the full complexity of psychological and cognitive processes. By integrating heterogeneous data types, ranging from physiological signals to behavioural cues, these approaches enable richer, more ecologically valid assessments that extend beyond the limitations of traditional psychometric methods.

One of the most active research areas is the use of Extended Reality (XR) technologies, which include virtual, augmented, and mixed reality platforms. These immersive environments enable real-time multimodal data collection by combining measures such as galvanic skin response, electroencephalography, eye and hand tracking, and motion capture. Compared with conventional tasks, XR assessments embed cognitive demands into dynamic and realistic scenarios, enhancing ecological validity. For example, memory tasks that typically involve simple recall can, in XR settings, require participants to navigate spatial arrangements or perform sequential actions in interactive worlds, such designs not only yield more naturalistic assessments but also open avenues for adaptive, personalised interventions (Gonzalez-Erena et al., 2025). Creativity assessment represents a particularly relevant domain where multimodal integration can address long-standing limitations of traditional methods.

Beyond immersive environments, multimodal learning has been applied to integrate diverse data modalities for cognitive assessment. Multimodal learning analytics (MMLA) has gained prominence in educational and psychological contexts, combining video, audio, physiological signals, and digital interaction data to analyse behaviour in detail (Guerrero-Sosa et al., 2025). Studies have identified natural language, vision, sensor data, human-centred signals, and activity logs as the primary modality groups. Vision-based and human-centred modalities, such as gaze, posture, or prosodic speech, are especially prevalent. Most investigations employ between two and five modalities, with results consistently demonstrating that multimodal integration outperforms unimodal approaches (Cohn et al., 2024).

A particularly impactful direction involves the integration of physiological and behavioural data. Research shows that combining EEG signals with observable behavioural indicators, such as facial expressions and body posture, yields predictive accuracies significantly higher than those achieved with any single modality. This synergy highlights the complementary value of physiological and behavioural signals in understanding cognitive and affective states, suggesting that integrated approaches provide a more comprehensive picture of human functioning (Babu et

al., 2018).

Multimodal methods have also proven useful for examining collaborative learning and group interactions. By analysing video and audio recordings, researchers can infer patterns of communication, observation, and resource management within groups. Process mining techniques applied to these datasets provide insights beyond what traditional statistical methods reveal, showing, for instance, that groups with strong collaboration often cycle between monitoring progress and taking collective action. Such approaches demonstrate the potential of multimodal analytics to capture social as well as individual dimensions of cognitive performance (Khan, 2017).

Central to this field are the fusion strategies used to combine modalities. Mid-level fusion, which merges features extracted from different sources before joint analysis, is the most common, followed by hybrid fusion approaches that combine elements of early and late fusion. Structured, model-based analytical approaches dominate, often supported by classification and statistical methods to predict outcomes and explore correlations. Qualitative techniques, such as thematic coding, remain valuable complements, particularly in educational and psychological contexts where interpretability is crucial (Cohn et al., 2024). Such methodological advances are particularly relevant for the dissertation, which explores multimodal integration of visual and textual data for creativity assessment.

Building on these methods, researchers are now developing frameworks for multimodal cognitive assessment. These systems integrate streams of data such as audio, video, and activity logs, modelling temporal dynamics to provide objective measures of behaviour and cognition (Khan, 2017). Prototypes include tools that combine handwriting captured through digital pens with electrodermal activity sensors to assess fine-grained motor and affective processes, exemplifying the practical implementation of multimodal frameworks in cognitive science and clinical domains (Niemann et al., 2018).

Despite encouraging progress, significant challenges and limitations remain. Many XR applications underutilise the full potential of multimodal integration, relying heavily on visual and auditory inputs while neglecting physiological or sensor data. Practical issues such as cybersickness, usability, and accessibility also hinder adoption in real-world settings. More broadly, the field suffers from data scarcity: studies are often conducted on small samples, limiting the generalisability of findings and restricting the use of data-intensive approaches such as deep learning. Issues of standardisation, interoperability, and scalability further complicate progress, highlighting the need for larger datasets, shared benchmarks, and interdisciplinary collaboration (Ramanarayanan, 2024).

Creativity assessment provides a compelling case study for multimodal learning. Like cognition and behaviour, creativity is multifaceted and not easily captured by a single modality. The following subsection reviews how multimodal methods have been applied to creativity assessment and highlights the remaining gaps.

2.3.1 Creativity Assessment: From Subjective Ratings to ML Automation

The assessment of creativity has traditionally relied on human experts, who evaluate artefacts such as drawings, essays, or designs according to subjective criteria. While expert judgement provides valuable insights, it is costly, time-consuming, and difficult to scale. In recent years, machine learning has emerged as an attractive alternative, offering automated methods that can support or even replace manual assessments in both educational and professional contexts. The introduction of multimodal learning has further enriched this field by enabling the integration of diverse data sources, such as sketches, text, code, images, and sound, to capture creativity in a more holistic and theoretically grounded manner.

One of the most prominent strands of research has been the application of multimodal learning frameworks to design and programming contexts. For example, Song et al. (2023) proposed an attention-enhanced multimodal model that processes both text and sketch data from early design concepts, predicting five key metrics: drawing quality, uniqueness, elegance, usefulness, and overall creativity. Their results show that multimodal models outperform unimodal approaches, improving explanatory power by 5–12%, with the cross-attention mechanism yielding a further 5–9% gain. In programming education, Kovalkov et al. (2021) developed a framework for Scratch projects that integrates three modalities, code, images, and sound. By applying distance-based measures of fluency, flexibility, and originality to each modality and then combining them through machine learning, the system achieved alignment with expert assessments that was as high, or higher, than agreement levels among experts themselves.

Machine learning has also been applied successfully to visual art and creativity testing. Zhang et al. (2024) employed a convolutional neural network to automatically evaluate human paintings, achieving accuracies of up to 90% and outperforming human raters in terms of speed. Similarly, Spee et al. (2023) used Random Forest regression models to predict creativity judgments based on 17 subjective art attributes, explaining around 30% of the variance in human ratings. In the context of established psychometric instruments, Cropley et al. (2024) developed an open-source, image-based, multi-label classification system for the Test of Creative Thinking – Drawing Production (TCT-DP), reporting accuracy levels comparable to those of

trained human assessors.

These systems demonstrate clear advantages over traditional assessment methods. Whereas conventional evaluations of creativity are labour-intensive and difficult to scale, ML-based approaches can deliver rapid, accurate, and cost-effective assessments that are easily integrated into educational or professional workflows. Importantly, performance comparisons show that the agreement of automated systems with expert assessments is at least as strong as the agreement between different human experts. In some cases, the combined ML models have achieved higher consistency than expert raters themselves, indicating that algorithmic scoring can provide a stable and replicable foundation for creativity evaluation.

The development of these approaches is grounded in creativity theory, which traditionally emphasises fluency, flexibility, and originality as core constructs. ML and multimodal methods extend these theoretical frameworks by going beyond simple counts of ideas or features, using distance-based metrics and representation learning to capture more nuanced relationships between creative outputs. In doing so, they not only automate assessment but also enrich the theoretical understanding of creativity.

However, while machine learning has been applied to automate the scoring of drawing-based creativity tests such as the TCT-DP, these efforts remain strictly unimodal, relying only on visual features of the drawings. By contrast, research in design and programming has demonstrated that multimodal approaches, which combine sketches or code with textual descriptions, significantly improve the validity and explanatory power of creativity assessment. Yet, no established multimodal framework jointly evaluates creativity through both drawings and accompanying text. This absence highlights a critical research gap: the potential of multimodal integration for drawing-based creativity assessment remains largely unexplored.

2.3.2 Challenges and Research Gap in Multimodal Creativity Assessment

Despite encouraging progress, the field of multimodal learning for human-centred assessment faces several persistent challenges. A first limitation concerns data availability and scalability. Most studies rely on small or homogeneous samples, which constrains generalisability and restricts the use of data-intensive methods such as deep learning. This scarcity is particularly problematic in creativity research, where collecting large annotated datasets is costly and subjective judgements introduce additional noise. Without larger, shared datasets and benchmarks, reproducibility and comparability across studies remain limited (Ramanarayanan, 2024).

Second, issues of standardisation and interoperability hinder integration across modalities. Heterogeneous sensor formats, inconsistent feature extraction pipelines, and domain-specific tools make it difficult to align multimodal signals in a coherent framework. This is evident in XR environments, where physiological, behavioural, and interaction data are often collected but rarely integrated beyond simple visual–auditory fusion. For creativity assessment, similar problems arise when attempting to combine artefacts such as drawings, text, and behavioural traces, leading to fragmented analyses that overlook important cross-modal relationships.

Third, practical and usability concerns restrict ecological validity. XR-based multimodal systems, for example, are affected by cybersickness, high equipment costs, and accessibility barriers. More generally, complex multimodal frameworks can be difficult to deploy in educational or clinical settings, where interpretability and ease of use are essential. Creativity assessment shares these challenges: while automated scoring methods outperform human raters in speed and consistency, most frameworks remain unimodal and focus narrowly on visual features, limiting adoption in practice.

Finally, methodological gaps persist in fusion strategies and theoretical grounding. While mid-level and hybrid fusion approaches have demonstrated strong performance, there is little consensus on which strategy best balances predictive accuracy and interpretability. Moreover, many multimodal studies emphasise empirical performance but neglect the theoretical constructs, such as fluency, flexibility, and originality, that underlie creativity research. This misalignment risks producing technically accurate models that lack explanatory value for psychology and education.

Taken together, these limitations highlight a clear research gap. Although multimodal learning has advanced fields such as XR-based cognitive assessment and collaborative learning analytics, its potential remains underexploited in creativity research. Existing drawing-based tests, such as the TCT-DP, continue to rely almost exclusively on unimodal visual features, while evidence from design and programming contexts demonstrates that integrating sketches or code with textual descriptions can substantially enhance validity. To date, however, there is no established multimodal framework that jointly evaluates creativity through both drawings and accompanying text. Addressing this gap requires scalable, interpretable, and theoretically grounded multimodal models that integrate heterogeneous artefacts into coherent and actionable assessments.

2.4 Explainability in Clinical Machine Learning

The integration of explainable artificial intelligence in healthcare is now a critical research priority, driven by the need to make complex models transparent, interpretable, and trustworthy (Arjunan, 2021). Over the past decade, interest has grown rapidly, particularly in clinical contexts where interpretability is essential because decisions carry life-critical consequences (Chaddad et al., 2023).

Current applications span diagnosis, prognosis, and treatment personalisation. In intensive care, explainability supports early sepsis detection. In radiology, visual rationales help verify that models attend to clinically meaningful regions. Similar approaches support models that predict postoperative complications, assess disease risk, and inform treatment planning. Deep learning systems in imaging are increasingly paired with explanation layers that provide feature-level or region-level justifications (Karako y Tang, 2024).

Common approaches include feature attribution, attention mechanisms, rule-based learners, and model-agnostic post-hoc methods. The most widely used techniques, such as SHAP, LIME, and saliency mapping, are reviewed in detail in the next subsection (Kumar et al., 2024; Falvo y Cannataro, 2024).

Despite progress, obstacles remain. Clinician trust is fragile when predictions lack intelligible rationales. Post-hoc methods approximate model behaviour rather than reveal it, raising concerns about faithfulness and stability. Workflow integration is hampered by computational costs, usability issues, and limited end-user involvement in design and validation (Petch et al., 2022). Moreover, evaluating explanation quality requires costly expert judgement, and explanation outputs inherit dataset biases, raising questions of accountability and privacy (Rasheed et al., 2021).

Beyond cultural and ethical concerns, there are also technical and implementation challenges. Explanations produced by methods such as SHAP and LIME may not always reflect the true internal logic of the underlying models (Chaddad et al., 2023). Their stability has also been questioned, with small changes in data sometimes producing markedly different explanations. Integrating these techniques into existing clinical workflows is further complicated by computational overheads and the need for collaboration among a wide range of stakeholders, including clinicians, AI developers, and regulatory bodies (Petch et al., 2022; Falvo y Cannataro, 2024).

Evaluating the quality and effectiveness of explanations remains another unresolved issue. While quantitative metrics are available, assessing whether explanations are clinically meaningful often requires expert evaluation, which is costly and time-consuming. Studies have shown that clinicians may remain dissatisfied with ex-

planation outputs even when these are technically accurate, underscoring the gap between what current methods provide and what end-users actually require (Chaddad et al., 2023).

Regulatory and ethical concerns further complicate deployment. Data privacy, accountability, and liability are ongoing issues, and the healthcare sector has called for clearer frameworks to govern the use of AI in sensitive contexts. At the same time, XAI models inherit the biases embedded in their training datasets. Given that patient populations are heterogeneous, biased models risk producing inequitable outcomes that could worsen disparities in healthcare provision (Kumar et al., 2024).

The field is shifting toward human-centred approaches that emphasise usability, reliability, and ethical soundness. Priorities include hybrid pipelines that combine transparent components with targeted post-hoc explanations, clinically grounded evaluation frameworks, and fairness-aware modelling. The overarching aim is not only accuracy but explanations that are interpretable within clinical practice (Arjunan, 2021; Chaddad et al., 2023). To illustrate how these challenges manifest, the following subsection reviews post-hoc and user-centred explanation methods, which dominate current clinical applications.

2.4.1 Post-Hoc and User-Centred Explanation Methods

Post-hoc explanation methods remain the most widely adopted strategy for making complex machine learning models interpretable. Unlike inherently transparent models, these methods are applied after training, providing approximations of how black-box models arrive at their outputs. Among them, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) dominate the literature, together accounting for the vast majority of XAI applications in healthcare and beyond. Their popularity is evident not only in research, where, for instance, nearly 70% of studies in Alzheimer's disease detection rely on one of the two, but also in open-source uptake, where both frameworks have mature open-source ecosystems (Vimbi et al., 2024).

SHAP is grounded in cooperative game theory and calculates the marginal contribution of each feature to a model's prediction by systematically considering different feature coalitions. It provides both global explanations, which rank overall feature importance, and local explanations for individual predictions . LIME, in contrast, is a purely local method that approximates the behaviour of a complex model with a simpler surrogate in the neighbourhood of a given instance. This makes LIME intuitive for case-level interpretation, particularly when clinicians require insights for individual patients (Salih et al., 2024). Extensions such as TreeSHAP and DeepSHAP adapt these methods to specific model families or large-scale data (Alvanpour

et al., 2025; Chen et al., 2022).

These techniques have been applied across diverse domains. In healthcare, SHAP and LIME are frequently used to interpret multimodal disease models based on MRI, EEG, genetic, and clinical data, as well as to highlight influential image regions in radiology or identify biomarkers in sepsis prediction Vimbi et al. (2024). Outside medicine, they are used in robotics to explain grasp failures, in finance for consumer scoring and fraud detection, in cybersecurity to interpret intrusion detection systems, and in education to improve transparency in predictive models of student performance (Alvanpour et al., 2025; Chen et al., 2022; Gaspar et al., 2024; Hooshyar y Yang, 2024). Such breadth illustrates both their technical versatility and their perceived value for fostering trust in AI systems.

Despite these advances, significant limitations remain. Explanations are model- and data-dependent. Assumptions such as feature independence can yield misleading attributions with correlated predictors (Salih et al., 2024). Computational cost is substantial for complex models, and feature rankings often diverge across methods challenging consistency (Alvanpour et al., 2025). Moreover, studies report poor correlation between the feature rankings of SHAP and LIME, raising concerns about consistency and reliability (Hooshyar y Yang, 2024; Vimbi et al., 2024). Explanations may also fail to reflect the structured knowledge captured by models, producing outputs that are plausible but misleading.

Beyond these technical concerns lies a deeper issue: the limited integration of user-centred approaches in predictive pipelines. Most implementations treat explainability as a post-hoc add-on rather than a design principle. Outputs are mathematically complex and often poorly aligned with clinical reasoning, which limits their usefulness at the point of care. As the literature notes, while clinicians cannot be expected to understand the minutiae of XAI algorithms, they must at least be provided with clear and meaningful frameworks that situate predictions within familiar clinical reasoning (Salih et al., 2024).

In light of these considerations, the selection of explainability techniques in this dissertation was guided by three methodological criteria. First, cross-model compatibility: preference was given to approaches that could be consistently applied across heterogeneous model families, including regression-based models, tree ensembles, support vector machines, and deep neural networks. Second, dual-level interpretability: methods were prioritised if they allowed both global feature ranking and instance-level explanations, enabling alignment with clinical reasoning practices. Third, empirical maturity: techniques with established validation and broad adoption in healthcare research were favoured to ensure comparability and methodological robustness. These criteria informed the systematic use of SHAP

and Permutation Importance across the case studies.

The absence of medical expert involvement in the design and validation of explanations compounds the problem. Few studies place clinicians in the loop to design, refine, and validate explanations. Ground-truthing of explanations is rare, so plausibility may be mistaken for correctness, which is problematic in high-stakes contexts such as psychiatry or oncology, where erroneous interpretations could directly impact patient care (Vimbi et al., 2024).

Some attempts at more user-centred approaches are emerging. In educational AI, for example, researchers have combined SHAP and LIME within human-computer interaction frameworks to deliver more accessible explanations, showing that intuitive interfaces can improve both trust and predictive accuracy. In healthcare, the concept of incorporating user-defined baseline distributions into feature attribution has been proposed, allowing explanations to answer contrastive questions that clinicians naturally pose (e.g. why this prediction instead of another?). Such methods indicate a move toward tailoring explanations to end-user reasoning, though these remain isolated examples rather than standard practice (Vimbi et al., 2024; Chen et al., 2022).

Ultimately, the challenge of real-world implementation is to align explanation methods with the needs of diverse stakeholders, such as clinicians, patients, and regulators, rather than treating explainability as a purely technical problem. In practice, this requires interdisciplinary collaboration, routine integration of medical expertise, and the development of validation strategies that ensure explanations are trustworthy (Vimbi et al., 2024). Until then, XAI in healthcare will remain limited by a paradox: tools designed to make black-box models interpretable risk becoming opaque themselves, informative for algorithm developers but less usable for practitioners who must rely on them in clinical decision-making. Addressing these issues requires closer stakeholder involvement and validation strategies to ensure explanations are meaningful in practice.

2.4.2 Challenges and Research Gap in Clinical Explainability

The current landscape of explainability in clinical machine learning is marked by an imbalance between technical innovation and practical utility. While methods such as SHAP and LIME dominate the literature and have demonstrated their versatility across multiple domains, they remain predominantly post-hoc tools that are only loosely connected to the realities of clinical practice. The explanations they provide often lack validation against ground truth data and are not consistently interpretable to medical experts. Moreover, clinicians are rarely involved in their development or evaluation. As a result, the gap is not one of methodological availability but of

clinical integration.

What is missing are approaches that embed explainability into predictive pipelines in ways that are aligned with the workflows, cognitive needs, and decision-making practices of clinicians. Without such user-centred integration, explanation frameworks risk remaining useful mainly to AI researchers rather than healthcare practitioners. Bridging this gap requires systematic involvement of stakeholders, rigorous external validation, and the design of explanation strategies that are transparent, trustworthy, and interpretable within real-world medical settings.

2.5 Summary of Research Gaps

This chapter has reviewed machine learning applications in healthcare, psychiatry and psychology, explainability, and multimodal human-centred assessment. Across these domains, several recurring limitations emerge that constrain progress and motivate the present dissertation.

A first challenge is generalisability and validation. In healthcare, predictive models for chronic conditions such as heart failure are typically trained on small, single-site datasets with little external validation, undermining their applicability across populations. Similar issues appear in psychiatry, where eating disorder studies rely heavily on narrow demographic cohorts, often young white women, which reduces representativeness and fairness. Without larger, standardised datasets and rigorous validation protocols, the reproducibility and robustness of findings remain limited.

A second challenge concerns interpretability and trust. Post-hoc methods such as SHAP and LIME are widely used to explain black-box models, yet their outputs are often unstable, inconsistent across algorithms, and poorly aligned with clinical or psychological reasoning. Explanations that appear plausible may not provide actionable insights, and the absence of systematic user involvement in the design and evaluation of explainability frameworks limits their credibility in practice. Building trust requires approaches that embed interpretability directly into predictive pipelines and present explanations in ways that resonate with domain expertise.

A third challenge lies in the underuse of multimodal integration. Although evidence from healthcare, education, and psychology shows that combining heterogeneous modalities improves predictive accuracy and ecological validity, most deployed systems remain unimodal. Data scarcity, interoperability issues, and the absence of agreed fusion strategies limit progress. As a result, many models capture only a partial view of human cognition and behaviour, missing the richness that multimodal signals can provide.

Taken together, these themes highlight a consistent gap: there is no established framework that combines robust validation, user-centred interpretability, and multimodal integration in a way that is both theoretically grounded and practically actionable. This gap is particularly visible in creativity assessment. Traditional instruments such as the Test of Creative Thinking – Drawing Production (TCT-DP) continue to rely on subjective scoring of visual artefacts, which is labour-intensive, inconsistent, and difficult to scale. While recent work has applied machine learning to automate creativity scoring, these approaches remain unimodal and restricted to visual features. By contrast, evidence from design and programming shows that integrating sketches, text, and other artefacts can substantially improve validity and explanatory power. Yet no framework currently exists that evaluates creativity through both drawings and accompanying text while ensuring interpretability and alignment with psychometric theory.

This dissertation addresses these gaps by developing interpretable models for human health and cognitive assessment. It draws on lessons from healthcare (rigorous validation), psychology (psychometric grounding), explainable AI (clinician- and user-centred interpretability), and multimodal analytics (fusion strategies). By integrating visual and textual artefacts into transparent and theoretically informed models, the dissertation contributes both methodological advances and applied tools for human-centred evaluation. In doing so, it bridges a critical divide between algorithmic innovation, psychometric tradition, and practical implementation.

The best way to make dreams come true is to wake up.

MAE JEMISON (1956 A.D. –)

CHAPTER

3

Predicting 30-Day Hospital Readmission in Patients with HF

This chapter contributes to the overall dissertation by establishing the methodological foundation for the prediction of 30-day hospital readmissions in patients with heart failure. It introduces the data structure, modelling framework, and explainability strategy that will later support the empirical validation of interpretable machine learning models in subsequent chapters.

The case study is motivated by the clinical and societal relevance of hospital readmissions, which are associated with worse patient outcomes and high healthcare costs. The proposed approach focuses on the construction of machine learning models capable of integrating clinical, demographic, and psychosocial variables, with particular attention to interpretability and applicability in real-world healthcare settings.

The contributions of this chapter are based on the development of a structured pipeline that encompasses dataset preparation, problem definition, model design, and integration of explainability techniques. A central challenge in this context is the heterogeneity of patient data and the difficulty of building models that are both predictive and interpretable. To address this challenge, an approach is proposed that combines ensemble learning with feature attribution methods, supported by clinician-guided feature selection. The methodological decisions made in this

chapter provide the foundation for the evaluation presented later in Chapter 7.

The chapter is divided into five sections. Section 3.1 describes the dataset, including its multi-centre origin, structure, and preprocessing steps. Section 3.2 defines the prediction task, discusses its inherent difficulties, and contrasts the proposed approach with traditional baselines. Section 3.3 details the modelling pipeline, including the selected algorithms, feature engineering, and validation methodology. Section 3.4 presents the integration of explainability through SHAP, highlighting its role in bridging predictive models and clinical reasoning. Finally, Section 3.5 concludes the chapter with a summary of the methodology and a discussion of its relevance within the overall dissertation framework.

3.1 ReIC Dataset

The dataset analysed in this case study originates from the Predictive Models of Readmission in Heart Failure (ReIC) cohort study (ClinicalTrials.gov Identifier: NCT03300791) (Garcia-Gutierrez et al., 2023). The ReIC study adhered to the principles of the Declaration of Helsinki and was approved by the Basque Country Ethics Committee (Approval Number: PI2015151). All patients provided informed consent prior to enrolment, and all data were anonymised before integration into the research database. For the use of these data in the present dissertation, an amendment to the original protocol was submitted to and approved by the same ethics committee.

The ReIC was a prospective, multicentre study conducted in five Spanish hospitals and coordinated through the REDISSEC network. The aim was to identify factors associated with early readmission after an index episode of acute decompensated heart failure (ADHF). To address this, a nested case–control design was applied, in which cases were consecutive patients readmitted within one month of discharge and controls were matched on age, sex, and ejection fraction from the same cohort who were not readmitted.

3.1.1 Eligibility and Study Flow

Patients were eligible if they were over 18 years of age and admitted for acute heart failure syndrome, either de novo or decompensated chronic heart failure, as defined by ICD-9-CM codes (428.x; 402.x). Exclusion criteria included ADHF occurring during admission for another cause, transfer from another hospital, myocardial infarction or stroke within the preceding four weeks, life expectancy below one year due to terminal illness, and inability to complete patient-reported outcome measures (PROMs) due to severe sensory or cognitive impairment or language barriers.

The flow of patients through the study is shown in Figure 3.1. A total, 1092 patients were recruited at the index admission; 96% of them were included in the analysis. During the follow-up period, 15% of patients were readmitted.

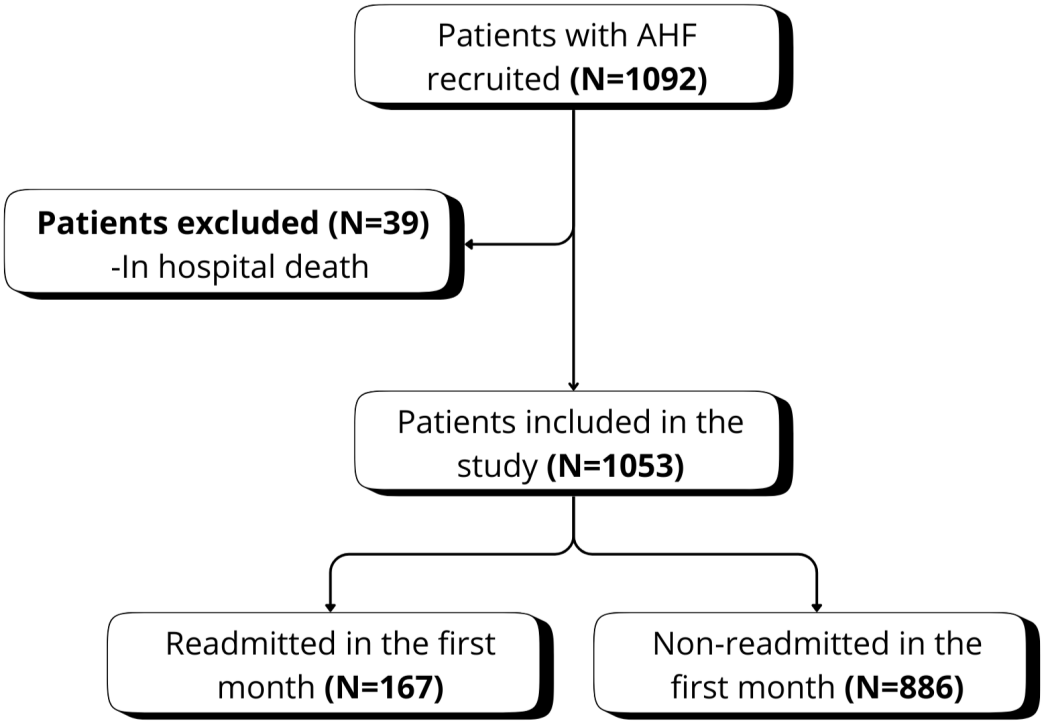


Figure 3.1: Flowchart of patient selection and inclusion in the ReIC dataset.

3.1.2 Baseline Characteristics

The cohort consisted of elderly people, with a mean age of approximately 79 years, and 47% were women. Most patients had chronic rather than de novo heart failure. Compared with controls, cases had a greater frequency of prior admissions and emergency department visits, and a higher prevalence of arrhythmias. No substantial differences were observed between groups in comorbidities such as hypertension, diabetes, or COPD, nor in the precipitating factors for the index admission. These characteristics are summarised in Table 3.1.

Table 3.1: Baseline sociodemographic and clinical characteristics.

Variable	Overall (N=1053)	No (N=886)	Yes (N=167)
Sociodemographic			

Continued on next page

Table 3.1 (continued)

Variable	Overall (N=1053)	No (N=886)	Yes (N=167)
Sex (Female)	484 (46.0)	406 (45.8)	78 (46.7)
Age (years), mean ± SD	77.98 ± 22.37	77.78 ± 23.99	79.10 ± 9.57
Heart failure history			
Known heart failure	469 (44.5)	406 (45.8)	63 (37.7)
Years of HF evolution, median [IQR]	0.0 (0.0–3.0)	0.0 (0.0–3.0)	1.0 (1.0–4.0)
Lifestyle			
Smoking status			
Never smoker	650 (61.7)	556 (62.8)	94 (56.3)
Ex-smoker	304 (28.9)	247 (27.9)	57 (34.1)
Smoker	99 (9.4)	83 (9.4)	16 (9.6)
History of alcoholism	112 (10.6)	93 (10.5)	19 (11.4)
Comorbidities			
Cardiovascular risk factors	12 (1.1)	12 (1.4)	0 (0.0)
Obesity	315 (29.9)	265 (29.9)	50 (29.9)
Dyslipidaemia	551 (52.3)	450 (50.8)	101 (60.5)
Arterial blood pressure (history)[a]	856 (81.3)	714 (80.6)	142 (85.0)
Hyperuricaemia	141 (13.4)	109 (12.3)	32 (19.2)
Anaemia	223 (21.2)	177 (20.0)	46 (27.5)
Ischaemic heart disease	46 (4.4)	38 (4.3)	8 (4.8)
Angina	103 (9.8)	84 (9.5)	19 (11.4)
Arrhythmia	554 (52.6)	461 (52.0)	93 (55.7)
Valvular disease	298 (28.3)	247 (27.9)	51 (30.5)
Congestive HF (history)	433 (41.1)	349 (39.4)	84 (50.3)
Cerebrovascular disease	177 (16.8)	143 (16.1)	34 (20.4)
Peripheral vascular disease	140 (13.3)	111 (12.5)	29 (17.4)
Chronic pulmonary disease (mild)	122 (11.6)	92 (10.4)	30 (18.0)

Continued on next page

Table 3.1 (continued)

Variable	Overall (N=1053)	No (N=886)	Yes (N=167)
Chronic pulmonary disease (moderate–severe)	131 (12.4)	108 (12.2)	23 (13.8)
Kidney failure (mild)	73 (6.9)	63 (7.1)	10 (6.0)
Kidney failure (moderate–severe)	88 (8.4)	75 (8.5)	13 (7.8)
Hepatic impairment (mild)	31 (2.9)	25 (2.8)	6 (3.6)
Hepatic impairment (moderate–severe)	18 (1.7)	14 (1.6)	4 (2.4)
Tumour	102 (9.7)	86 (9.7)	16 (9.6)
Lymphoma	7 (0.7)	5 (0.6)	2 (1.2)
Leukaemia	12 (1.1)	11 (1.2)	1 (0.6)
Metastatic cancer	9 (0.9)	8 (0.9)	1 (0.6)
Cardiac investigations and index admission			
Previous echocardiographic parameters available	815 (77.4)	670 (75.6)	145 (86.8)
Admission to ICU	17 (1.6)	14 (1.6)	3 (1.8)

[a] Data are presented as n(%) unless otherwise indicated.

3.1.3 Patient-Reported Outcome Measures

A comprehensive set of PROMs was collected at baseline, covering quality of life, frailty, functional status, anxiety and depression, self-care, and social support. Compared to controls, cases reported worse quality of life on the Minnesota Living with Heart Failure Questionnaire (MLHFQ), particularly in the emotional dimension, and higher frailty scores on the Tilburg Frailty Indicator (TFI). They also had slightly higher levels of anxiety and depression measured by the Hospital Anxiety and Depression Scale (HADS). These measures are summarised in Table 3.2.

3.1.4 Follow-Up at One Month

At one month, cases and controls were reassessed either during readmission or via telephone/postal follow-up. Additional PROMs were collected, along with measures

Table 3.2: Patient-reported outcomes at baseline.

Instrument	Overall (N=1053)	No (N=886)	Yes (N=167)
MLHFQ total (0–105)	58.78 ± 26.13	58.28 ± 25.93	61.50 ± 27.06
Barthel Index (0–100)	62.38 ± 39.12	63.18 ± 38.94	57.87 ± 39.92
TFI (0–15)	7.30 ± 3.04	7.08 ± 3.14	7.94 ± 2.88
HADS–Anxiety (0–21)	10.36 ± 4.50	10.44 ± 4.52	9.86 ± 4.30
HADS–Depression (0–21)	9.78 ± 4.22	9.86 ± 4.30	9.52 ± 4.00

of symptoms, healthcare use, and medication adherence. Compared with controls, cases more frequently reported worsening dyspnea, orthopnea, oedema, dizziness, poor appetite, and weight gain. They also had more contacts with healthcare providers during follow-up, although treatment adjustments did not differ between groups.

3.1.5 Outcome Definition

The primary outcome was hospital readmission for ADHF within one month of discharge. Cases were defined as readmitted patients, while controls were non-readmitted patients matched by age, sex, and left ventricular ejection fraction.

3.2 Problem Definition

This use case addresses the prediction of unplanned hospital readmission within 30 days following an index admission for acute decompensated heart failure (ADHF). From a clinical perspective, early readmissions represent a major quality and safety concern, are associated with excess morbidity and mortality, and carry substantial operational cost. From a machine learning (ML) perspective, the task is a supervised binary classification problem with strong constraints on temporal validity, interpretability, and deployability within multi-centre workflows.

3.2.1 Formal Statement

The problem addressed in this use case can be formally stated as follows. The dataset comprises individual patient records obtained at the time of discharge after an index hospitalisation for acute decompensated heart failure. Each record contains a set of features, the outcome of interest, the hospital site where the patient was recruited, and the time of discharge. The outcome is defined as whether the patient was readmitted to hospital within 30 days of discharge, recorded as either

readmitted or not readmitted.

The features available for prediction are anchored to the point of discharge and drawn from several domains:

- *Sociodemographic and clinical history*, such as age, sex, comorbidities, previous healthcare utilisation, and cardiovascular history.
- *Index admission context*, including precipitating factors, length of stay, and investigations carried out at or near discharge.
- *Patient-reported outcome measures (PROMs)*, such as the Minnesota Living with Heart Failure Questionnaire (MLHFQ), the Hospital Anxiety and Depression Scale (HADS), the Tilburg Frailty Indicator (TFI), as well as measures of functional status and self-care.
- *Follow-up* information at one month is explicitly excluded, as the prediction task is defined strictly at the time of discharge.

The learning objective is to develop a probabilistic model that estimates the likelihood of 30-day readmission based solely on information available at discharge. This task is subject to two critical constraints: the need for explainability, ensuring that model outputs can be understood and trusted by clinicians, and the need for clinical applicability, guaranteeing that the selected predictors are feasible to collect in real-world healthcare workflows.

Model training seeks to minimise predictive error while addressing class imbalance, applying regularisation to prevent overfitting, and ensuring that only features observable at discharge are used. This avoids temporal leakage and ensures that the resulting system is valid for deployment in prospective clinical settings.

3.2.2 Data-Generating and Temporal Structure

Each instance corresponds to a unique index admission and its subsequent 30-day window. All predictors must reflect information available by (or immediately prior to) discharge time. Variables measured after discharge (e.g., one-month symptoms, post-discharge healthcare contacts) are strictly excluded from model inputs to maintain causal and operational validity. Site indicator captures heterogeneity arising from multi-centre recruitment (different clinical practices, population mix, and data capture), which must be considered during validation and in the deployment discussion.

3.2.3 Clinical and Operational Constraints

Beyond predictive accuracy, the model must satisfy:

1. **Interpretability**: predictions should be explainable to clinicians via post-hoc methods (e.g., SHAP) and/or through constrained feature sets aligned with clinical reasoning.

2. **Applicability**: feature sets should privilege variables routinely available and reliable at discharge; reduced models with clinically curated predictors are prioritised for integration.

3. **Robustness**: performance should be stable across sites and clinically relevant subgroups; evaluation will consider variability and potential drift.

4. **Ethical use**: avoid unintended harms from biased predictions; fairness and transparency considerations guide feature curation and interpretation.

3.2.4 Specific Methodological Challenges

Despite the apparent simplicity of binary classification, several domain and data issues complicate learning and evaluation:

1. **Class imbalance.** The positive class (30-day readmission) is relatively infrequent, potentially leading to biased decision thresholds and optimistic accuracy. Loss reweighting, calibrated thresholds, and metrics insensitive to prevalence (AUROC, AUPRC) are required.

2. **Missingness and measurement heterogeneity.** PROMs and certain clinical variables exhibit non-random missingness, and capture conventions vary across centres. Missingness mechanisms must be considered; imputation strategies should be auditable and compatible with downstream explainability.

3. **Multi-centre heterogeneity.** Differences across hospitals can induce distributional shift. Validation must reflect this; models should avoid learning spurious site proxies.

4. **Risk of information leakage.** Variables temporally downstream of discharge (or summarising post-discharge events) must be excluded. Derived features should be computed using only data available up to discharge time.

5. **Temporal alignment and index selection.** Patients with multiple admissions require consistent index selection to avoid correlated samples and label contamination. The present study adopts a single index episode per patient as defined in Section 3.1.

6. **Collinearity and clinical redundancy.** Strongly correlated predictors (e.g., overlapping functional scores) can destabilise feature attributions. Regularisation and careful feature grouping are used to mitigate attribution artefacts.

7. **Calibration and decision support.** For triage or care planning, calibrated probabilities are preferable to raw scores. Post-hoc calibration will be assessed, with clinically meaningful operating points reported.

8. **Explainability fidelity and cognitive load.** Post-hoc methods may produce unstable attributions in correlated settings. Explanations will be constrained to clinically meaningful feature subsets and accompanied by uncertainty awareness.

3.2.5 Scope and Exclusions

The predictive target is limited to *ADHF* 30-day readmission following ADHF. Mortality prediction, all-cause readmission, other cause readmission, and cost prediction are out of scope for model training in this chapter. Causal inference (e.g., treatment effect estimation) is not attempted; the focus is prognostic prediction to support risk stratification and anticipatory care planning.

3.2.6 Success Criteria and Evaluation Plan

Success is defined by (i) discriminative performance better than baseline clinical heuristics, (ii) clinically acceptable calibration at relevant operating points, (iii) stability across centres, and (iv) delivery of faithful, concise explanations aligned with expert knowledge and based on routinely available variables. Formal evaluation protocols, metrics, and comparative analyses are detailed in Chapter 7; this chapter focuses on the problem set-up, data definitions, and modelling design.

The methodological challenges identified above directly shaped the design of the modelling pipeline, described in the next section.

3.3 Solution Design

The methodological design of this use case was guided by the challenges described in Section 3.2, namely the presence of missing values, severe class imbalance, heterogeneous multicentre data, and the need to ensure both predictive performance and clinical applicability. To address these constraints, a multi-stage pipeline was developed that integrated preprocessing strategies, feature refinement in collaboration with clinicians, and an ensemble learning framework designed to balance

robustness and interpretability.

The issue of missing data, a pervasive problem in clinical cohorts, was tackled through two complementary strategies. On the one hand, several of the models considered, such as decision trees, random forests, and gradient boosting variants, are capable of handling missing values internally and could therefore exploit incomplete data directly. On the other hand, for algorithms requiring complete matrices, imputation techniques were applied. These included constant-value injection, mean or median substitution, and mode imputation depending on the variable type. By combining these approaches, the dataset could be retained in its entirety while maintaining consistency across different modelling families (Garcia-Gutierrez et al., 2023).

Another major challenge was the imbalance between readmitted and non-readmitted patients, with 84% of patients not being readmitted versus only 16% being readmitted. This imbalance was first addressed using conventional strategies such as synthetic oversampling (SMOTE) to generate artificial minority class samples, undersampling to reduce the number of majority class instances, and class-weight adjustment in algorithms that permitted cost-sensitive learning. However, to overcome the limitations of these isolated techniques and to stabilise predictions across different resampling procedures, a bagging-based ensemble framework was specifically designed for this study (see Figure 3.2).

In this framework, a balanced test set was first created by randomly sampling rows from the original dataset with replacement, ensuring equal numbers of readmitted and non-readmitted patients. As a result, some patients could appear multiple times, while others might not be included, but the overall class distribution was perfectly balanced. From the remaining data, n independent training subsets were generated following the same principle of random sampling with replacement and class balancing. Each training subset was then used to train m base learners, which could include decision trees, naïve Bayes classifiers, logistic regression models, support vector classifiers, or other algorithms of choice. This procedure produced a total of $n \times m$ trained models, each capturing slightly different perspectives of the imbalanced data distribution.

Predictions for the test set were obtained by applying all models in the ensemble and then aggregating their outputs through a soft voting scheme, in which predicted probabilities were averaged across models before determining the final class label. This approach combines the variance reduction properties of bagging with the robustness of probability-based consensus, thereby mitigating the bias toward majority cases while maintaining sensitivity to minority cases. The resulting ensemble thus constitutes a key methodological contribution of this dissertation, providing a

principled way of balancing real-world data and stabilising predictive performance in the highly imbalanced setting of 30-day readmissions in heart failure.

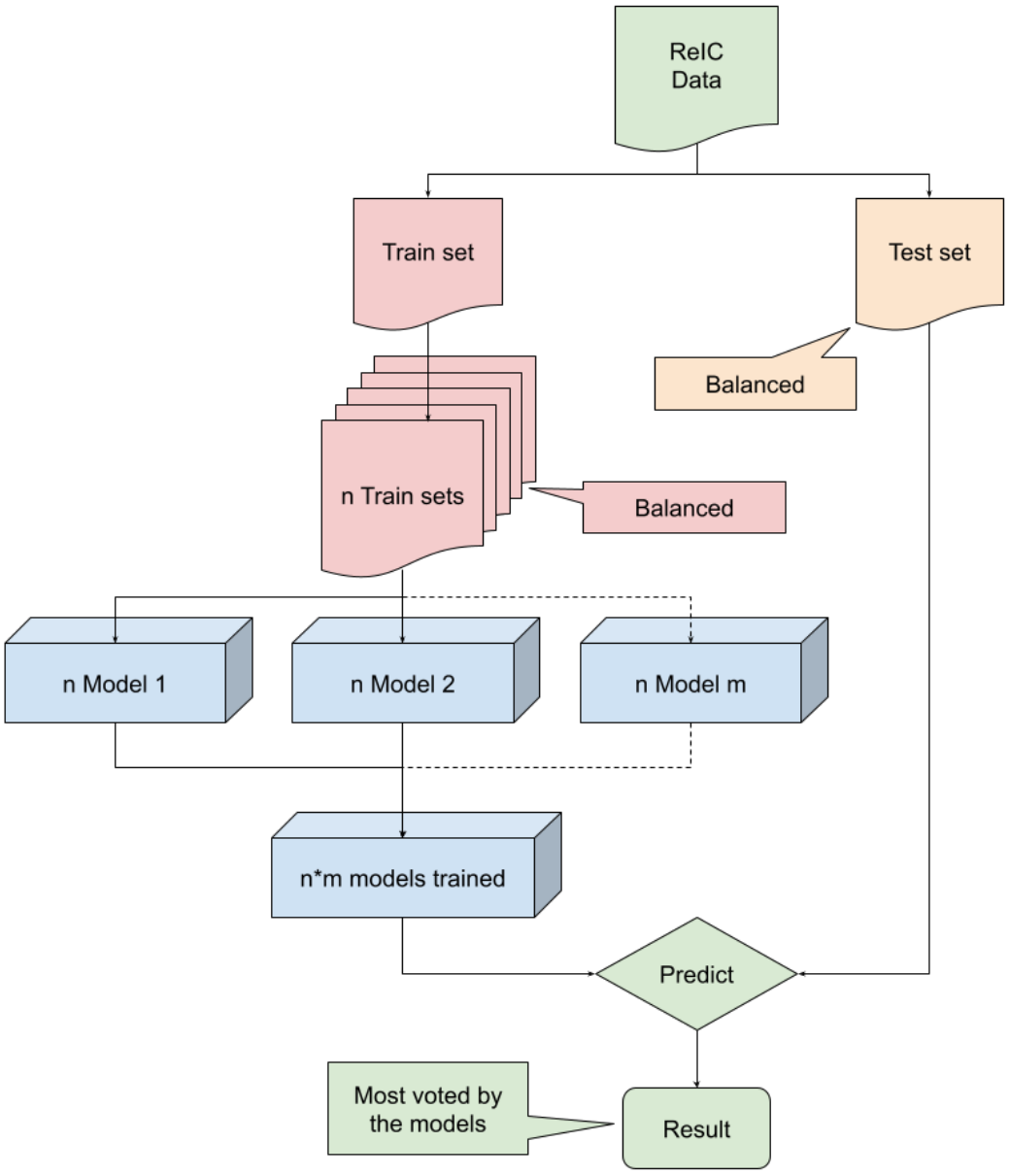


Figure 3.2: Bagging and ensemble strategy for mitigating the bias toward majority class prediction.

Since the dataset combined variables measured on very different scales, such as questionnaire scores, laboratory values, and clinical indices, a standardisation step was applied to continuous predictors. This prevented scale-sensitive algorithms,

including logistic regression, support vector classifiers, and neural networks, from being biased toward features with larger magnitudes, while leaving scale-invariant tree-based methods unaffected.

To further refine the modelling process, feature selection was carried out with the dual aim of improving predictive accuracy and ensuring clinical usability. Models trained on the full feature set were first used to estimate variable importance, and SHapley Additive exPlanations (SHAP) provided a global ranking of predictors. From the top fifty features identified in this way, a panel of clinicians selected those considered most relevant and feasible to collect in routine practice. The resulting subset reduced the dimensionality of the data by approximately 96% while preserving much of the predictive signal, thereby enhancing both interpretability and applicability in healthcare settings.

The ensemble framework therefore acted as the integrating component of the pipeline, combining balanced resampling strategies, diverse base learners, and probability-based aggregation to deliver stable predictions. By design, this approach addressed the main methodological challenges of the dataset, imbalance, missingness, and heterogeneity, while maintaining clinical applicability through the use of a refined and interpretable feature set.

Overall, the solution design mitigates the inherent limitations of real-world clinical data and provides a principled path toward reliable prediction of 30-day readmissions in heart failure. While the ensemble framework provides robustness and predictive performance, its clinical utility depends on transparency. The next section therefore introduces the explainability framework embedded throughout the pipeline, focusing on how SHAP values were used to interpret the ensemble models and evaluate their alignment with clinical reasoning.

3.4 Explainability Framework

A central element of this dissertation is the integration of explainability into the predictive pipeline, ensuring that the models developed are not only technically robust but also clinically interpretable and usable. To this end, SHapley Additive exPlanations (SHAP) were systematically applied throughout the modelling process. SHAP provides a theoretically grounded way of quantifying the marginal contribution of each feature to individual predictions, while also allowing the estimation of global feature importance across the dataset. Its suitability for heterogeneous data and compatibility with different model families made it particularly appropriate for this use case. Unlike most studies, where SHAP is applied only at the end to interpret a trained model, this framework embeds SHAP iteratively into the pipeline. Feature attribution guides model refinement, and clinician feedback ensures that

interpretability directly informs both feature selection and final ensemble design.

Explainability was incorporated at two key stages. First, SHAP values were used to generate a ranking of predictors in models trained on the full feature set. This ranking guided the feature selection process, where the fifty most influential variables were identified. Clinical experts then reviewed this shortlist, selecting those predictors that were both highly informative and feasible to obtain in routine practice. In this way, the explainability framework functioned as a bridge between statistical relevance and clinical utility, ensuring that the reduced subset of variables retained predictive power while also being grounded in medical reasoning.

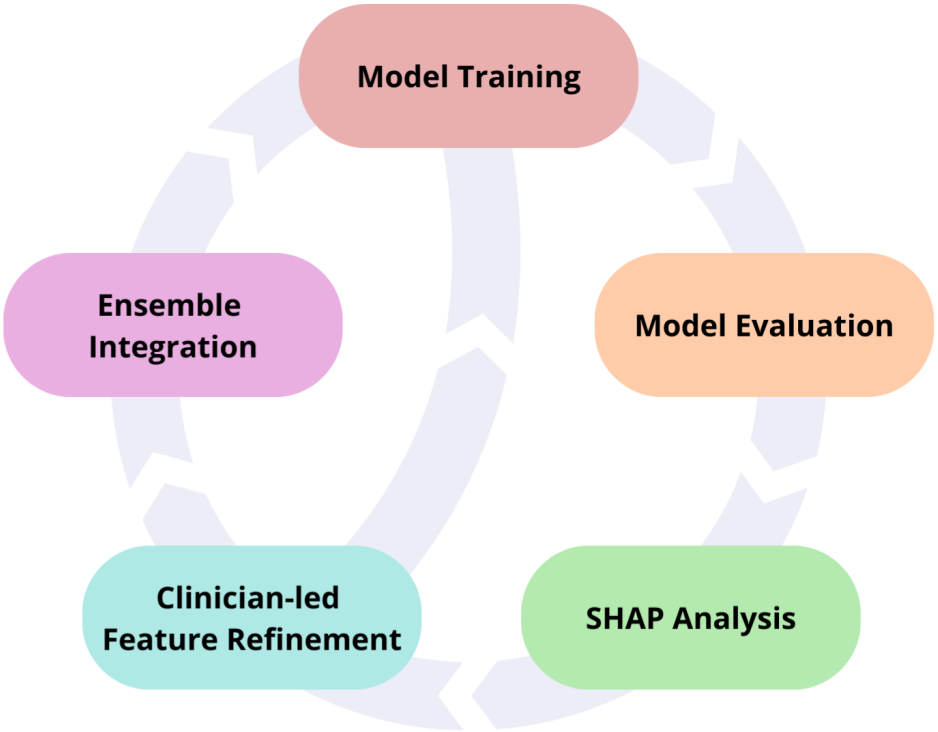


Figure 3.3: Cyclic explainability pipeline showing the integration of SHAP with clinician-driven feature refinement and model selection.

Second, SHAP was employed to interpret the behaviour of the ensemble models. For each of the base learners, local explanations were produced to highlight the factors driving predictions for individual patients, while aggregated global explanations identified consistent patterns across the cohort. These outputs were then discussed with clinicians to validate whether the models' reasoning aligned with established knowledge of heart failure readmission risk. Where discrepancies emerged, adjustments were made to the final set of models incorporated into the ensemble, privileging those whose explanatory patterns were both technically sound

and clinically credible. This iterative process, in which SHAP analysis and clinical expertise informed both feature refinement and model selection, is summarised schematically in Figure 3.3.

By embedding explainability into both feature selection and model validation, the approach ensured that predictive models were not treated as opaque black boxes. Instead, they became transparent tools capable of providing actionable insights into patient care. The continuous involvement of clinicians in interpreting SHAP outputs reinforced the trustworthiness of the models, while also demonstrating the viability of aligning machine learning predictions with medical expertise. This integration of data-driven and human-centred perspectives constitutes a core contribution of the dissertation and represents a step toward the development of predictive systems that are not only accurate but also ethically and clinically meaningful.

3.5 Summary and Conclusions

This chapter has described the methodological framework developed to address the problem of predicting 30-day readmissions in patients with acute decompensated heart failure. The work built upon the multicentre ReIC dataset, which incorporated both clinical variables and patient-reported outcomes, extending beyond conventional hospital data sources commonly employed in prior studies. This design reflects the broader aim of advancing predictive modelling in heart failure by integrating biomedical and psychosocial dimensions that are rarely analysed jointly in the literature.

Compared with existing state-of-the-art approaches, which primarily rely on regression or single-centre designs, the proposed framework introduced a unified pipeline that systematically manages the inherent challenges of real-world, multicentre data. The methodology addressed the high degree of missingness, class imbalance, and variable heterogeneity through a structured combination of data imputation, feature scaling, resampling techniques, and ensemble learning. In particular, a bagging-based ensemble strategy was designed to overcome the limitations of individual classifiers and reduce the bias towards majority outcomes, representing a methodological contribution that advances beyond standard single-model approaches used in prior heart failure studies.

Feature selection was guided by both data-driven and expert-informed processes. The incorporation of SHAP-based importance rankings together with clinical validation ensured that the resulting feature space remained parsimonious, interpretable, and clinically meaningful. This dual strategy not only enhances methodological rigour but also ensures that the models remain aligned with clinical reasoning, thus

promoting their potential integration into decision-support systems.

Explainability was embedded throughout the pipeline rather than treated as a post-hoc step. SHAP analysis provided a consistent mechanism for understanding model behaviour and for refining the predictor set through iterative collaboration with clinicians. This approach exemplifies the dissertation's broader objective of transforming predictive models into transparent, trustworthy, and human-centred tools capable of supporting clinical decision-making.

The heart failure use case directly contributes to the four objectives of this dissertation. It addresses the first by developing a dedicated machine learning pipeline for predicting 30-day readmissions that integrates both clinical and patient-reported variables. The second objective is met through the design of a reproducible framework comparable with traditional models such as logistic and Cox regression. The third is achieved by embedding SHAP-based explainability and clinician-guided refinement to ensure transparency and trustworthiness. Finally, by involving clinical experts throughout the modelling process, the study fulfils the fourth objective of ensuring real-world applicability and relevance in healthcare settings.

The evaluation chapter will assess the empirical performance, calibration, and interpretability of this framework. The methods introduced here are detailed in the peer-reviewed publication Machine learning approaches for predicting heart failure readmissions (Postgraduate Medical Journal, Oxford University Press, 2025) (Pikatza-Huerga et al., 2025a).

It is not our differences that divide us. It is our inability to recognize, accept, and celebrate those differences.

AUDRE LORDE (1934 A.D. – 1992 A.D.)

CHAPTER

4

Predicting 30-Day Hospital Readmission in Patients with MCC

This chapter contributes to the overall dissertation by extending and validating the methodological framework developed in the previous case study to a more complex and heterogeneous population. It demonstrates how explainable machine learning can be applied to patients with multiple chronic conditions (MCC), integrating clinical, functional, and social determinants of health to enhance the understanding and prediction of 30-day hospital readmissions.

The chapter addresses the methodological design for the prediction of hospital readmissions in patients with multiple chronic conditions (MCC). The case study is motivated by the clinical and societal relevance of hospital readmissions in this highly vulnerable population. Patients with MCC account for a significant proportion of healthcare utilisation and costs, and they experience elevated risks of adverse outcomes after discharge. Readmissions are particularly concerning as they often signal clinical instability, inadequate continuity of care, or complex interactions between multimorbidity and social determinants of health. Thus, predictive modelling in this context is not only a methodological challenge but also a public health priority.

The contributions of this chapter build upon the development of a structured pipeline that integrates heterogeneous data sources, including clinical indicators, functional status, patient-reported outcomes, and social determinants of health. The design of predictive models for MCC patients requires addressing critical issues such as high comorbidity burden, imbalanced event rates, and the limited external validity of traditional statistical approaches. To confront these challenges, machine learning methods are employed alongside survival analysis, supported by strategies for missing data handling, feature selection, and class imbalance management. A central component of this work is the integration of explainability techniques, which ensure that the models provide interpretable outputs aligned with clinical reasoning and patient-centred care.

This chapter is organised into five sections. Section 4.1 introduces the RePluris dataset, describing the eligibility criteria, baseline characteristics of the cohort, and the prospective data collection strategy. Section 4.2 formulates the problem definition, including the outcome of interest, the data-generating and temporal structure, and the main methodological challenges. Section 4.3 details the solution design, encompassing preprocessing strategies, model construction, and validation procedures. Section 4.4 presents the explainability framework, focusing on the interpretation of predictive factors and the clinical utility of the models. Finally, Section 4.5 summarises the findings and discusses their implications for managing patients with multiple chronic conditions.

4.1 RePluris Dataset

The RePluris project (ClinicalTrials.gov Identifier: NCT05574790) is a multicentre prospective cohort study designed to identify predictors of hospital readmission and mortality in patients with multiple chronic conditions (MCC) (Garcia-Gutierrez, 2025). Recruitment took place between 2023 and 2024 across five public hospitals in Spain, located in the Basque Country, Catalonia, and Andalusia. All centres belong to the Spanish National Health System, which guarantees universal access to care and covers both urban and semi-urban catchment areas.

4.1.1 Eligibility and Study Flow

Eligible participants were adults aged 18 years or older admitted for acute exacerbations or decompensations of pre-existing chronic diseases and classified as pluripathological, or MCC, according to the Ollero criteria. Exclusion criteria included admissions for end-of-life care, scheduled procedures, or acute events unrelated to chronic disease, as well as absence of informed consent, untraceable residence, or withdrawal during follow-up. Treatments administered during the index admission were not considered as predictors, as the study emphasised baseline, functional, and

social characteristics assessed at discharge.

A total of 1,003 patients were screened for participation, of whom 485 met the inclusion criteria. Fifty patients were excluded due to death within 30 days, leaving 435 patients for the final analysis. Of these, 388 were allocated to the training cohort (56 readmissions) and 97 to the test cohort (14 readmissions). An external validation cohort comprising 173 patients was recruited from Galdakao-Usansolo Hospital. Figure 4.1 shows the study flow.

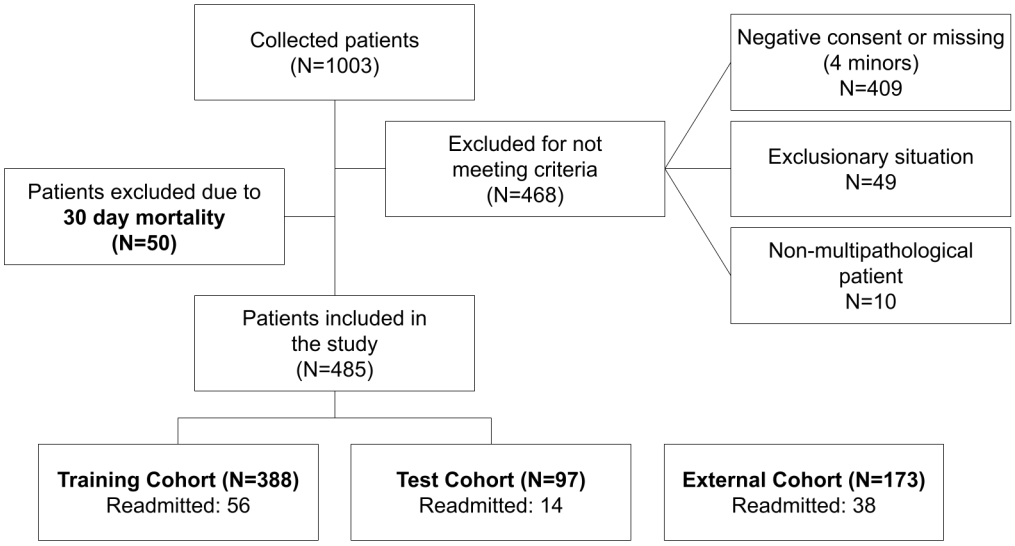


Figure 4.1: Flowchart of patient selection and inclusion in the RePluris dataset.

4.1.2 Baseline Characteristics

The mean age of the cohort was 81.9 years (±9.9), with a range between 41 and 101 years. Slightly more than half of the participants were male (52.6%). The prevalence of major chronic conditions was high: heart failure (52.2%), diabetes mellitus (48.2%), chronic kidney disease (49.3%), and chronic obstructive pulmonary disease (25.6%). Patients had a median of two chronic conditions, and almost all (97.3%) presented polypharmacy, defined as five or more drugs at discharge.

Functional and social vulnerability were also notable. Nearly half of the participants presented moderate to severe functional dependence on the Barthel Index, and the median score on the Gijón Social and Family Evaluation Scale suggested significant social risk. The mean PROFUND index at discharge was 6.6 (±5.1), reflecting a moderate to high degree of clinical complexity and frailty.

Table 4.1: Baseline sociodemographic, clinical, functional, and social characteristics of the RePluris dataset according to 30-day readmission status[a].

Variable	Overall (N=485)	No (N=415)	Yes (N=70)
Sociodemographic			
Age, years (mean ± SD)	81.9 ± 9.9	81.7 ± 10.0	83.1 ± 9.3
Male sex	255 (52.6)	211 (50.8)	44 (62.9)
Comorbidities			
Heart failure	253 (52.2)	215 (51.8)	38 (54.3)
COPD[b]	124 (25.6)	106 (25.5)	18 (25.7)
Diabetes mellitus	234 (48.2)	197 (47.5)	37 (52.9)
Chronic kidney disease	239 (49.3)	203 (48.9)	36 (51.4)
Cognitive impairment/dementia	125 (25.8)	107 (25.8)	18 (25.7)
Active neoplasia	126 (26.0)	104 (25.1)	22 (31.4)
Number of chronic diseases, median (IQR)	2 (2–3)	2 (1–3)	2.5 (2–3)
Functional status			
Barthel Index ≥95	138 (28.4)	122 (29.4)	16 (22.9)
Barthel 90–65	121 (24.9)	97 (23.4)	24 (34.3)
Barthel 60–25	143 (29.5)	125 (30.1)	18 (25.7)
Barthel ≤20	66 (13.6)	56 (13.5)	10 (14.3)
Social situation			
Lives with family/partner without conflict	211 (49.1)	178 (49.0)	33 (49.3)
Lives alone without support	17 (4.0)	15 (4.1)	2 (3.0)
Adequate social support	268 (55.3)	226 (60.3)	42 (62.7)
Gijón scale, median (IQR)	7 (5–9)	7 (5–9)	7 (5–9)
Healthcare use (previous year)			
Previous admissions, median (IQR)	2 (1–2)	1 (1–2)	2 (1–3)
Emergency visits, median (IQR)	2 (1–4)	2 (1–4)	2 (1–3)

Continued on next page

Table 4.1 (continued)

Variable	Overall (N=485)	No (N=415)	Yes (N=70)
Index admission			
Main diagnosis: heart failure	197 (40.6)	173 (41.7)	24 (34.3)
Main diagnosis: COPD exacerbation	41 (8.5)	37 (8.9)	4 (5.7)
Main diagnosis: respiratory infection	90 (18.6)	82 (19.8)	8 (11.4)
Length of stay, median (IQR)	7 (5–11)	7 (5–11)	6.5 (4–12)
ICU admission	12 (2.5)	12 (2.9)	0 (0.0)
Number of discharge medications, mean ± SD	11.6 ± 3.8	11.5 ± 3.9	11.7 ± 3.3
Polypharmacy (≥5 drugs)	472 (97.3)	403 (97.1)	69 (98.6)

[a] Data are presented as n (%) unless otherwise especified.

[b] COPD: chronic obstructive pulmonary disease.

4.1.3 Patient-Reported Outcome Measures

Patient-reported outcome measures (PROMs) were systematically collected at discharge to capture health-related quality of life, functional dependency, and treatment adherence. Table 4.2 shows a summary of the statistics of each of the them.

Table 4.2: Patient-reported outcomes at baseline reported as (mean ± SD).

Instrument	Overall (N=485)	No (N=415)	Yes (N=70)	Missing (n (%))
Frail-VIG index	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.1	0 (0.0%)
EQ-5D-5L index	0.5 ± 0.3	0.5 ± 0.3	0.5 ± 0.3	98 (20.2%)
ARMS score	16.8 ± 4.8	15.4 ± 5.3	14.9 ± 4.2	224 (46.2%)
Gijón scale	7.9 ± 7.9	7.7 ± 7.6	8.5 ± 9.5	42 (8.7%)
PROFUND index	6.6 ± 5.1	6.5 ± 5.1	7.2 ± 5.1	0 (0.0%)

The EQ-5D-5L questionnaire was employed to evaluate quality of life, with a mean index score of 0.5. Functional status was additionally assessed through the Barthel

Index, and medication adherence was measured using the Adherence to Refill and Medications Scale (ARMS). Caregiver-reported outcomes were included, with the Zarit Burden Interview capturing caregiver strain. These PROMs provided complementary information to clinical data, highlighting the multidimensional burden experienced by MCC patients and their families.

4.1.4 Follow-Up at One Month

Follow-up was performed one month after discharge through electronic health records and clinical contact to ascertain readmission events. For patients lost to follow-up due to residence changes or dropout, baseline sociodemographic and clinical information was preserved to characterise attrition. The prospective design enabled reliable tracking of readmissions and mortality within the critical 30-day post-discharge period.

4.1.5 Outcome Definition

The primary outcome was all-cause hospital readmission within 30 days after discharge. Readmissions were identified through the participating hospitals' electronic health records, capturing both planned and unplanned events. Mortality within the same timeframe was considered a competing event but excluded from the readmission analysis cohort.

4.2 Problem Definition

This use case addresses the prediction of unplanned hospital readmission within 30 days following an index admission for acute decompensation of multiple chronic conditions (MCC). From a clinical perspective, early readmissions in this population represent a major concern, as they are associated with excess morbidity, frailty progression, and high care costs. From a machine learning (ML) perspective, the task is a supervised binary classification problem with strict constraints on temporal validity, interpretability, and deployability across multiple hospitals.

4.2.1 Formal Statement

The dataset comprises patient records obtained at the point of discharge after an index hospitalisation due to exacerbations or complications of chronic multimorbidity. Each record contains a set of features, the outcome of interest, the hospital site where the patient was recruited, and the time of discharge. The outcome is defined as whether the patient was readmitted to hospital within 30 days of discharge, recorded as either readmitted or not readmitted.

The features available for prediction are anchored to the discharge time and span several domains:

- *Sociodemographic and clinical history*, including age, sex, comorbidity burden, previous admissions, and polypharmacy.
- *Index admission context*, such as primary diagnosis, length of stay, complications, and discharge destination.
- *Patient-reported outcome measures (PROMs)*, including the EQ-5D-5L for quality of life, the Barthel Index for functional status, and the ARMS scale for medication adherence.
- *Caregiver-reported measures*, such as the Zarit Burden Interview, capturing caregiver strain.
- *Social determinants*, including the Gijón social scale, household composition, and adequacy of social support.

The modelling objective is to estimate the probability of 30-day readmission based solely on data available at discharge. The two critical constraints are the requirement for explainability, so that predictions are interpretable by clinicians, and applicability, ensuring that features are routinely available in real-world workflows.

4.2.2 Data-Generating and Temporal Structure

Understanding the temporal and generative structure of the dataset is essential for ensuring validity of the models. In this case: each instance corresponds to one index hospitalisation and the subsequent 30-day observation window. Predictors are strictly limited to information available up to discharge. Post-discharge variables (e.g., follow-up consultations or new treatments) are excluded to prevent temporal leakage and maintain causal validity. The multi-centre design captures heterogeneity in clinical practice, case mix, and data collection, which must be considered during training and validation.

4.2.3 Clinical and Operational Constraints

In addition to predictive accuracy, the modelling framework was designed with several practical requirements in mind. These requirements reflect the conditions under which predictive systems could be deployed in real-world hospital workflows and the expectations of clinicians regarding transparency and usability. Accordingly, the framework must satisfy:

1. **Interpretability**: predictions must be explainable using domain-grounded subsets (e.g., frailty, polypharmacy, social support) and model-agnostic methods such as SHAP.

2. **Applicability**: predictors must be reliably collected at discharge; reduced models privileging routine clinical and social indicators are prioritised.

3. **Robustness**: performance should remain stable across hospitals and clinically relevant subgroups, reflecting heterogeneity in MCC populations.

4. **Fairness**: social and demographic features must be carefully considered to avoid unintended bias and to preserve equitable decision support.

4.2.4 Specific Methodological Challenges

Several methodological issues identified in the heart failure (ReIC) use case are also present in RePluris. These include class imbalance due to the low prevalence of 30-day readmissions, non-random missingness in both clinical and patient-reported data, and the need for well-calibrated probabilities and explainable predictions to support clinical decision-making. However, the multimorbidity context introduces additional and distinctive challenges that extend beyond those observed in single-disease populations. These are summarised below:

1. **Multidimensional and cross-domain heterogeneity.** Patients with multiple chronic conditions exhibit diverse clinical trajectories shaped by the interaction of medical, functional, and social factors. Integrating these heterogeneous sources into a single predictive framework requires harmonising variables with different scales, semantics, and reliability while maintaining interpretability across domains.

2. **Overlap and interdependence among vulnerability indicators.** Constructs such as frailty, dependency, cognitive decline, and social isolation are conceptually related and frequently co-occur, generating strong collinearity. This redundancy complicates attribution analysis and obscures the identification of distinct risk pathways. Regularisation and domain-informed feature grouping are necessary to ensure stable and meaningful explanations.

3. **Fragmentation of multimorbid profiles.** The wide variety of disease combinations and comorbidity patterns produces sparse and fragmented feature distributions. This sparsity limits the availability of comparable cases within specific subgroups, constraining the model's ability to generalise and increasing predictive uncertainty.

4. **Interpretability across clinical and social domains.** The inclusion of social and environmental determinants of health introduces new interpretive demands. Explanations must bridge medical, functional, and social reasoning to remain understandable and actionable for multidisciplinary teams involved in patient care.

4.2.5 Scope and Exclusions

The predictive task is limited to all-cause 30-day readmission following an index hospitalisation for MCC. Mortality, longer-term readmissions, cost prediction, and causal inference are outside the scope. Patients hospitalised for surgical, obstetric, or acute non-chronic events are excluded. One index admission per patient is analysed to prevent label contamination.

4.2.6 Success Criteria and Evaluation Plan

The success of the modelling pipeline was defined not only in terms of predictive accuracy but also in relation to clinical relevance and usability. Specifically, evaluation considered four criteria: (i) predictive performance surpassing baseline risk indices such as LACE, (ii) clinically acceptable calibration at operating points relevant for discharge planning, and (iii) provision of faithful, concise explanations aligned with expert reasoning, particularly regarding frailty, polypharmacy, and social determinants. Evaluation follows the protocols described in Chapter 7, with internal validation on a held-out cohort and external validation on an independent hospital dataset.

4.3 Solution Design

The solution design for the RePluris case study was conceived as a structured pipeline to ensure methodological rigour, reproducibility, and clinical applicability. The modelling process began with data preprocessing, where the dataset was split into training and test partitions, maintaining the class distribution of the outcome. An external cohort was held out for independent validation to evaluate generalisability across hospitals. Numerical features were standardised to improve comparability across variables, and missing data were handled using multiple imputation strategies, allowing the inclusion of predictors that otherwise would have been excluded.

Feature selection and dimensionality reduction techniques were applied to mitigate redundancy and multicollinearity. Approaches based on statistical criteria, projection methods, and clinically guided subsets were explored alongside models trained

on the complete feature set to establish a benchmark. This systematic process allowed the trade-off between predictive performance and interpretability to be assessed.

The modelling stage incorporated two complementary families of approaches. Survival models were developed to estimate readmission risk as a time-to-event process, while classification models addressed the binary prediction of 30-day readmission. Within both families, algorithms of varying complexity were considered, ranging from interpretable linear and rule-based methods to more flexible ensemble and neural models.

Given the relatively low proportion of readmissions, class imbalance was explicitly addressed. Oversampling, undersampling, and hybrid resampling techniques were evaluated in parallel with models trained on the original data to determine their impact on sensitivity and overall predictive stability. For models requiring parameter tuning, systematic search procedures were applied within cross-validation to optimise configurations while limiting overfitting.

Performance was assessed using multiple complementary metrics capturing discrimination, calibration, and classification quality, and confidence intervals were reported to quantify statistical uncertainty. Internal evaluation was complemented by testing on the external cohort, ensuring that the models could be assessed not only on technical grounds but also on their robustness across heterogeneous clinical contexts.

The pipeline was designed to provide a flexible and transparent framework in which different methodological choices could be systematically compared. Its design emphasised methodological comparability and clinical feasibility, laying the foundation for subsequent sections focused on model interpretability and clinical insights.

4.4 Explainability Framework

The explainability framework adopted in this case study followed the same rationale as in the ReIC case study, with SHAP as the central technique for interpretation. The overarching goal was to ensure that the predictive models could be understood in terms of clinically meaningful variables and that modelling decisions were informed by transparent analyses of feature contributions.

The process was organised in several stages, which can be seen in Figure 4.2. First, SHAP was applied to models trained with the complete feature set to obtain an overview of which variables contributed the most to the prediction of 30-day read-

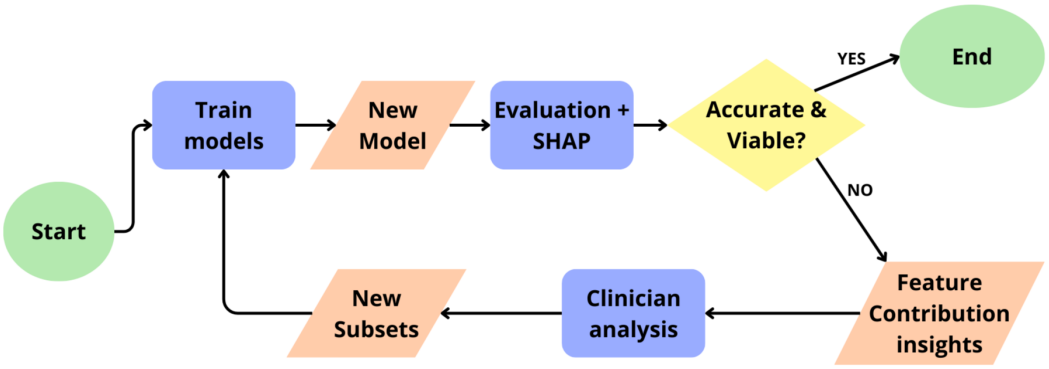


Figure 4.2: Explainability framework for the RePluris case study integrating SHAP analyses for feature selection and model validation.

mission. The same procedure was repeated for base subsets of variables that had been predefined by clinicians on the basis of their clinical relevance and viability of collection. This comparison provided an initial assessment of the added value of extending models beyond strictly clinical features to include functional, patient-reported, and social dimensions.

Second, the insights from this first stage were used to guide the construction of new subsets of features. By identifying redundant or marginally informative predictors, it was possible to design more subsets that preserved interpretability while reducing noise and improving stability.

Finally, once the final model was selected, SHAP was again employed to provide detailed explanations of how each feature contributed to the risk estimation and in which direction. This analysis allowed the identification of variables that exerted a positive influence, increasing the likelihood of readmission, as well as those whose presence or higher values reduced the predicted risk. In this way, the most relevant predictors could be consistently interpreted both at the global level, revealing general patterns across the cohort, and at the local level, indicating the specific drivers of readmission risk for each individual patient.

This iterative use of SHAP, similar to the methodology followed in the ReIC study, ensured that explainability was embedded throughout the modelling pipeline: first as a tool for understanding the data and guiding feature selection, and later as a means of validating and interpreting the final predictive models.

4.5 Summary and Conclusions

This chapter has presented the methodological framework developed for predicting 30-day hospital readmissions in patients with multiple chronic conditions (MCC) using data from the RePluris project. The study addressed a major gap in the current state of the art, where most readmission models are disease-specific, rely exclusively on clinical data, and overlook the contribution of functional, psychological, and social determinants of health. The RePluris design thus expands the modelling perspective by incorporating a multidimensional set of variables that more accurately reflects the lived reality of multimorbid patients and the complex interplay of medical and contextual risk factors.

The proposed pipeline followed a structured and reproducible process that encompassed data preprocessing, feature selection, model construction, and validation. Building on the methodology established in the ReIC study, this framework extended it to a broader and more heterogeneous population. Specific attention was devoted to handling missing values, class imbalance, and variable interdependence through a combination of imputation strategies, resampling techniques, and dimensionality reduction methods. These methodological choices were guided by the dual objective of ensuring generalisability and maintaining interpretability across clinical environments characterised by data variability and limited standardisation.

Feature selection combined model-driven importance rankings with clinical and social-domain expertise, enabling the identification of predictors that are both statistically robust and meaningful for practice. Explainability was embedded at each stage of the modelling process, allowing global and local inspection of model behaviour through SHAP-based analyses. This approach ensured that the emerging predictors of readmission risk, such as spanning clinical indicators, functional dependency, and social support, could be evaluated and refined in collaboration with clinicians and researchers.

By uniting methodological rigour with clinical and social relevance, the chapter contributes to the dissertation's overarching objectives: developing predictive systems for a representative use case (Objective 1), benchmarking their methodological advances against existing approaches (Objective 2), embedding explainability into the modelling workflow (Objective 3), and including clinicians throughout the process to produce models applicable in real-world contexts (Objective 4). The framework presented here illustrates how machine learning can be systematically adapted to capture the multidimensional nature of multimorbidity, offering a scalable foundation for ethically grounded, human-centred predictive analytics.

The forthcoming evaluation chapter will assess the predictive performance and interpretability of these models. The methodological components reported in this chapter correspond to a study conducted within the RICAPPS research network (2025).

Owning our story and loving ourselves through that process is the bravest thing that we'll ever do.

Brené Brown (1965 A.D. –)

CHAPTER

5

Eating Disorder Risk and Recovery Prediction

This chapter contributes to the overall dissertation by adapting the general methodological framework to the mental health domain, illustrating how explainable machine learning can be applied to psychometric data for the prediction of clinically meaningful outcomes. It extends the principles of interpretability and data-driven modelling to a context where psychological constructs, rather than biomedical indicators, serve as the primary predictors of patient status.

The chapter presents the methodological framework developed for predicting two critical outcomes in the context of eating disorders: the risk of developing an ED within one year and the degree of recovery among diagnosed patients. The case study is motivated by the clinical and societal impact of eating disorders, which are associated with high chronicity, elevated relapse rates, and significant burdens on quality of life and healthcare systems. The work presented here focuses on the construction of machine learning models that operate exclusively on psychometric data obtained through validated questionnaires, with particular emphasis on interpretability and clinical usability.

The contributions of this chapter are centred on the design of a dual-task predictive pipeline capable of addressing both risk and recovery. The approach incorporates a systematic preprocessing strategy tailored to questionnaire-based data, the training of classification and regression models for complementary predictive tasks, and

the integration of a two-tier explainability framework. A central challenge in this context is the reliance on self-reported measures, which are prone to missing data, response bias, and interpretability issues. To address these challenges, the proposed framework combines careful preprocessing with questionnaire- and item-level feature attribution, ensuring that the models remain both accurate and transparent to clinicians. These methodological decisions provide the foundation for the evaluation of model performance and interpretability presented later in Chapter 7.

The chapter is structured as follows. Section 5.1 describes the dataset, including its structure, scope, and limitations. Section 5.2 defines the prediction tasks, discusses the main challenges of psychometric-only modelling, and outlines the exclusions applied in this use case. Section 5.3 presents the design of the predictive pipeline, covering preprocessing, model training, and outcome definition. Section 5.4 details the integration of explainability at both questionnaire and item level, highlighting its role in bridging machine learning models and psychological theory. Finally, Section 5.5 concludes the chapter with a summary of the methodology and a discussion of its contribution to the overall dissertation framework.

5.1 ED Dataset

The dataset analysed in this use case was specifically collected for the research programme that led to the development and validation of the Resilience in Eating Disorders (RED) scale and its abbreviated version (RED-5). Unlike secondary datasets drawn from external sources, this cohort was designed and implemented from the outset to capture the multidimensional nature of eating disorder (ED) psychopathology and recovery. Its design followed COSMIN recommendations for patient-reported outcome measure (PROM) development, ensuring that data acquisition was closely aligned with the psychometric objectives of the project (Las-Hayas et al., 2025). In order to use the data collected within this project, the study protocol was reviewed and approved by the Ethics Committee of the University of Deusto, guaranteeing compliance with ethical standards for research involving human participants.

5.1.1 Study Population

Three groups of participants were recruited to provide a broad spectrum of clinical and non-clinical profiles:

- **Current patients (CP):** individuals diagnosed with an ED according to DSM-IV-TR criteria, recruited from four specialised mental health facilities. Diagnosis and inclusion were confirmed by treating psychiatrists.

- **Former patients (FP):** individuals with a history of ED who had been symptom-free for at least one year.
- **General population (GP):** participants without any ED history, recruited from the community (primarily the University of Deusto campus) to provide a normative comparison.

At baseline (T0), the total sample included 571 respondents, distributed across three distinct groups: 165 current patients (CP), 57 former patients (FP), and 349 participants from the general population (GP). This diverse composition allowed for comprehensive comparisons between clinical and non-clinical populations. After one year, a follow-up assessment (T1) was conducted, yielding retention rates of 73.9% for current patients (116 participants), 57.9% for former patients (29 participants), and 42.4% for the general population (148 participants). The overall study flow and participant progression across assessment points are illustrated in Figure 5.1. Consequently, the resulting longitudinal dataset encompasses both cross-sectional and prospective components, providing a robust framework to investigate treatment outcomes, symptom evolution, and recovery trajectories over time. The main sociodemographic and diagnostic characteristics of the sample at baseline and follow-up are summarized in Table 5.1.

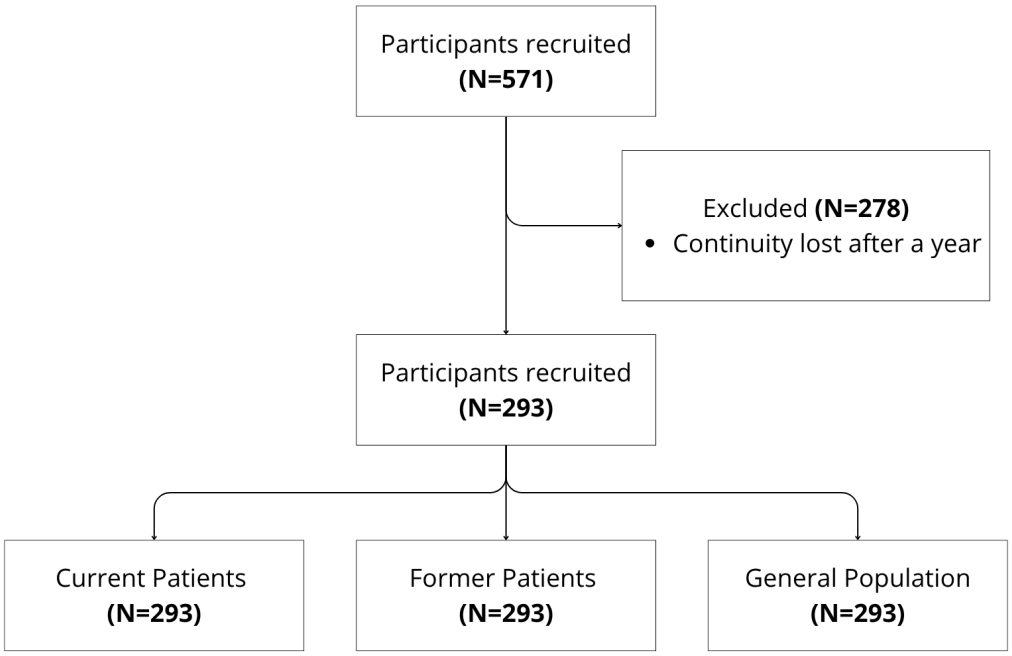


Figure 5.1: Flowchart of participant selection for eating disorders use case.

Table 5.1: Sociodemographic and diagnostic characteristics of the eating disorder dataset at baseline and follow-up.

Characteristics	Baseline (T0)			Follow-up (T1)		
	CP	FP	GP	CP	FP	GP
Female (%)	95.1	95.0	100	94.2	93.3	100
Age, mean (SD)	30.0 (9.5)	29.0 (7.4)	27.9 (7.9)	30.8 (9.8)	30.6 (8.1)	29.6 (8.4)
Age at onset, mean (SD)	19.1 (6.8)	17.2 (3.5)	–	18.6 (6.4)	17.3 (3.5)	–
Years in psychiatric treatment, mean (SD)	5.9 (6.3)	5.2 (4.4)	–	5.8 (5.8)	5.5 (4.9)	–
DSM-IV-TR diagnosis (%)						
Anorexia Nervosa (AN)	37.8	57.1	–	46.9	58.6	–
Bulimia Nervosa (BN)	28.8	17.9	–	22.1	20.7	–
EDNOS	13.5	12.5	–	12.4	6.9	–
Binge Eating Disorder (BED)	3.2	10.7	–	4.4	3.4	–
AN and BN	16.7	0.0	–	12.4	6.9	–

5.1.2 Data Collection and Instruments

All participants completed an identical set of standardised questionnaires at both baseline and follow-up, providing consistent cross-group comparisons. The battery included:

- WHOQOL-BREF for health-related quality of life (physical, psychological, social, and environmental domains).
- Hospital Anxiety and Depression Scale (HADS).
- Eating Attitudes Test (EAT-26).
- Resilience Scale (RS-25).
- Resilience in Eating Disorders scale (RED and RED-5).
- SEIQoL, an idiographic quality of life assessment based on self-defined domains.

In addition, clinical information such as age at ED onset, illness duration, number of years in treatment, and self-perceived recovery level (reported on a visual analogue scale from 0 to 100) was recorded for CP and FP groups.

5.2 Problem Definition

The predictive challenge in this use case is framed around the use of self-reported psychometric questionnaires as the sole source of information. Unlike hospital-based datasets that include laboratory results or detailed clinical records, here the emphasis lies on subjective but structured self-assessments that capture symptoms, resilience, well-being, and quality of life. The aim is to determine whether these instruments, which are inexpensive and easy to deploy, can be used to predict clinically meaningful outcomes: the onset of eating disorder symptoms in the general population and the degree of recovery in affected patients.

5.2.1 Formal Statement

Two predictive tasks are considered. The first is a **binary classification** task: predicting whether an individual will meet clinical thresholds for an eating disorder one year after the baseline assessment. The second is a **regression** task: predicting the degree of recovery among patients with a current or past eating disorder, expressed on a continuous 0–100 scale (where 0 indicates no recovery and 100 full recovery). Together, these outcomes capture both prevention and treatment perspectives.

5.2.2 Data-Generating and Temporal Structure

All predictors are drawn exclusively from baseline (T0) questionnaires completed at enrolment. These instruments include validated scales such as the Eating Attitudes Test (EAT-26), the Hospital Anxiety and Depression Scale (HADS), resilience measures (RED, RED-5), and broader quality-of-life tools such as WHOQOL-BREF. By restricting predictors to baseline, the models reflect real-world situations where only first-contact information is available and future data cannot be accessed.

The outcomes are anchored at the one-year follow-up (T1). For classification, participants are labelled as positive if their responses cross-validated thresholds for eating disorder symptoms, and negative otherwise. For regression, recovery is self-reported on the 0–100 scale. This prospective structure prevents information leakage and ensures that all predictors are available at the time of decision-making.

5.2.3 Clinical and Operational Constraints

The reliance on self-reported questionnaires brings several practical constraints:

- **Subjectivity:** responses reflect personal perception, mood, and willingness to disclose symptoms, which can introduce variability.
- **Compliance:** longer instruments increase the likelihood of incomplete responses, raising the risk of missing data.
- **Interpretation:** psychometric scores are context-dependent and may differ in meaning across populations or cultural groups.
- **Feasibility:** while questionnaires are scalable, their deployment must remain realistic in terms of patient burden and clinical workflow.

5.2.4 Specific Methodological Challenges

From a methodological perspective, four key challenges stand out:

- **Missing data:** incomplete responses are common and require imputation strategies or models robust to missingness.
- **Class imbalance:** relatively few participants develop eating disorder symptoms at follow-up, risking bias toward majority predictions.
- **High dimensionality:** questionnaires produce a large number of items relative to the sample size, making feature selection essential.
- **Variability in responses:** the subjective nature of self-reports can amplify noise, potentially reducing predictive stability.

In light of these methodological and clinical challenges, the scope of the present use case was deliberately restricted to questionnaire data alone, as outlined below.

5.2.5 Scope and Exclusions

In response to the challenges outlined above, the scope of this use case is deliberately restricted to self-reported psychometric data. Clinical history variables such as illness duration, diagnostic subtype, or body mass index were excluded, despite their availability in the original dataset. This restriction ensures that the analysis tests the predictive sufficiency of questionnaires alone and reflects contexts where clinical data are unavailable, such as large-scale prevention programs or digital health interventions.

5.2.6 Success Criteria and Evaluation Plan

Success in this use case is evaluated along three dimensions.

- **Predictive performance:** meaningful discrimination for the classification task (AUC ideally above 0.70) and adequate fit for the regression task (assessed by mean absolute error, root mean squared error, and explained variance).
- **Interpretability:** explanations should identify predictors that clinicians recognise as clinically relevant, such as eating attitudes, resilience, or anxiety.
- **Usability:** the final predictor set must be feasible to collect in routine practice, avoiding excessive patient burden.

By defining these boundaries, it is ensured that the predictive modelling task remains both viable and clinically interpretable. With this scope established, the next step is to design a solution pipeline that can address the methodological challenges identified earlier while working within these constraints.
A nested cross-validation framework is employed to provide unbiased estimates of generalisation. For classification, resampling ensures balanced folds and results are compared with simple baselines (e.g., logistic regression with all items). For regression, predictions are compared against trivial baselines such as mean recovery. In this way, improvements are measured against realistic benchmarks, ensuring that reported gains are both robust and clinically interpretable.

5.3 Solution Design

Building on the defined scope, the solution design focuses on the design of the predictive pipeline for the eating disorder use case, which was structured to address both tasks: prediction of one-year risk and estimation of recovery level, within a unified framework. Since both tasks relied exclusively on questionnaire responses, preprocessing steps were shared across them to ensure consistency and comparability. The questionnaires were originally completed in Spanish, with answers expressed on ordinal scales such as *nunca* (never), *casi nunca* (almost never), *a veces* (sometimes), *muy a menudo* (very often), and *siempre* (always), or similar. To render these responses suitable for machine learning models, they were first translated into numeric values and then rescaled to a common range. This step was essential to prevent certain models, such as support vector machines or neural networks, from being biased by differences in measurement scales across questionnaires.

Missing data represented a key challenge, as the dataset consisted of an extensive battery of instruments, and not all participants completed every item. Rather than

discarding incomplete records, which would have severely reduced the sample size, different imputation strategies were tested. These included replacing missing values with the mode or median, injecting constant values, and applying k-nearest neighbours imputers. Such approaches are particularly suitable for questionnaire data, since many items are categorical or ordinal in nature, and respondents often skip questions systematically (e.g., due to fatigue or perceived irrelevance). By carefully imputing rather than removing these cases, it was possible to retain the statistical power of the dataset while mitigating the biases introduced by missingness.

Given that the extensiveness of the questionnaire battery itself likely contributed to the presence of missing responses, a second layer of analysis focused on clinical viability. Specifically, the predictive contribution of individual questionnaires, aggregated questionnaire scores, and the full battery was systematically assessed. This strategy not only allowed the identification of the most informative instruments but also helped streamline the predictive process by emphasising measures that could realistically be implemented in clinical practice without overwhelming patients or clinicians.

A pipeline was created then for the model training and evaluation for the two predictive tasks formulated. The first aimed to classify whether an individual would be at risk of having an eating disorder one year after baseline. This was treated as a binary classification task. The second task predicted the degree of recovery on a continuous 0–100 scale, formulated as a regression problem. Although both tasks were trained on the same questionnaire inputs, they captured complementary perspectives of the patient's condition. The binary model provided a clear risk signal that is useful for screening, while the regression model offered a more nuanced representation of clinical status, avoiding the limitations of a strict "yes/no" categorisation. Together, these dual outputs provided a richer understanding of individual trajectories, aligning better with clinical needs. Figure 5.2 illustrates the dual-task pipeline, highlighting the parallel formulation of classification and regression.

The solution design directly addressed the methodological challenges identified earlier. Preprocessing ensured that heterogeneous questionnaire data could be integrated without scale-induced distortions. The use of imputation preserved the size and representativeness of the dataset while respecting the nature of questionnaire-based data. Assessing predictive power at both the item and score level offered a pragmatic response to the problem of data burden, showing how streamlined models can still provide reliable predictions. Finally, the dual-task formulation mitigated the limitations of purely binary outcomes, offering predictions that are clinically actionable and better aligned with patient-centred care.

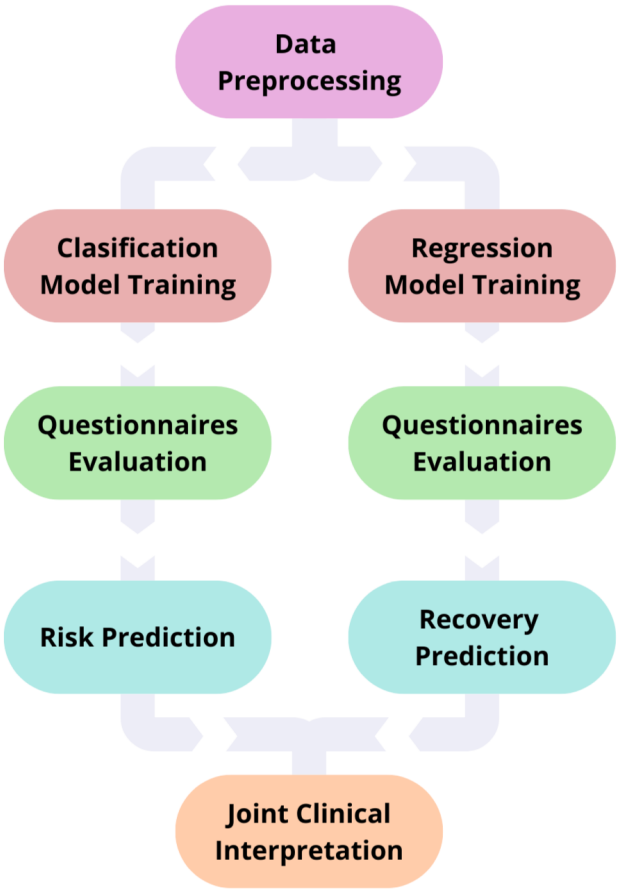


Figure 5.2: Predictive pipeline for eating disorder outcomes.

Briefly, the design combined practical preprocessing, clinically focused feature evaluation, and complementary prediction tasks into a single framework. By doing so, it not only enhanced predictive performance but also ensured that the resulting models remained interpretable and feasible for deployment in real-world mental health contexts. The next section therefore introduces the explainability framework embedded throughout the pipeline, focusing on how the different questionnaires contribute to the results and evaluate their alignment with clinical reasoning.

5.4 Explainability Framework

The explainability framework for the eating disorder use case was designed to ensure transparency in models built exclusively from questionnaire data. Given the large number of items and the heterogeneity of the instruments, interpretability was considered a fundamental requirement from the outset of model development. The framework was structured to provide insights both at the level of entire question-

naires and at the level of individual items.

The first step consisted of systematically assessing the predictive power of each questionnaire. Models were trained using: (i) all questionnaire items simultaneously, (ii) each questionnaire independently, and (iii) aggregated questionnaire scores instead of individual items. This strategy allowed a direct comparison between full item-level information and summary-level constructs, clarifying the added value of each representation. By doing so, the predictive framework was able to quantify the relative contribution of each instrument to the overall task, providing a principled basis for model selection in both the risk and recovery prediction settings.

After identifying the most suitable questionnaire representation for each predictive task, the analysis advanced to the feature level. Here, permutation feature importance was applied. This model-agnostic technique estimates the contribution of each variable by randomly permuting its values and measuring the impact on predictive performance. In this case, permutation importance was preferred over SHAP values for both theoretical and practical reasons. The dataset comprised psychometric items with substantial intercorrelations across questionnaires, which can lead additive explanation methods such as SHAP to distribute importance inconsistently among related variables. Permutation importance, by contrast, provides a stable and model-independent estimation of each feature's marginal contribution to predictive accuracy, offering clearer interpretability under multicollinearity. Moreover, its computational efficiency and consistency across model types align with the goal of assessing variable influence in a domain characterised by numerous correlated indicators. Figure 5.3 illustrates how permutation importance identifies the variables with the strongest predictive contributions. In this context, it enabled the ranking of questionnaire items according to their predictive relevance while avoiding biases associated with model-specific weighting schemes.

Together, these steps ensured that explainability was embedded at multiple levels of the modelling pipeline. The framework was not limited to global model accuracy, but explicitly examined which questionnaires and which individual items carried predictive value. This structure ensures that the insights derived from feature attribution are not only technically robust but also recognisable to clinicians, thereby enhancing their practical value.

5.5 Summary and Conclusions

This chapter has presented the methodological framework developed to predict two complementary outcomes in the context of eating disorders: the one-year risk of developing an eating disorder and the level of recovery among diagnosed patients. In contrast with the state of the art, which typically focuses on diagnostic clas-

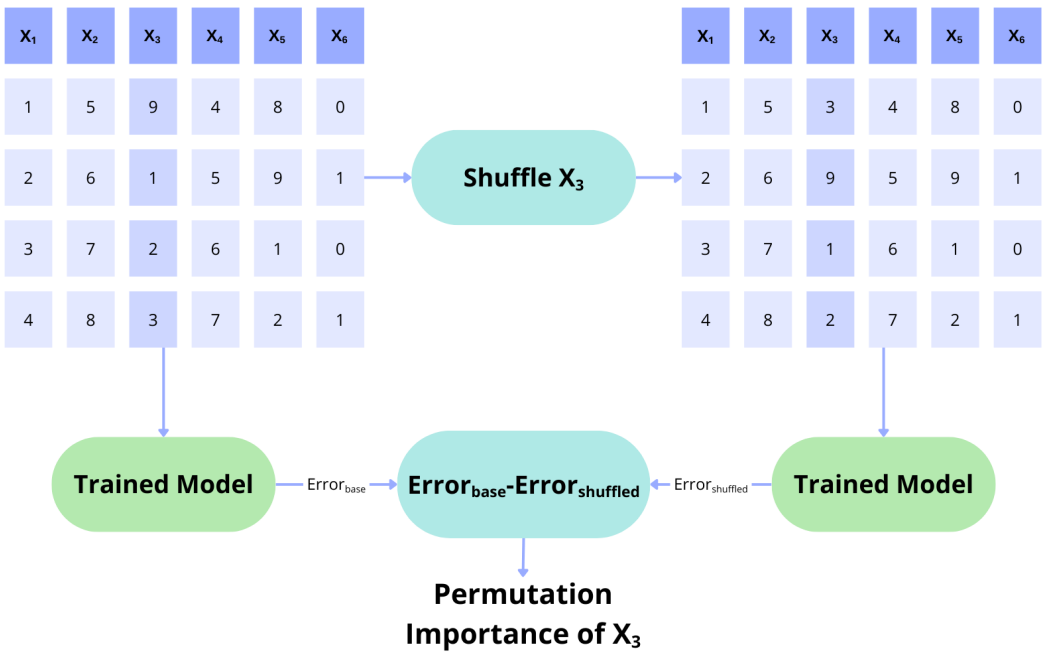


Figure 5.3: Calculation of feature importance using permutation.

sification or symptom detection using limited clinical variables, the present work proposed a comprehensive and interpretable predictive pipeline based exclusively on psychometric data. By integrating a broad set of validated questionnaires covering behavioural, cognitive, emotional, and quality-of-life dimensions, the study advanced current approaches by demonstrating that complex psychological constructs can be modelled quantitatively while retaining clinical interpretability.

The methodological design followed the unified structure proposed throughout the dissertation, encompassing data preprocessing, feature selection, model construction, and explainability. Ordinal questionnaire responses were systematically transformed and scaled, while multiple imputation strategies were applied to address item-level missingness. These steps ensured methodological robustness despite the inherent heterogeneity of self-reported data, surpassing the preprocessing depth and reproducibility standards of previous studies. The modelling pipeline was implemented for both classification and regression tasks, enabling the simultaneous analysis of risk and recovery as complementary yet distinct processes. This dual design represents a methodological advance over most existing works, which tend to address only one outcome and seldom consider longitudinal recovery prediction.

Explainability was embedded as a core component of the framework, allowing interpretation at multiple levels of granularity. Which questionnaires contributed most to prediction accuracy were identified, and permutation-based importance

scores revealed which specific items within those instruments were most influential. This multi-level interpretability structure advances beyond traditional correlational analyses commonly found in the literature, transforming psychometric data into actionable insights that can be discussed and validated by clinicians and psychologists.

The proposed framework directly contributes to the overarching objectives of this dissertation by developing interpretable machine learning models tailored to representative human-centred use cases (Objective 1), and embedding explainability to enhance psychological plausibility and expert trust (Objective 3). It also exemplifies how collaborative validation with domain specialists can ensure that model outputs remain clinically meaningful and ethically sound (Objective 4).

The forthcoming evaluation chapter will assess the empirical performance, and interpretability of this framework. The methods and design described here correspond to the peer-reviewed publication Predicting one-year risk and recovery in eating disorders using explainable machine learning (Computational and Structural Biotechnology Journal, Elsevier, 2025) (Pikatza-Huerga et al., 2025b).

I found I could say things with color and shapes that I couldn't say any other way.

GEORGIA O'KEEFFE (1887 A.D. – 1986 A.D.)

CHAPTER

6

Multimodal Prediction of Human Creativity

THIS chapter contributes to the overall dissertation by extending the methodological framework beyond clinical and psychological contexts to the domain of human cognition. It demonstrates how explainable and multimodal machine learning can be applied to the automatic assessment of creativity, integrating visual and textual information to capture complex cognitive and expressive processes.

The chapter describes the methodological design developed for the automatic assessment of creativity. Creativity is a multidimensional construct, typically evaluated through expert judgment, but such assessments are often subjective and difficult to scale. In this context, machine learning offers a promising way to complement traditional evaluation methods by providing systematic, multimodal analyses of creative artefacts.

The case study presented here focuses on drawings and their accompanying titles, which were collected under controlled conditions and subsequently annotated by experts along several creativity dimensions. This dataset allows the exploration of models that integrate visual and textual information, aiming to provide both predictive capability and interpretable insights.

The chapter is organised as follows. Section 6.1 introduces the dataset, including participants, collection procedures, and annotated creativity dimensions. Sec-

tion 6.2 defines the prediction problem and the main methodological challenges. Section 6.3 outlines the solution design, describing the modelling approaches developed. Section 6.4 presents the explainability framework, highlighting how the relative contributions of different modalities were analysed. Finally, Section 6.5 summarises the work and its implications for computational creativity research.

6.1 Creativity Dataset

The dataset analysed in this case study was originally collected as part of a research project designed to investigate the impact of intracranial stimulation on creative expression. Although the initial clinical objective was different, the resulting material, drawings paired with participant-provided titles, constitutes a rich resource for machine learning approaches to creativity assessment. For the purposes of this dissertation, the data were reused under appropriate anonymisation protocols, focusing specifically on multimodal prediction of creativity dimensions.

6.1.1 Participants

A total of 53 individuals took part in the study, contributing both before and after a controlled stimulation procedure. Participants represented a wide age span, from 10 to 60 years, with an almost balanced gender distribution. In addition to age and sex, the participant database contains metadata on:

- **Mother tongue**, predominantly Spanish,
- **Educational background**, ranging from compulsory secondary education to postgraduate studies,
- **Lifestyle variables**, including stimulant consumption (coffee, energy drinks), tobacco use, and number of hours slept before the experimental session.

These variables provide useful context for interpreting the dataset, even though the primary modelling tasks in this dissertation focused on the creative outputs themselves. The sociodemographic characteristics are summarised in the Table 6.1.

Table 6.1: Sociodemographic characteristics of participants of creativity dataset.

Characteristic	Value (N=53)
Gender (Female)	27 (50.9%)
Age, mean (SD)	29.7 (11.8)
Mother tongue	

Continued on next page

Table 6.1 (continued)

Characteristic	Value (N=53)
Spanish	50 (94.3%)
Portuguese	2 (3.8%)
Romanian	1(1.9%)
Education	
Secondary school	14 (26.4%)
High School	12 (22.6%)
Higher Vocational	9 (17.0%)
University Degree	7 (13.2%)
Medium Vocational	6 (11.3%)
Master	2 (3.8%)
Basic	1 (1.9%)
Doctorate	1 (1.9%)
Specialised Tracks	1 (1.9%)
Smoker (Yes)	27 (50.9%)
Stimulant consumption (Yes)	35 (66.0%)
Sleep (previous night), mean (SD)	7.25 (1.37)

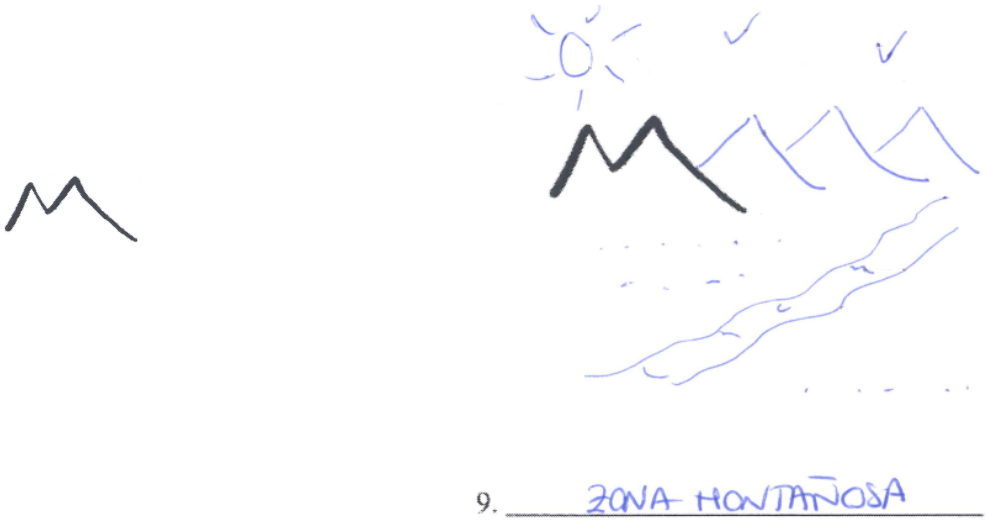
6.1.2 Data Collection Procedure

Each participant was asked to complete drawing tasks during two phases: (i) pre-stimulation and (ii) post-stimulation. For every drawing, participants were instructed to provide a title or short textual description that captured their interpretation or concept. Figure 6.1 shows the base drawing that the participants were provided as well as the drawing completed with its title. This resulted in a multimodal dataset combining visual (scanned images of drawings) and textual (titles in Spanish) information. In total, the dataset comprises 486 samples, accounting for both phases across all participants.

6.1.3 Dataset Content and Annotations

Each entry in the creativity dataset includes both the creative production itself and associated contextual information. Specifically, the dataset contains:

- The **drawing image**, collected in pre- and post-stimulation conditions and digitised for analysis,
- The corresponding **title text** (in Spanish),



(a) Baseline drawing **(b)** Completed drawing

Figure 6.1: Comparison of baseline (a) and completed (b) drawings — "ZONA MONTAÑOSA" (mountainous area).

- **Expert annotations** of four creativity-related dimensions,
- **Participant metadata**, including age, sex, mother tongue, educational level, stimulant/tobacco consumption, and sleep duration.

The expert annotations were performed by trained psychologists and provide the supervised ground truth for subsequent modelling. Four creativity-related target variables were assessed:

- **Originality (O):** binary label (0 = not original, 1 = original) evaluating the novelty of the drawing,
- **Flexibility (FLE):** categorical score capturing the thematic diversity represented across categories,
- **Elaboration (E):** ordinal score quantifying the level of detail and complexity in the drawing,
- **Title creativity (T):** binary label assessing whether the accompanying title was judged to be creative.

Table 6.2 presents a descriptive summary of the target variables, including their distributions and ranges. This overview provides an initial understanding of the structure of the annotated outcomes and highlights potential issues such as skewness, imbalance, or restricted variability. Identifying these characteristics at the

outset is essential to inform the subsequent methodological design and to incorporate appropriate strategies aimed at mitigating their impact on model development and evaluation.

Table 6.2: Summary of target variables.

Target	**Value (N=486)**
Creativity (yes)	198 (40.7%)
Title Creativity (yes)	277 (57.0%)
FLE (range)	1 – 63
E (range, mean (SD))	0 – 12, 2.07 (1.89)

The integrated structure enables different analytical configurations: unimodal approaches based on either image or text alone, as well as multimodal approaches that combine both modalities. In this way, the dataset offers a rich and ecologically valid basis for exploring how artificial intelligence can be used to assess creativity in human productions.

6.2 Problem Definition

This use case addresses the automatic assessment of creativity in human productions. From a psychological perspective, creativity is a complex construct that cannot be captured by a single definition, but is often studied through observable characteristics such as novelty, variety, and richness of ideas (Dow, 2022). Traditional assessments rely heavily on expert judgement, which is labour-intensive and subject to variability, limiting their scalability. From a machine learning (ML) perspective, the challenge lies in modelling these subjective evaluations while ensuring interpretability, validity, and applicability across experimental and educational settings.

6.2.1 Formal Statement

The predictive task is formally defined as supervised learning on multimodal data consisting of hand-drawn sketches and their accompanying textual titles. Expert annotations provide four distinct targets: (i) *Originality*, a binary indicator of whether the drawing presents novel ideas; (ii) *Flexibility*, a categorical measure of thematic variety across elements; (iii) *Elaboration*, an ordinal score quantifying detail and complexity; and (iv) *Title creativity*, a binary label of inventiveness in the title.

The objective is to design models that map the combined image–text input to these four outcomes. Beyond predictive accuracy, the models must offer interpretable insights into the features driving expert evaluations, thereby supporting a scalable and transparent alternative to purely manual creativity assessments.

6.2.2 Data-Generating and Temporal Structure

The dataset was collected as part of a structured experimental study in which participants were asked to produce drawings under controlled conditions. Each participant completed two drawing tasks: one before and one after a stimulation activity designed to enhance creative engagement. In both cases, participants were instructed to provide a title for their drawing. Sociodemographic and lifestyle data were also collected, including age, gender, education level, smoking and stimulant use, and sleep patterns.

Expert psychologists subsequently annotated the creative artefacts, assigning scores for originality, flexibility, elaboration, and title creativity. Thus, while data generation followed a synchronous and standardised procedure, the target labels were obtained retrospectively. Importantly, there is no longitudinal follow-up beyond this pre–post structure, making the temporal framework episodic rather than continuous.

6.2.3 Clinical and Operational Constraints

Although not a clinical dataset, model design and evaluation are subject to several operational constraints that parallel those found in applied predictive settings:

- **Ecological validity:** The data were generated in a controlled experimental environment. While this design ensures standardisation, it limits generalisability to more spontaneous forms of creative expression.
- **Expert dependency:** Ground-truth labels rely on subjective evaluations by psychologists. Even with careful protocols, inter-rater variability and differing theoretical perspectives inevitably introduce noise into the annotations.
- **Multimodal heterogeneity:** Input data include both visual artefacts and textual descriptions. Their integration requires deliberate modelling choices that balance modality-specific information with cross-modal synergies.
- **Sample size:** With under 500 annotated creative products, the dataset restricts the use of high-capacity models such as deep neural networks and necessitates careful regularisation to mitigate overfitting.
- **Language specificity:** Titles were predominantly produced in Spanish. Models may therefore not generalise directly to broader linguistic contexts without adaptation.

6.2.4 Specific Methodological Challenges

The prediction of creativity differs from clinical or behavioural prediction tasks in several ways:

- **Subjectivity of annotations:** Creativity does not have a universally agreed definition. Labels inevitably reflect theoretical biases of the annotators, which may propagate into trained models.
- **Balanced but diverse targets:** As shown in Table 6.2, the binary targets (originality and title creativity) are not severely imbalanced, but the continuous and categorical variables (elaboration and flexibility) span wide ranges. This diversity requires tailored modelling strategies.
- **Multimodal fusion:** Combining text and image inputs is non-trivial. Naïve fusion risks diluting signals, while overly complex strategies may overfit given the dataset's size.
- **External validity:** Models trained in this controlled setting may capture artefacts of the experimental design rather than general principles of creativity, limiting their generalisability.

6.2.5 Scope and Exclusions

The scope of this study is limited to supervised prediction of the four annotated creativity dimensions. Alternative approaches to creativity assessment, such as divergent thinking tests, self-reported questionnaires, or the generation of novel creative artefacts, fall outside its remit. No external datasets are incorporated, and transfer learning is not employed, in order to focus strictly on the interpretability of models trained on the collected dataset. Importantly, the models are developed exclusively for research purposes and are not intended for educational or clinical evaluation.

6.2.6 Success Criteria and Evaluation Plan

Success is defined by a dual focus on predictive performance and interpretability. For binary targets (originality and title creativity), evaluation will rely on ROC–AUC and F1-score, among other metrics. Flexibility, as a categorical variable, will be assessed using macro-F1, while elaboration will be evaluated using regression metrics such as mean squared error and correlation with expert scores.

Beyond accuracy, interpretability is considered essential. Models must reveal which features, either visual or textual, drive predictions in a way that aligns with expert

reasoning. Evaluation will follow a cross-validation strategy, stratified where relevant, to ensure robustness despite the dataset's limited size. Finally, both unimodal and multimodal pipelines will be compared to assess the added value of integrating different sources of information.

6.3 Solution Design

The methodological design for the creativity assessment use case was centred on the development of multimodal predictive pipelines capable of integrating visual and textual information. The dataset consisted exclusively of pairs of artefacts: drawings produced by participants and their corresponding textual titles. No clinical history, sociodemographic characteristics, or questionnaire data were used in the modelling stage. Instead, the entire design focused on extracting and combining meaningful representations from the images and titles themselves.

Two complementary approaches were explored. The first relied on neural network–based feature extraction and multimodal fusion. For each modality, embeddings were obtained separately before being combined or used independently depending on the configuration tested. For the image branch, pre-trained convolutional neural networks such as ResNet50, InceptionV3, and Xception were employed to generate high-level embeddings of the drawings. These models, trained on ImageNet, provided transferable representations of visual attributes such as structure, composition, and detail. Their convolutional backbones were used without the final classification head, and the resulting feature maps were flattened into one-dimensional embeddings that were subsequently processed through dense layers with non-linear activations and dropout regularisation to mitigate overfitting.

For the textual branch, pre-trained language models such as FastText and BETO (a Spanish-adapted BERT model) were used to transform each title into dense semantic embeddings. These representations captured lexical and contextual relationships between words and were refined through additional fully connected layers to model higher-order semantic abstractions. Depending on the configuration, three main modelling settings were implemented: (i) text-only models, in which the textual embedding branch alone was connected to the prediction layers; (ii) image-only models, using exclusively the visual embeddings; and (iii) multimodal models, where the last hidden layers of both branches were concatenated into a joint feature vector. This fused embedding was then passed through additional dense layers to learn cross-modal dependencies before producing the final output for each creativity dimension. Figure 6.2 illustrates this architecture.

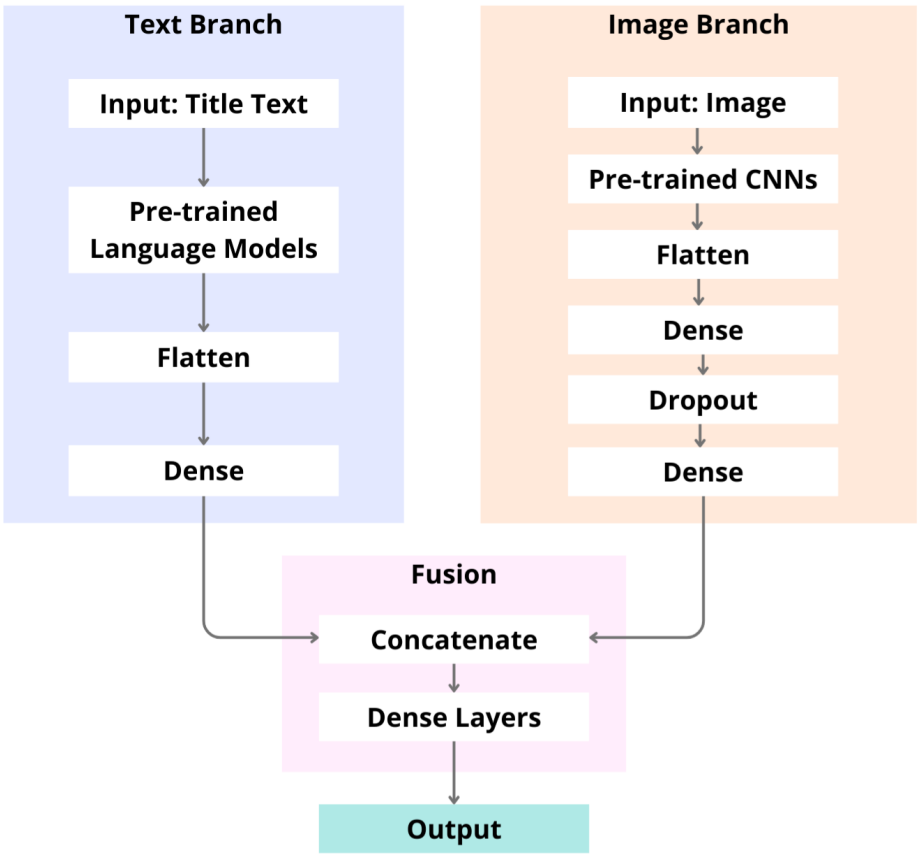


Figure 6.2: Multimodal architecture combining text and image branches through feature fusion for creativity prediction.

The second approach adopted a unified multimodal embedding strategy through the use of CLIP (Contrastive Language–Image Pre-training). CLIP jointly encodes drawings and titles into a shared multimodal latent space, where semantically aligned image–text pairs are positioned close together (see Figure 6.3). Each drawing–title pair was represented as a pair of 512-dimensional vectors derived from the CLIP image and text encoders (ViT-B/32). When both modalities were used, their embeddings were concatenated to form a single fused vector. These CLIP-based embeddings were then employed as input to a range of conventional machine learning classifiers and regressors (Figure 6.4), including logistic regression, naïve Bayes, support vector machines, random forests, and multilayer perceptrons. The same representation was used for both classification and regression tasks, depending on whether the creativity dimension was binary classification(originality and title creativity), multiclass classification (flexibility) or continuous (elaboration). This strategy leveraged CLIP's large-scale cross-modal alignment while maintaining simplicity and interpretability in the predictive layer.

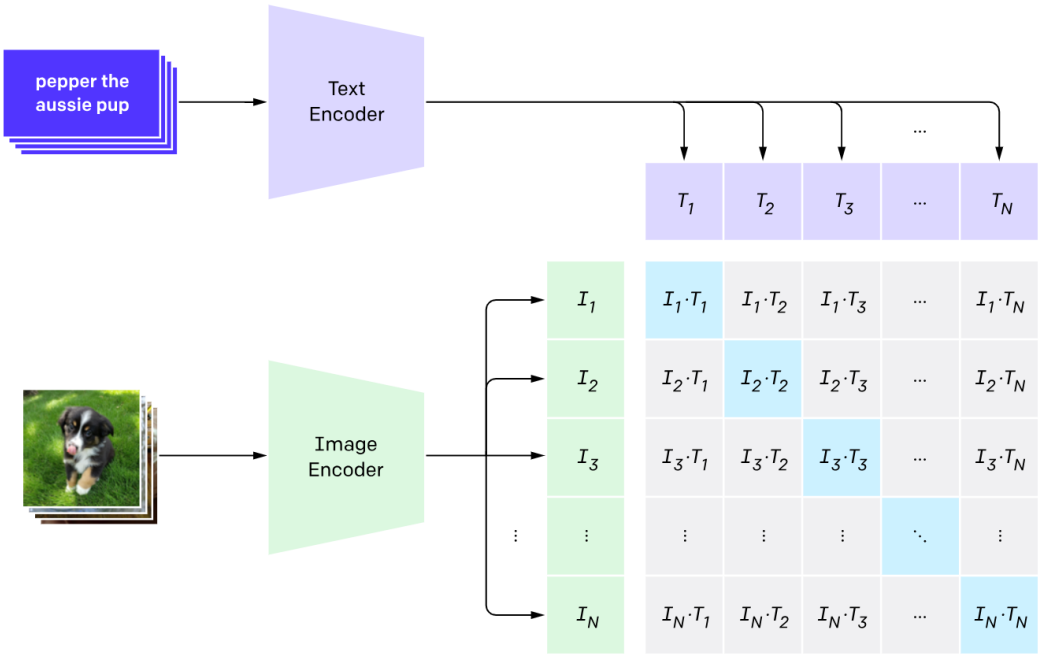


Figure 6.3: CLIP model (Radford et al., 2021).

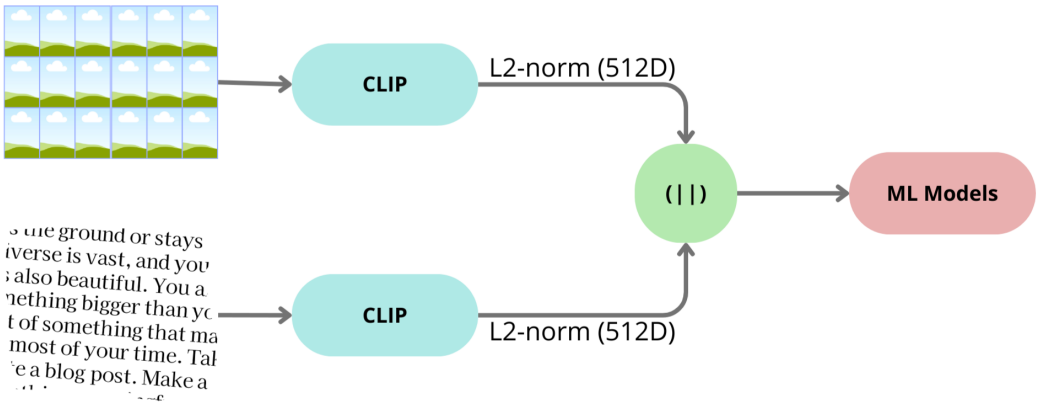


Figure 6.4: CLIP strategy for predicting creativity dimensions.

Together, these two approaches provided complementary perspectives on the problem. The neural network–based fusion architecture enabled fine-grained, dataset-specific integration of image and text features, allowing joint optimisation of the modality-specific embeddings. In contrast, the CLIP-based pipeline capitalised on extensive pre-training to generate robust, general-purpose multimodal representations suitable for lightweight classifiers. By systematically comparing the performance of these approaches across classification and regression tasks, the design sought to balance predictive accuracy with interpretability and to evaluate the re-

lative benefits of bespoke multimodal fusion versus general-purpose embedding strategies for the automatic assessment of creativity.

6.4 Explainability Framework

In contrast to the clinical use cases presented earlier in this dissertation, the goal of explainability in the creativity assessment task was not to identify the contribution of individual predictors, but rather to disentangle the relative value of different input modalities.

Although alternative explainability strategies specific to deep neural networks exist, such as gradient-based saliency maps, attention visualisation, or Grad-CAM for image-based models, the objective of this work was not to inspect intermediate activation patterns but to preserve a unified interpretability framework across model families. For this reason, model-agnostic techniques were prioritised. This decision ensured comparability between classical machine learning models and deep architectures, allowing interpretability to be embedded as a transversal design principle rather than as architecture-specific analysis.

Since the dataset consisted exclusively of visual artefacts (drawings) and textual artefacts (titles), explainability was centred on evaluating how text-only, image-only, and multimodal (text + image) models behaved when predicting each of the annotated creativity dimensions. This approach enabled the assessment of whether drawings, titles, or their integration carried the strongest signal for expert judgments of originality, flexibility, elaboration, and title creativity.

For each prediction task, three pipelines were trained and compared. The first relied exclusively on the textual titles, represented through pre-trained embeddings such as BETO and FastText. The second used only the drawings, represented through image embeddings obtained from models such as ResNet50 and InceptionV3. The third combined both sources in a multimodal framework, either through neural network fusion or through CLIP-based multimodal embeddings. By systematically evaluating performance across these three configurations, it was possible to determine the extent to which each modality contributed to the task and whether combining them improved prediction.

This comparative design served as the main interpretability tool. For example, the good performance of image-only models in predicting originality would suggest that visual cues dominate expert judgements about novelty, while improvements in predicting flexibility when including titles would indicate that textual explanations add critical information about the subject matter. Conversely, limited gains from multimodal integration would suggest redundancy between the two modalities, in-

sufficient complementarity, or the inclusion of noise for a given task. Thus, the analysis provided specific information about the relative contribution of drawings and titles.

The explainability framework therefore rests on cross-modality comparisons rather than feature-level attributions. By aligning differences in predictive performance with psychological theories of creativity, the approach allowed us to draw conclusions about the role of imagery and language in expert evaluation. This not only improved the transparency of the models but also offered a structured way of interpreting AI-based creativity assessment in terms that are meaningful for psychological research.

6.5 Summary and Conclusions

This chapter has presented the methodological framework developed to address the problem of automatically assessing human creativity from multimodal artefacts, specifically drawings accompanied by textual titles. The study was grounded in a curated dataset that combined image and text information with expert annotations of originality, flexibility, elaboration, and title creativity. The annotation process, carried out by psychologists using established creativity criteria, provided a reliable yet inherently subjective ground truth that reflects the interpretative nature of creativity assessment.

The work advanced the current state of the art by integrating multimodal machine learning techniques into a domain traditionally dominated by qualitative or manually coded evaluations. Existing approaches to computational creativity have typically relied on unimodal representations, most often visual features extracted from artworks or linguistic analysis of written content. In contrast, this study proposed a multimodal learning framework capable of jointly modelling visual and semantic cues, thereby providing a more comprehensive understanding of creative expression.

The methodological design addressed the key challenges of this domain, including the heterogeneity of image and text modalities, limited dataset size, and subjectivity in expert judgments. To this end, a dual modelling strategy was implemented. The first branch employed deep neural network architectures to extract modality-specific embeddings and fused them through a multimodal neural network to predict creativity dimensions. The second branch leveraged the CLIP (Contrastive Language–Image Pre-training) model to embed drawings and titles into a shared latent space, from which traditional machine learning algorithms were trained for classification and regression tasks. This design provided complementary perspectives, contrasting end-to-end multimodal deep learning with transfer learning using pre-

trained, semantically aligned representations.

Explainability was incorporated through modality-level interpretability rather than feature-level attribution, a necessary adaptation given the high-dimensional and abstract nature of the inputs. By systematically comparing text-only, image-only, and text+image configurations, the framework enabled transparent analysis of the relative contributions of visual and linguistic information to expert evaluations. This approach goes beyond most existing computational creativity models, offering an interpretable basis for examining the interplay between imagery and language in the perception of creative quality.

The proposed framework aligns with the overall objectives of this dissertation. It demonstrates the development of domain-relevant machine learning models (Objective 1). Explainability techniques were embedded to provide interpretable insights consistent with creativity theory (Objective 3), while collaboration with psychologists ensured the methodological and conceptual validity of the models (Objective 4).

The evaluation of this framework, presented in Chapter 7, will examine its predictive performance and interpretability across creativity dimensions. The methods described here correspond to the publication Analysing the impact of images and text for predicting human creativity through encoders (Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health, ICT4AWE 2025).

We don't see things as they are, we see them as we are.

ANAÏS NIN (1903 A.D. – 1977 A.D.)

CHAPTER

7

Evaluation

This chapter contributes to the overall dissertation by providing a comprehensive evaluation of the predictive systems developed across all use cases. It consolidates the methodological advances introduced in previous chapters and assesses their effectiveness, robustness, and interpretability across heterogeneous domains. Through this evaluation, the generalisability and practical relevance of the proposed framework are critically examined.

The chapter evaluates the predictive systems developed across four applied contexts that share the common objective of supporting decision making in human-centred settings. The evaluation addresses performance, calibration, robustness, and interpretability, and examines how these properties transfer across domains with heterogeneous data, outcomes, and operational constraints. The cases considered are the prediction of 30-day hospital readmission in patients with heart failure, prediction of 30-day hospital readmission in patients with multiple chronic conditions, prediction of eating disorder risk and one-year recovery, and multimodal assessment of creativity from hand-drawn sketches and titles.

Evaluation is treated as an iterative component of the modelling pipeline rather than a terminal step. Results reported here informed preprocessing choices, feature set curation, imbalance handling, and the selection and configuration of learning algorithms. In all cases, discrimination and calibration are reported alongside sensitivity analyses that test robustness to alternative imputations, sampling strategies, and decision thresholds. Where applicable, external validation is included to assess transportability and fairness. Interpretability is evaluated through feature attribution and model inspection to determine whether predictions align with clinical or

psychological plausibility and to support expert review.

The four use cases enable a cross domain view of the proposed methodology. The two readmission tasks provide complementary clinical settings. The heart failure cohort stresses integration of clinical indicators with patient reported measures at discharge. The multiple chronic conditions cohort incorporates functional status and social determinants, which tests the capacity of the models to capture risk in patients with complex needs. The eating disorder task focuses on prediction from psychometric instruments and therefore examines how well questionnaire based features support both classification of future risk and regression of self rated recovery. The creativity task is multimodal and evaluates whether text and image information contribute distinct and complementary signal. Together, these settings allow a unified analysis of model behaviour under missingness, imbalance, and domain shift, and of the trade off between accuracy and interpretability.

The remainder of the chapter is organised as follows. Section 7.1 presents the evaluation of the heart failure readmission models, including discrimination, calibration, and explainability analyses with patient reported outcomes. Section 7.2 evaluates the models for readmission in patients with multiple chronic conditions, with emphasis on the role of functional and social variables and on transportability across hospitals. Section 7.3 reports results for eating disorder risk and recovery, detailing classification and regression performance and the contribution of resilience and quality of life measures. Section 7.4 presents the evaluation of the multimodal creativity models and quantifies the complementarity of textual and visual inputs. Section 7.5 provides a cross case synthesis that compares methodological patterns and failure modes across domains, reports ablations, and discusses fairness and calibration. Section 7.6 summarises threats to validity and limitations. Section 7.7 concludes the chapter with the main findings that motivate the final conclusions and future work.

7.1 Prediction of 30-Day Readmission in HF Patients

Hospital readmission after acute decompensated heart failure (HF) remains a major clinical and organisational concern. Thirty-day readmission rates are widely used as indicators of care quality because they reflect post-discharge stability, continuity of follow-up, and adequacy of discharge planning. Their increase is linked to higher mortality, poorer quality of life, and elevated healthcare costs. In Spain, as in many other countries, the ageing population contributes to the growing prevalence of HF and, consequently, to the pressure on health systems.

Traditional methods such as logistic regression and Cox proportional hazards have long been applied to estimate readmission risk. However, their linear assumptions

limit the ability to model the complex interactions among clinical, functional, and psychosocial predictors. Consequently, their discriminative power is typically modest, with reported AUC values rarely exceeding 0.6 (Pikatza-Huerga et al., 2025a), which restricts their practical use for early identification of high-risk patients.

To address these limitations, this use case investigated machine learning (ML) as a flexible alternative for short-term readmission prediction. ML techniques can handle non-linearities, missing data, and class imbalance, while ensemble approaches enhance stability and accuracy. Additionally, SHAP-based explanations were used to ensure that model outputs remain interpretable and clinically meaningful.

The analysis was conducted on the multicentre ReIC cohort of patients discharged after HF hospitalisation, with 30-day readmission as the target outcome. Predictors included sociodemographic information, clinical history, functional measures, and patient-reported outcomes. The dataset was divided into training and test sets (80/20 split), and model parameters were optimised using grid search with internal cross-validation. Table 7.1 summarises the full evaluation protocol, including data partitioning, imbalance handling, baseline and ML models, ensemble design, metrics, and reproducibility settings. The evaluation focused on benchmarking traditional statistical baselines against a diverse set of ML classifiers and ensembles under consistent validation conditions.

Table 7.1: Evaluation protocol for 30-day readmission modelling in the ReIC cohort.

Aspect	Description
Prediction target	30-day hospital readmission (binary).
Data split	Random 80/20 train–test split with preserved class proportions; all model selection confined to the training set.
Hyperparameter search	Grid search on the training set with internal cross-validation; best configuration per algorithm carried forward to test evaluation.
Missing data	Comparison of: (i) no explicit imputation (models that tolerate missingness), (ii) simple imputations (zero, mode, mean) prior to training.
Class imbalance	Comparison of SMOTE, random undersampling, and cost-sensitive training; performance aggregated across repeated runs.
Baselines	Cox proportional hazards and logistic regression.

Continued on next page

Table 7.7 (continued)

Aspect	Description
ML models	Decision Tree, Random Forest, Gradient Boosting, LGBM, KNN, SVC, Gaussian Naïve Bayes, Neural Network.
Ensembles	Bagging-style ensembles with balanced resamples; hard and soft voting schemes evaluated; heterogeneous base learners (trees + Gaussian NB, with and without NNs).
Feature scaling	Standardisation applied..
Feature selection	SHAP-based global ranking on full models; clinician-guided reduction used to retrain and re-evaluate ensembles.
Metrics	Accuracy, ROC AUC, Sensitivity, Specificity, F1; mean ± SD reported over repetitions.
Calibration	Reliability curves and Brier score computed on the test set.
Reproducibility	Fixed random seed for dataset splitting, resampling, and model initialisation; identical evaluation pipeline across all configurations.

The following subsections present the results obtained for baseline and ML models and discuss their relative strengths in terms of discrimination, calibration, and interpretability. The section concludes with a synthesis of the findings and their implications for clinical decision support and healthcare resource optimisation.

7.1.1 Handling Missing Values

The first analysis assessed model behaviour when training directly on the incomplete dataset. This setting exposes how learners react to systematic missingness in an imbalanced cohort, where non-readmissions dominate. Table 7.2 summarises the results for the three baseline classifiers considered in this configuration.

The three algorithms converge to near-trivial solutions. Specificity is very high and sensitivity collapses towards zero, which means that models almost always predict non-readmission. Reported accuracy is inflated by the class imbalance and is not informative about minority class detection. ROC AUC remains close to 0.5, indicating a lack of discriminative capacity. These findings show that unaddressed missingness compounds the imbalance problem and leads to degenerate decision rules dominated by the majority class.

Table 7.2: Performance without explicit handling of missing values.

Model	Accuracy	ROC AUC	Sensitivity	Specificity	F1
Decision Tree	0.73	0.49	0.15	0.83	0.15
Gradient Boosting	0.83	0.49	0.00	0.99	0.00
LGBM	0.83	0.51	0.03	0.98	0.05

The next step evaluated three simple imputation strategies before training: filling missing entries with zero, with the mode, and with the mean. Each strategy was applied across a diverse set of classifiers. Tables 7.3 report the results in terms of sensitivity (Sens.), specificity (Spec.) and F1-score (F1) to see if it makes any difference in solving the imbalance problem.

Table 7.3: Performance of classifiers under different imputation strategies.

	Zero imputation			**Mode imputation**			**Mean imputation**		
Model	Sens.	Spec.	F1	Sens.	Spec.	F1	Sens.	Spec.	F1
Decision Tree	0.15	0.89	0.17	0.03	0.96	0.05	0.00	0.98	0.00
Gradient Boosting	0.00	0.98	0.00	0.03	0.98	0.05	0.00	1.00	0.00
LGBM	0.03	0.99	0.06	0.00	0.98	0.00	0.00	1.00	0.00
Logistic Regression	0.00	0.98	0.00	0.00	0.99	0.00	0.03	1.00	0.06
KNN	0.00	0.99	0.00	0.03	0.97	0.05	0.03	0.97	0.05
Random Forest	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
Naïve Bayes	**0.67**	0.39	0.27	**0.73**	0.35	0.28	**0.21**	0.78	0.18
SVC	0.30	0.87	0.30	0.00	1.00	0.00	0.00	0.99	0.00
Neural Network	0.27	0.89	0.27	0.09	0.96	0.14	0.00	1.00	0.00

The three strategies produce similar overall patterns. Most classifiers converge towards trivial solutions dominated by the majority class: specificity remains extremely high (often above 0.95) while sensitivity collapses close to zero, resulting in poor identification of readmitted patients. Only Naïve Bayes consistently avoids this collapse, reaching sensitivities between 0.21 and 0.73 across imputations, although this comes at the cost of markedly reduced specificity.

Zero imputation appears slightly more favourable for decision trees, SVC, and neural networks, which recover moderate sensitivity values while maintaining acceptable specificity. Mode and mean imputation, in contrast, push almost all classifiers towards the degenerate solution of predicting non-readmission for nearly all cases. These results indicate that simple imputation does not solve the fundamental im-

balance problem. While some strategies can preserve partial minority-class signal, the overall discriminative ability remains insufficient for clinically meaningful application, highlighting the necessity of combining imputation with more advanced techniques such as resampling, cost-sensitive learning, or ensemble modelling.

7.1.2 Class Imbalance Handling

As the ReIC dataset exhibits a strong imbalance between readmitted and non-readmitted patients (16% vs. 84%), three balancing techniques were applied in combination with the models: (i) Synthetic Minority Oversampling Technique (SMOTE), (ii) random undersampling, and (iii) cost-sensitive learning based on class weights. Performance for each configuration was aggregated across repetitions, reporting mean values and standard deviations (Table 7.4).

The results with SMOTE show that several classifiers converged towards trivial solutions. Both gradient boosting and random forest obtained specificity values close to 1 but sensitivities and F1 scores near zero, reflecting their systematic tendency to predict non-readmission. Logistic regression achieved moderate sensitivity (0.43 ± 0.21) but low F1, while KNN and Naïve Bayes displayed some balance between recall and specificity, though with modest discriminative capacity overall. These findings indicate that simple oversampling of the minority class was insufficient to restore meaningful recall without overfitting to duplicated samples.

In contrast, random undersampling substantially increased sensitivity across almost all classifiers. Decision tree, gradient boosting, logistic regression, and random forest models reached sensitivities close to 0.45–0.50, with F1 values in the range of 0.45–0.49. The neural network obtained the highest sensitivity (0.57 ± 0.16), even though with reduced specificity. Naïve Bayes also benefited, achieving balanced performance with F1 around 0.43. Importantly, undersampling produced more consistent improvements across classifiers, with lower variance compared with the other techniques.

The results for cost-sensitive training were heterogeneous. While the decision tree and neural network achieved high sensitivities (up to 0.66 ± 0.24), these gains were offset by marked reductions in specificity, often below 0.50. Other models, such as random forest, remained degenerate with F1 values close to zero. The variability observed across repetitions was also considerably higher than for undersampling, suggesting instability in the optimisation of class-weighted losses.

Overall, these results demonstrate that random undersampling was the most consistent strategy for recovering discriminative capacity, achieving a better trade-off between sensitivity and specificity across classifiers. For this reason, undersampling

Table 7.4: Summary of model performance under different imbalance handling strategies. Values are reported as mean ± standard deviation.

	SMOTE		
Model	Sens.	Spec.	F1
Decision Tree	0.14 ± 0.02	0.77 ± 0.02	0.12 ± 0.03
Gradient Boosting	0.00 ± 0.02	0.99 ± 0.02	0.00 ± 0.04
LGBM	0.00 ± 0.04	0.98 ± 0.02	0.00 ± 0.07
Logistic Regression	0.43 ± 0.21	0.54 ± 0.02	0.22 ± 0.08
KNN	0.35 ± 0.02	0.64 ± 0.02	0.21 ± 0.00
Random Forest	0.00 ± 0.00	0.99 ± 0.02	0.00 ± 0.00
SVC	0.24 ± 0.00	0.75 ± 0.02	0.19 ± 0.00
Neural Network	0.18 ± 0.00	0.89 ± 0.02	0.20 ± 0.01
Naïve Bayes	0.26 ± 0.02	0.81 ± 0.02	0.23 ± 0.05
	Undersampling		
Model	Sens.	Spec.	F1
Decision Tree	0.44 ± 0.06	0.54 ± 0.07	0.46 ± 0.05
Gradient Boosting	0.45 ± 0.09	0.53 ± 0.05	0.46 ± 0.08
LGBM	0.44 ± 0.12	0.51 ± 0.07	0.45 ± 0.11
Logistic Regression	0.44 ± 0.06	0.48 ± 0.07	0.45 ± 0.06
KNN	0.37 ± 0.14	0.49 ± 0.06	0.39 ± 0.10
Random Forest	0.49 ± 0.05	0.50 ± 0.11	0.49 ± 0.06
SVC	0.27 ± 0.23	0.52 ± 0.21	0.31 ± 0.15
Neural Network	0.57 ± 0.16	0.31 ± 0.25	0.52 ± 0.06
Naïve Bayes	0.41 ± 0.08	0.51 ± 0.06	0.43 ± 0.06
	Cost-Sensitive Training		
Model	Sens.	Spec.	F1
Decision Tree	0,59 ± 0.17	0,44 ± 0.04	0,26 ± 0.05
LGBM	0,13 ± 0.39	0,79 ± 0.22	0,13 ± 0.20
Logistic Regression	0,46 ± 0.15	0,55 ± 0.03	0,24 ± 0.06
Random Forest	0,00 ± 0.13	0,90 ± 0.13	0,00 ± 0.12
Neural Network	0,66 ± 0.24	0,33 ± 0.39	0,30 ± 0.05
Naïve Bayes	0,17 ± 0.02	0,90 ± 0.01	0,19 ± 0.02

was selected as the basis for subsequent experiments. To further enhance robustness and reduce the variability inherent to undersampling, a bagging ensemble strategy was adopted, allowing the integration of multiple undersampled models into a stable and more generalisable predictor.

7.1.3 Bagging Ensemble Strategies

To further enhance predictive stability after undersampling, four bagging ensemble configurations were evaluated: **C1** (3 decision trees + 3 Gaussian Naïve Bayes), **C2** (5 decision trees + 5 Gaussian Naïve Bayes), **C3** (10 decision trees + 10 Gaussian Naïve Bayes), and **C4** (5 decision trees + 5 Gaussian Naïve Bayes + 5 neural networks). These ensembles were designed to exploit the complementary strengths of trees, probabilistic models, and neural networks while mitigating variance through resampling.

Table 7.5 reports the performance of ensemble models compared with the statistical baselines. Cox regression and logistic regression confirmed their limited predictive utility, with ROC AUC values of 0.58 ± 0.06 and 0.50 ± 0.05, respectively, and F1 scores of only 0.27 and 0.18. Sensitivity for both methods remained below 0.50, underlining their inability to reliably detect patients who were readmitted within 30 days.

The previous section demonstrated that random undersampling improved sensitivity for several single classifiers. Decision trees, logistic regression, and random forest models reached sensitivities in the 0.44 - 0.49 range, with F1 values around 0.45 - 0.49. Although these represented improvements compared with unbalanced models, specificity tended to decline, and performance remained inconsistent across techniques.

By contrast, the bagging ensembles that combined decision trees with Gaussian Naïve Bayes yielded marked and consistent gains. The configuration with three classifiers of each type achieved a ROC AUC of 0.73 ± 0.05, sensitivity of 0.90 ± 0.08, and an F1 score of 0.76 ± 0.04, significantly outperforming not only Cox and logistic regression but also the best results obtained with undersampling alone. Increasing the number of base learners to five of each type further improved ROC AUC to 0.74 ± 0.06 and F1 to 0.77 ± 0.04, maintaining high sensitivity (0.90 ± 0.08) with only a minor change in specificity.

Overall, bagging ensembles delivered a step-change in performance. Compared with Cox and logistic regression, the ensembles improved ROC AUC by more than 0.15 and doubled the F1 score. Relative to the best undersampling results, the ensembles enhanced sensitivity by more than 0.40 while maintaining a higher

Table 7.5: Performance of ensemble models and baselines. Data are presented as mean (SD)

	Accuracy	ROC AUC	Sensitivity	Specificity	F1 Score
CoX	0.67 ± 0.03	0.58 ± 0.06	0.41 ± 0.09	0.72 ± 0.03	0.27 ± 0.06
Log. Reg.	0.57 ± 0.03	0.50 ± 0.05	0.31 ± 0.08	0.62 ± 0.04	0.18 ± 0.05
Hard Voting					
C1	0.72 ± 0.05	0.73 ± 0.05	0.90 ± 0.08	0.56 ± 0.13	0.76 ± 0.04
C2	0.74 ± 0.06	0.74 ± 0.06	0.90 ± 0.08	0.58 ± 0.14	0.77 ± 0.04
C3	0.79 ± 0.04	0.79 ± 0.04	0.91 ± 0.06	0.67 ± 0.10	0.81 ± 0.03
C4	0.79 ± 0.05	0.80 ± 0.05	0.93 ± 0.03	0.66 ± 0.10	0.82 ± 0.03
Soft Voting					
C1	0.76 ± 0.05	0.76 ± 0.05	0.84 ± 0.10	0.68 ± 0.11	0.77 ± 0.05
C2	0.77 ± 0.05	0.77 ± 0.05	0.85 ± 0.11	0.70 ± 0.11	0.78 ± 0.05
C3	0.79 ± 0.04	0.79 ± 0.04	0.89 ± 0.07	0.70 ± 0.10	0.81 ± 0.03
C4	**0.81 ± 0.04**	**0.81 ± 0.04**	**0.93 ± 0.03**	**0.69 ± 0.09**	**0.83 ± 0.03**

F1, demonstrating their ability to capture minority-class signal without collapsing into trivial solutions. Although specificity decreased moderately, this trade-off was deemed acceptable in the clinical context, where missing true readmissions carries greater risks than false positives. For these reasons, bagging ensembles were selected as the definitive modelling approach for the ReIC use case.

7.1.4 Feature Selection Guided by SHAP and Clinicians

Explainability was incorporated into the modelling process to ensure that the predictors retained clinical plausibility and practical applicability. Feature importance was first quantified using SHAP values on the best-performing bagging ensemble (see Figure 7.1. This analysis identified frailty indicators, psychosocial dimensions, and patient-reported outcomes among the strongest contributors to the prediction of 30-day readmission. In particular, high levels of anxiety and depression, reduced functional capacity, and elevated frailty scores consistently emerged as key risk factors. Clinical variables such as previous hospitalisations and comorbidity burden also had significant impact, but were complemented by the psychosocial and functional dimensions in determining readmission risk.

While the full SHAP ranking highlighted a large set of candidate predictors, direct input from clinicians was sought to improve interpretability and facilitate potential

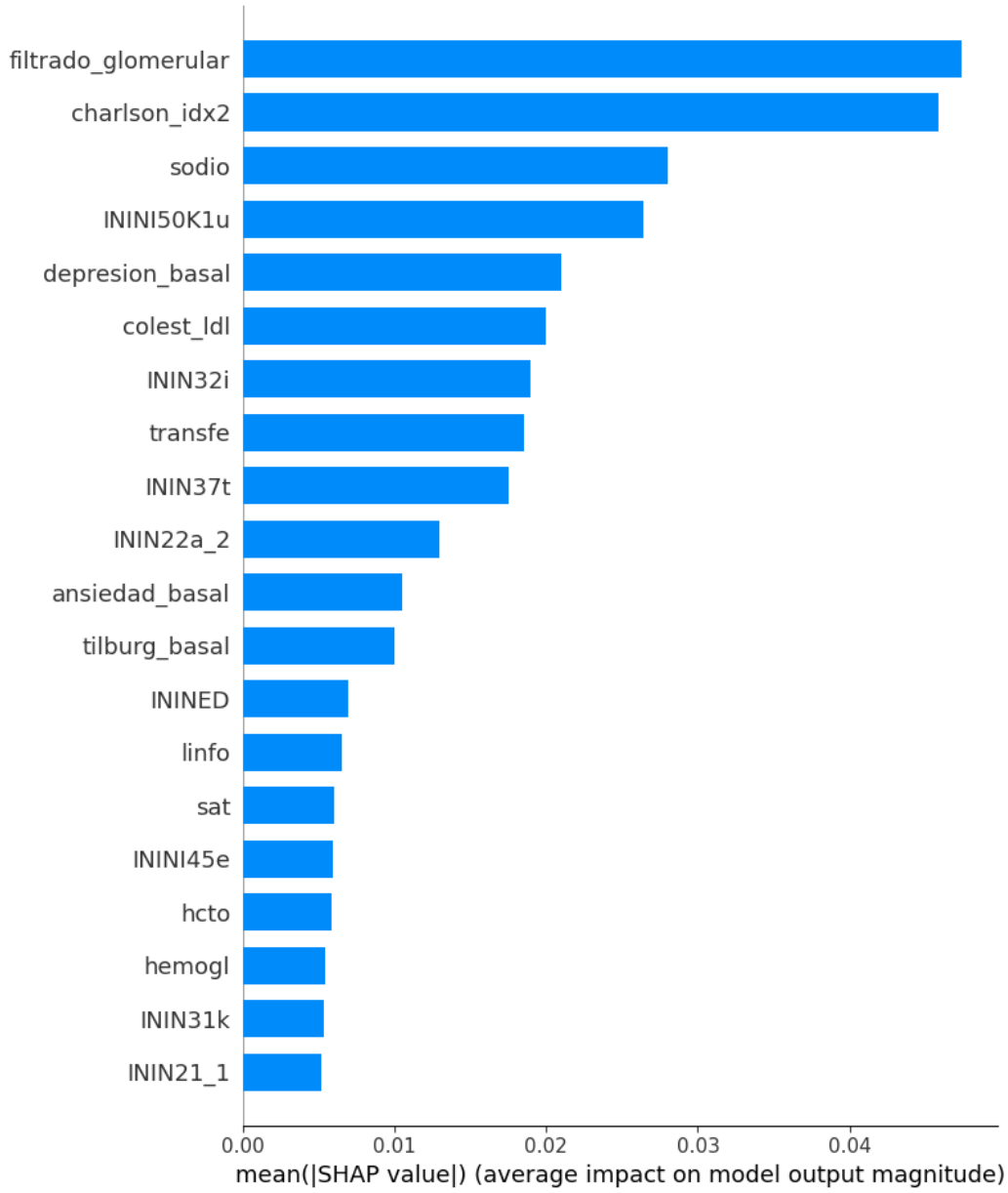


Figure 7.1: The 20 most important features of the best model to date. Go to Appendix A for the complete list of 50 variables that clinicians reviewed and refined.

integration into routine practice. A subset of features was therefore selected based on both predictive contribution and clinical alignment. This process prioritised variables that are routinely collected, easy to measure, and considered actionable in discharge planning. The final clinician-guided feature set included frailty, depression, anxiety, functional status, and number of previous admissions, as these factors jointly captured the multidimensional vulnerability associated with early

readmission while remaining feasible for real-world application. Table 7.6 lists the characteristics that were ultimately chosen, their meaning, and how they are measured.

Table 7.6: Final feature set selected based on SHAP importance and clinician guidance.

Variable	Description
filtrado_glomerular	Glomerular filtration rate (mL/min/1.73m^2)
sodio	Serum sodium concentration (mmol/L)
ININI50K1u	Presence of oedema up to the knee, assessed clinically
depresion_basal	HAD depression subscale score at admission (0–21)
colest_ldl	Low-density lipoprotein cholesterol (mg/dL)
transfe	Transferrin concentration (mg/dL)
ININ37t	Use of angiotensin receptor antagonists measured in the emergency department (binary)
ansiedad_basal	HAD anxiety subscale score at admission (0–21)
tilburg_basal	Tilburg Frailty Indicator score
ININED	Age in years
linfo	Lymphocyte count (10^9/L)
hemogl	Haemoglobin concentration (g/dL)
ININ21_1	Last troponin value before admission (ng/L)
vitd	Serum vitamin D level (ng/mL)
ININ1e_0	Altered acid-base type at admission (categorical)
ININI50sp	Systolic blood pressure at discharge (mmHg)
ININ22c_0	Baseline tricuspid regurgitation rate (echocardiography)
ININ17	Body mass index (kg/m^2)
mlhfq_basal	Minnesota Living with Heart Failure Questionnaire score at baseline
ININ29d_2	History of mitral stenosis (binary)
ININI50cp	Paroxysmal nocturnal dyspnoea at discharge (binary)
ININ20	NT-proBNP test conducted (Yes/No)
ININI52q	Use of calcium antagonists at discharge (binary)
ININ27c	Long-term treatment with beta-blockers at baseline (binary)
ININ30b_2	History of drug-eluting stent (binary)

Continued on next page

Table 7.6 (continued)

Variable	Description
dimerod	D-dimer test result (ng/mL)
ININ20_1	Baseline NT-proBNP level (pg/mL)
ININI45b	Pulmonary disease decompensation during admission (binary)
ININ25_0	Number of emergency department visits for HF in the last 2 years
hba1	Haemoglobin A1c (%)
colest_hdl	High-density lipoprotein cholesterol (mg/dL)
ININI500e_1	Bradyarrhythmia identified as a cause of heart failure (binary)

Retraining the bagging ensemble with the clinician-guided feature subset confirmed the robustness of the approach. As shown in Figure 7.2, all ensemble configurations clearly outperformed Cox regression (blue line) and logistic regression (orange line). The statistical baselines remained close to random prediction, with AUCs of 0.58 and 0.50, sensitivities below 0.50, and F1 scores below 0.30. In contrast, every ensemble variant achieved substantially higher discrimination, highlighting the benefit of combining undersampling with diverse learners.

The best-performing configuration, trained with the reduced feature set, reached a ROC AUC of **0.71**, compared with 0.74 for the full-feature model. Sensitivity remained high at **0.88**, and the F1 score exceeded **0.70**. These values represent only a modest decrease relative to the full model while retaining a marked advantage over traditional statistical approaches and even the best single-classifier results from the imbalance-handling experiments.

Calibration analysis further supported the reliability of the reduced model. The calibration plot (Figure 7.3) shows that predicted probabilities closely matched observed outcomes, with only minor deviations from the ideal diagonal. The Brier score of 0.24 confirmed good overall calibration, which is essential for clinical implementation where probability estimates inform post-discharge planning.

Figure 7.4 displays global feature importance and directional effects. Each point represents a subject, the horizontal axis encodes the SHAP contribution to the log-odds of readmission, and colour encodes the feature value within that subject (blue low, pink high). Rightward displacement indicates higher predicted risk, leftward displacement indicates lower predicted risk.

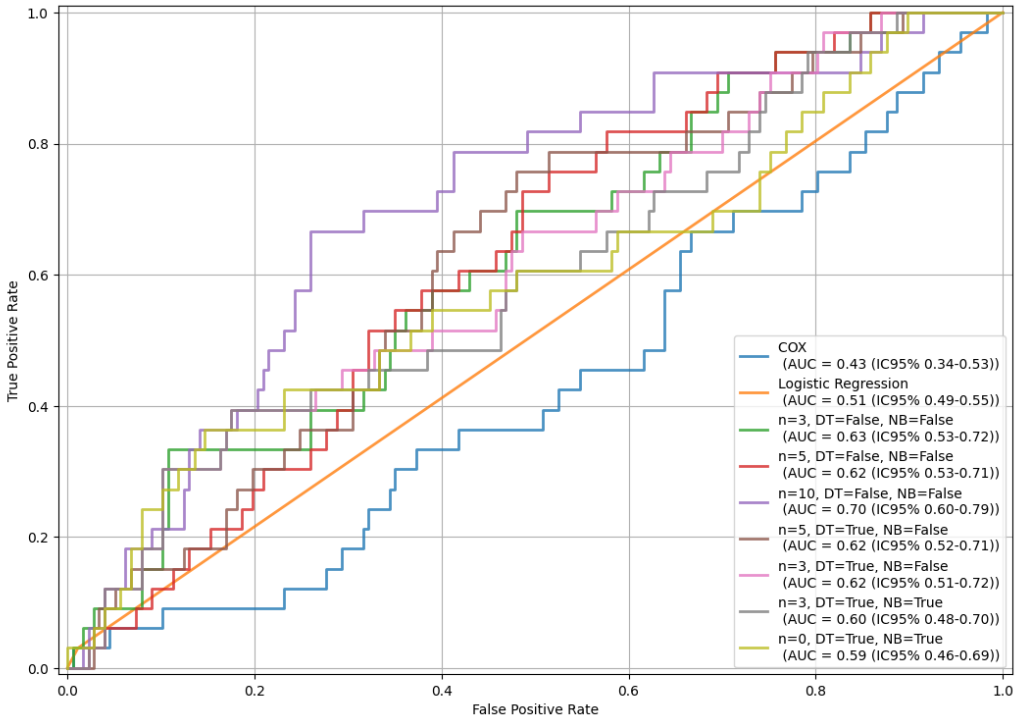


Figure 7.2: ROC curves of models trained with the data subset compared with CoX and Logistic Regression.

Renal function and biomarkers.

- **filtrado_glomerular** sits at the top of the ranking. Blue points are concentrated on the right while pink points cluster on the left, indicating that *lower* glomerular filtration rate increases readmission risk and *higher* values are protective. The wide horizontal footprint shows a sizeable marginal effect.

- **sodio** shows the same pattern, with low sodium values (blue) pushing risk upwards. This is consistent with the known association between hyponatraemia and adverse HF outcomes (Zhao et al., 2023).

- **hemogl** displays predominantly right-shifted blue points, indicating that *lower* haemoglobin increases risk, which aligns with anaemia burden in HF (Rahman et al., 2021).

- **linfo** has blue points on the right and pink on the left, suggesting that *lower* lymphocyte counts increase risk. This likely captures inflammatory and nutritional status.

- **transfe** manifests mixed behaviour. Both low and high values can contribute to risk, with a wider dispersion for higher values, which suggests non-linearity

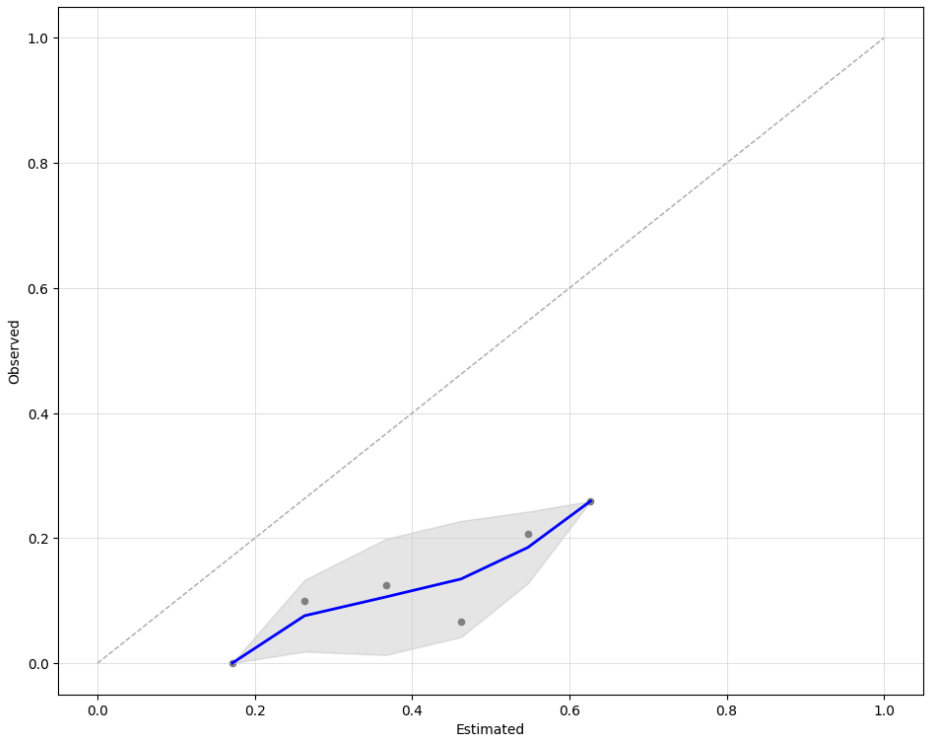


Figure 7.3: Calibration of the final model for ReIC.

consistent with iron metabolism where deficiency and inflammation can lead to different transferrin responses.

- **vitd** shows blue-to-right displacement, indicating that *lower* vitamin D levels increase risk.
- **hba1** shows a tendency for higher values to shift right, consistent with worse glycaemic control increasing risk (Yang et al., 2022).
- **dimerod** is dominated by pink points on the right, indicating that *higher* D-dimer levels increase risk, plausibly capturing pro-thrombotic or inflammatory activity.
- **ININ21_1** and **ININ20_1** (troponin and baseline NT-proBNP) exhibit rightward shifts for high values, reflecting myocardial injury and congestion burden as risk drivers.

Psychosocial and patient-reported outcomes.

- **depresion_basal** and **ansiedad_basal** show clear rightward shifts for higher scores, meaning that worse depressive or anxiety symptoms increase predicted risk. The vertical spread indicates heterogeneous effects across clinical strata,

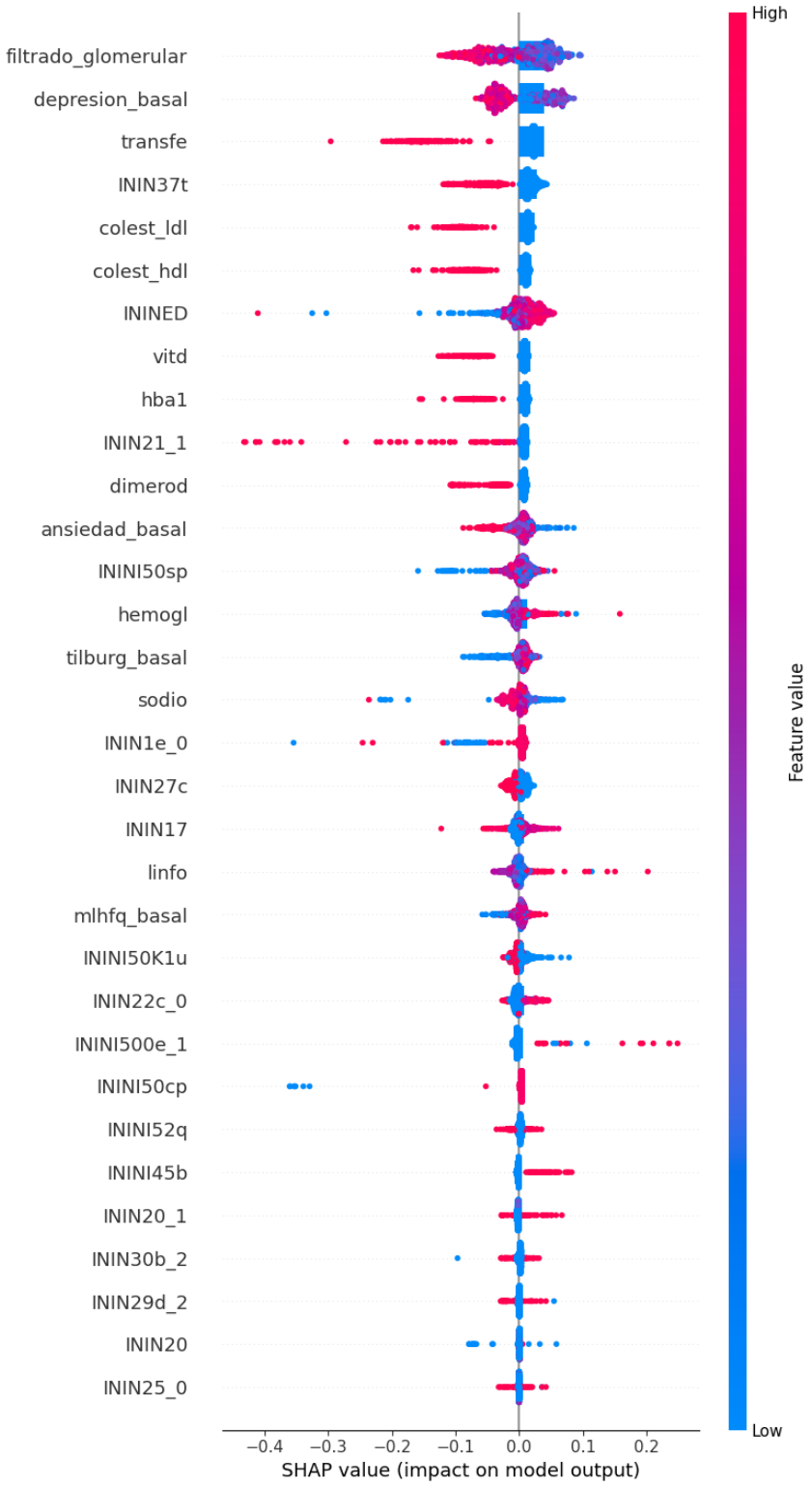


Figure 7.4: SHAP values of the final model for ReIC.

which is expected for psychosocial factors moderated by disease severity and social support.

- **tilburg_basal** and **mlhfq_basal** demonstrate that greater frailty and poorer self-reported quality of life shift SHAP values to the right. The effect is monotonic and clinically coherent, reinforcing the utility of patient-reported measures for short-term readmission risk.

Haemodynamics, signs, and vitals.

- **ININI50sp** (systolic blood pressure at discharge) presents blue-to-right behaviour, suggesting that lower discharge SBP increases risk, consistent with haemodynamic fragility after decompensation (Krzesiński et al., 2021).
- **ININI50K1u** (oedema up to the knee) and **ININI50cp** (paroxysmal nocturnal dyspnoea at discharge) show right-skewed positive SHAP values when present, representing residual congestion.
- **ININ1e_0** (altered acid-base at admission) pushes risk to the right for abnormal categories, consistent with metabolic or respiratory compromise.

Cardiac structure and history.

- **ININ22c_0** (tricuspid regurgitation), **ININ29d_2** (mitral stenosis history), and **ININ30b_2** (drug-eluting stent history) have positive right-tail contributions when present, marking structural or procedural history linked to recurrent instability.
- **ININED** (age) shows a predominantly right-shifted pattern for higher values but with visible spread, suggesting interactions with frailty and comorbidity.
- **ININ17** (BMI) suggests a non-linear association. Low BMI shows right-shifted blue points indicating higher risk, compatible with the cachexia phenotype (Valentova et al., 2022). Extremely high BMI does not systematically increase risk in this cohort, which fits the obesity paradox described in HF populations (Hobbach et al., 2024).

Treatments and care processes.

- **ININ27c** (long-term beta-blocker use at baseline) tends to show left-shifted pink points, indicating a protective association when treatment is present, after adjustment by the model for other covariates.
- **ININ52q** (calcium antagonists at discharge) and **ININ37t** (angiotensin receptor antagonists recorded in emergency department) display positive right-tail effects in a subset of patients. These likely reflect confounding by indication and residual risk in those discharged on more complex regimens rather than a causal effect of the drugs themselves.

- **ININ20** (NT-proBNP test performed) shows small positive contributions when present. This is plausibly a proxy for clinical suspicion of congestion and therefore correlates with risk.

Healthcare use.

- **ININ25_0** (emergency visits for HF in the previous two years) displays a right-skewed contribution as counts increase, capturing prior instability as a strong predictor of early readmission.

Three general patterns emerge from the plot. First, several predictors exhibit *directional monotonicity* with clear clinical meaning, for example lower renal function, lower sodium, and lower haemoglobin consistently increase risk. Second, a subset presents *non-linear* or *interaction-rich* effects, seen as vertical colour mixing and bilateral horizontal spread, for example transferrin, age, and BMI. Third, some care process variables such as testing flags and discharge medications behave as *proxies* of clinical concern rather than primary biological drivers, which the model exploits to refine individual risk. Together these patterns demonstrate that the reduced model concentrates signal on clinically interpretable pathways of congestion, frailty, inflammation, and prior instability.

In conclusion, despite a substantial reduction in the number of predictors, the ensemble retained most of the discriminative performance of the full feature configuration, improved calibration remained evident, and the SHAP analysis aligned with pathophysiological expectations. The reduction in dimensionality therefore increased applicability and transparency without a commensurate loss in predictive accuracy.

7.2 Prediction of 30-Day Readmission in MCC Patients

Building on the insights obtained from the prediction of readmissions in heart failure, the modelling strategy for patients with multiple chronic conditions (MCC) was designed to be more comprehensive from the outset. Instead of focusing sequentially on individual methodological challenges, the full pipeline was applied directly, encompassing missing data handling, dimensionality reduction, class imbalance correction, and feature subset evaluation. This decision reflected both the complexity of MCC populations, where clinical, social, and functional dimensions interact, and the prior lessons that underscored the need for systematic exploration of preprocessing strategies.

The analysis combined internal and external validation to evaluate both model performance and generalisability. A comprehensive comparison was conducted across

feature representations, imputation methods, balancing strategies, and learning algorithms, complemented by a clinical benchmark to contextualise results.

Table 7.7 summarises the evaluation design, including dataset partitioning, model optimisation, and validation settings. It also outlines the performance metrics, calibration procedures, and reproducibility controls adopted across all experiments. Together, these elements provided a rigorous and standardised framework to assess predictive performance, interpretability, and robustness in the MCC cohort.

Table 7.7: Evaluation protocol for 30-day readmission modelling in the RePluris cohort.

Aspect	Description
Prediction target	30-day hospital readmission in patients with multiple chronic conditions.
Data split	Random 80/20 train–test split with preserved class proportions. Model selection and preprocessing choices decided on the training set.
External validation	Independent cohort of 173 patients from a participating hospital used for out-of-sample validation of the final selected model.
Hyperparameter search	Grid search on the training set with internal cross-validation. Best configuration per algorithm evaluated on the held-out test set and then on the external cohort where applicable.
Feature sets	Complete feature space. Corrected feature set with reduced multicollinearity. TOP20 selected with SHAP. PROFUND subset with and without the Alcoholism variable.
Dimensionality reduction	Principal Component Analysis and SelectKBest applied where indicated, with the number of components or features chosen within the training-set search.
Missing data	KNN Imputer and Iterative Imputer compared. Imputation fitted on the training split only and applied to the test and external data.

Continued on next page

Table 7.7 (continued)

Aspect	Description
Class imbalance	Oversampling with SMOTE, Borderline-SMOTE, and ADASYN. Undersampling with Cluster Centroids, NearMiss, and Random Undersampling. Hybrid methods SMOTETomek and SMOTEENN. Results aggregated across repeated runs.
Baselines	Cox regression as statistical baseline. LACE index as a clinical benchmark for external comparison.
ML models	Decision Tree, Random Forest, XGBoost, Support Vector Classifier, KNN, Multi-layer Perceptron.
Ensembles	Bagging and heterogeneous ensembles explored during development. Final external validation performed with the selected single-model configuration.
Feature scaling	Standardisation applied.
Metrics	Accuracy, ROC AUC, Sensitivity, Specificity, Precision, F1. Precision–recall curves for threshold analysis.
Calibration	Reliability curves and Brier score on the test set and on the external cohort for the final model.
Reproducibility	Fixed random seeds for splitting, resampling, and model initialisation. Identical evaluation pipeline across all configurations.

The following subsections report the results of this systematic experimentation. First, the best-performing combinations of imputation, reduction, and balancing strategies are presented for each feature set. Next, the models selected for further interpretation are compared in terms of discrimination, calibration, and external validation. Finally, the contribution of individual features is analysed through SHAP values, highlighting the clinical and social dimensions that most strongly influence the risk of readmission in patients with MCC.

7.2.1 Results in readmissions within 30 days in MCC

The systematic exploration of models for predicting readmissions in MCC revealed distinct performance patterns across feature sets and preprocessing strategies. Table 7.8 provides the best combinations of model and preprocessing for each feature set, while Table 7.9 shows their results in terms of metrics. In the complete dataset, configurations that combined KNN imputation, dimensionality reduction with PCA, and oversampling methods such as ADASYN achieved very high sensitivity, successfully identifying most readmissions, but this came at the expense of

precision. By contrast, the corrected dataset, where highly correlated variables were grouped to reduce redundancy, yielded the highest overall accuracy and specificity with a decision tree, although sensitivity remained low and many readmissions were missed.

Table 7.8: Best-performing configurations for prediction of 30-day readmission in MCC.

Feature set	Model	Imputation	Balancing
Complete + PCA	KNN	KNN Imputer	ADASYN
Corrected	Decision Tree	KNN Imputer	–
TOP20 (SHAP)	SVC	Iterative Imputer	Undersampling
PROFUND + Alcoholism	Decision Tree	Iterative Imputer	SMOTE
PROFUND	XGBoost	–	Undersampling

Notes. Corrected: complete set with additional preprocessing to reduce multicollinearity. *TOP20 (SHAP)*: twenty most important predictors on the complete model.

Table 7.9: Performance metrics for the best configurations in MCC (Part I). Values are mean (95% CI).

Feature set	Accuracy	Precision	AUC
Complete	0.48 (0.39–0.58)	0.21 (0.13–0.29)	0.67 (0.58–0.76)
Corrected	**0.82 (0.75–0.90)**	**0.38 (0.29–0.48)**	0.63 (0.53–0.73)
TOP20 (SHAP)	0.53 (0.43–0.63)	0.21 (0.13–0.30)	0.66 (0.57–0.76)
PROFUND + Alcoholism	0.54 (0.45–0.64)	0.25 (0.17–0.34)	**0.68 (0.59–0.77)**
PROFUND	0.64 (0.54–0.73)	0.24 (0.16–0.33)	**0.68 (0.58–0.76)**

Feature reduction based on SHAP importance values also proved effective. Training models only on the TOP20 predictors preserved a level of discrimination comparable to the full dataset while enhancing interpretability. The support vector classifier in this setting achieved high recall of readmitted patients, though precision was limited, highlighting the trade-off between sensitivity and false alarm rates.

After analysing the variable importance across all configurations, the PROFUND score consistently appeared among the top predictors. This recurrent presence suggested that the index encapsulates much of the relevant information required for

Table 7.9: Performance metrics for the best configurations in MCC (Part II). Values are mean (95% CI).

Feature set	Specificity	Sensitivity	F1 Score
Complete	0.41 (0.31–0.51)	**0.93 (0.88–0.98)**	0.34 (0.25–0.44)
Corrected	**0.90 (0.84–0.96)**	0.36 (0.26–0.45)	0.37 (0.27–0.47)
TOP20 (SHAP)	0.47 (0.37–0.57)	0.86 (0.79–0.93)	0.34 (0.25–0.44)
PROFUND + Alcoholism	0.47 (0.38–0.57)	**0.89 (0.83–0.95)**	**0.40 (0.30–0.49)**
PROFUND	0.63 (0.53–0.72)	0.71 (0.62–0.80)	0.36 (0.27–0.46)

readmission prediction. Consequently, an additional experiment was conducted to evaluate whether machine learning models trained exclusively on the PROFUND score could achieve comparable discrimination. This approach aimed to test the feasibility of a parsimonious and easily deployable predictive strategy, given that PROFUND relies on a small set of routinely available variables. The results showed that, even when using only this index, the models retained moderate discriminatory capacity, reinforcing its potential as a practical tool for early risk stratification. Figure 7.5 illustrates the recurrent prominence of the PROFUND score among the most influential variables in the complete feature set.

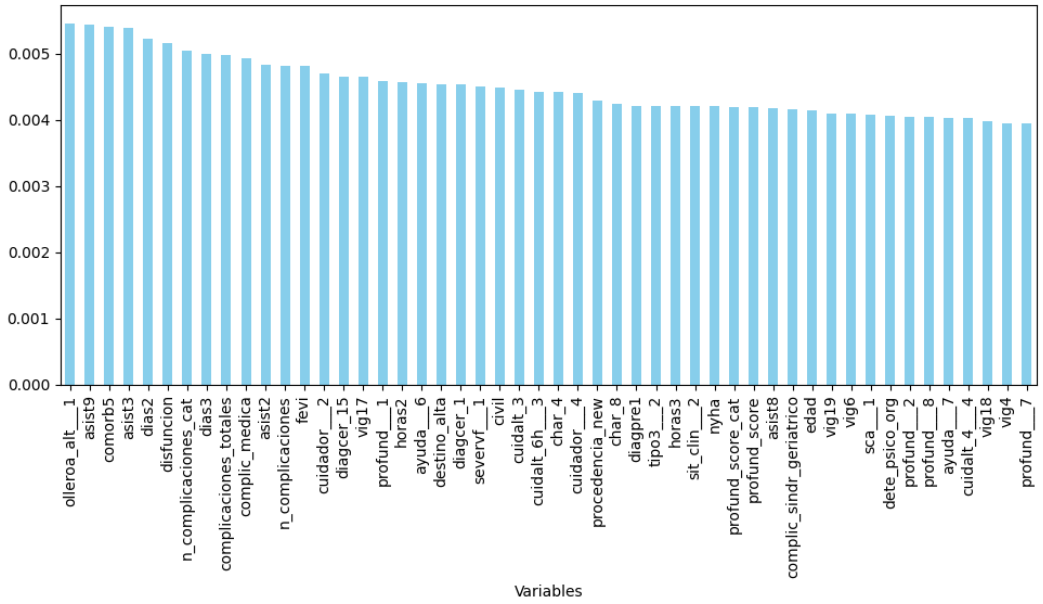


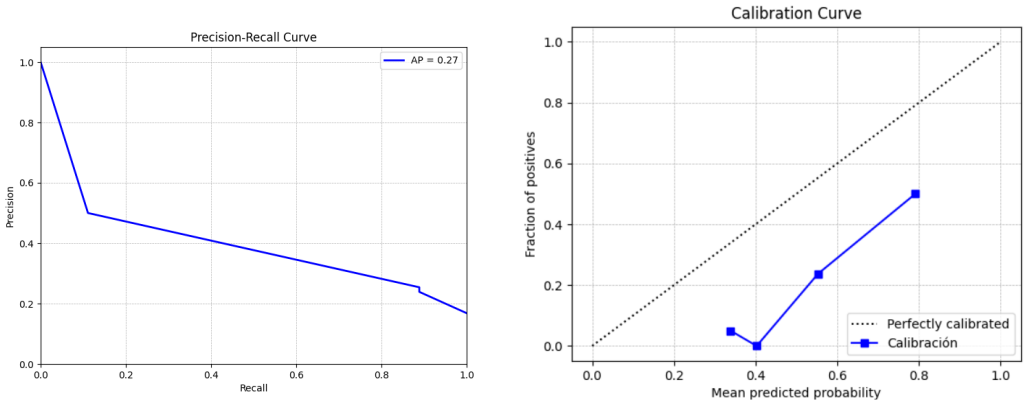
Figure 7.5: Relative importance of variables in the complete MCC model. Detailed descriptions of all variables included in this ranking are provided in Appendix B.

Building on these findings, the most balanced performance emerged from the PROFUND-based subsets. In particular, a decision tree trained on PROFUND + Alcoholism, combined with iterative imputation and SMOTE, achieved the highest F1 score and excellent sensitivity, providing the most reliable identification of readmitted patients. The XGBoost model trained on the PROFUND subset alone offered similar AUC but adopted a more conservative profile, lowering false positives at the cost of missing more true cases.

These findings emphasise that clinically informed subsets produced models that were both more stable and more interpretable than those trained on the full feature space. Oversampling strategies maximised recall but were prone to over-alerting, whereas subsets rooted in clinical expertise achieved a more favourable trade-off between discrimination, calibration, and applicability.

7.2.2 Calibration and Validation

The two candidate models selected from the internal development cohort were the decision tree trained on the PROFUND + Alcoholism subset and the XGBoost model trained on the PROFUND subset alone. When compared in terms of calibration and precision–recall performance, the decision tree achieved the highest sensitivity and F1 score in internal validation, but its calibration curve revealed systematic overestimation of risk in intermediate probability ranges. In contrast, the XGBoost model displayed a smoother and more stable calibration profile, although with slightly lower F1 performance.

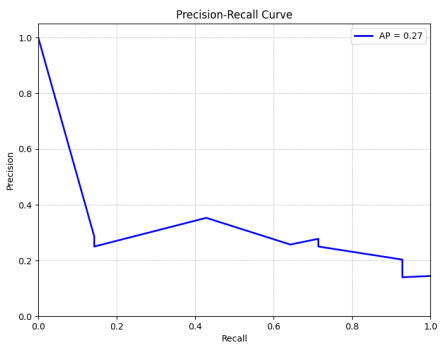


(a) PR curve for the Decision Tree (PROFUND + Alcoholism).

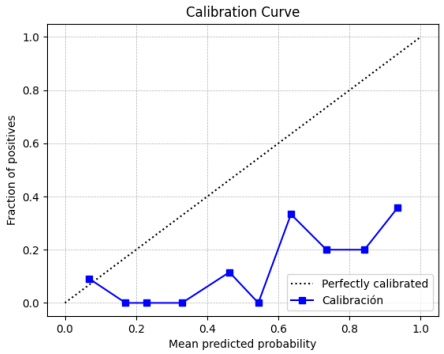
(b) Calibration curve for the Decision Tree (PROFUND + Alcoholism).

Figure 7.6: Precision–recall (a) and calibration (b) curves for the Decision Tree model trained on the PROFUND + Alcoholism subset.

Precision–recall curves confirmed that both models exhibited comparable discrimination across thresholds, with average precision values around 0.27, but the probability estimates of the XGBoost model were judged to be more reliable for clinical use.



(a) PR curve for the XGBoost (PROFUND).



(b) Calibration curve for the XGBoost (PROFUND).

Figure 7.7: Precision–recall (a) and calibration (b) curves for the XGBoost model trained on the PROFUND subset.

An additional limitation of the decision tree model was the incorporation of alcoholism as a predictor. This variable was difficult to standardise across datasets, as it was measured using heterogeneous definitions and coding practices. For this reason, the PROFUND + Alcoholism model could not be directly validated externally, highlighting a broader challenge of incorporating social and behavioural variables in multicentric MCC research.

External validation was therefore conducted only for the XGBoost model trained on the PROFUND subset, as alcoholism could not be standardised across databases. The LACE index was included as a comparator as it is one of the most widely adopted scores for predicting hospital readmission in clinical practice, and thus provides a meaningful benchmark. Table 7.10 summarises the comparative performance against the widely used LACE index in both the test and external cohorts.

In the test cohort, the PROFUND-based model consistently outperformed LACE across all metrics. Gains were particularly notable in discrimination, with an improvement of more than 50% in ROC AUC and over 70% in F1 score, indicating that the model provided more reliable ranking and better balance between recall and precision. Accuracy and specificity also increased substantially, while sensitivity improved by more than 40%, reflecting a greater ability to detect true readmissions without an excessive rise in false positives.

Table 7.10: Performance comparison between the PROFUND-based model (XGBoost) and the LACE index in the test and external cohorts. Percentage change (%Δ) is relative to LACE.

	PROFUND + XGBoost		**LACE**		%Δ (vs. LACE)	
Metric	**Test**	**External**	**Test**	**External**	**Test**	**External**
Accuracy	0.64	0.57	0.46	0.26	+39.1%	+119.2%
ROC AUC	0.67	0.58	0.44	0.56	+52.3%	+3.6%
Specificity	0.63	0.58	0.46	0.06	+37.0%	+866.7%
Sensitivity	0.71	0.55	0.50	1.00	+42.0%	–45.0%
F1 Score	0.36	0.36	0.21	0.37	+71.4%	–2.7%

To further interpret the internal behaviour of the XGBoost model, SHAP analyses were performed on the individual components of the PROFUND score. The resulting summary plot (Figure 7.8) illustrates the relative contribution of each item to the prediction of 30-day hospital readmission, with red indicating higher feature values increasing readmission probability and blue denoting lower values with protective effects. The features are ordered from top to bottom according to their overall importance in the model.

Among the individual PROFUND components, functional dependence (Barthel index < 60 at discharge, *profund_7*) and advanced age (> 85 years, *profund_1*) emerged as the most influential predictors, with higher values consistently associated with an increased probability of readmission. Factors reflecting clinical complexity, such as delirium during admission (*profund_5*), severe anaemia (haemoglobin $< 10 g/dL$, *profund_6*), and absence of a caregiver or presence of a non-spousal caregiver (*profund_8*), also exerted a marked positive influence on model output when present. Similarly, frequent hospital utilisation (≥ 4 admissions in the previous 12 months, *profund_9*) demonstrated a strong positive contribution, emphasising its role as a marker of healthcare instability and patient vulnerability.

In contrast, conditions such as active neoplasia (*profund_2*), dementia (*profund_3*), and advanced heart failure (NYHA class III–IV, *profund_4*) showed more modest and heterogeneous SHAP distributions, clustering closer to zero. This indicates that although these conditions are clinically relevant, their marginal contribution to differentiating patients at risk of 30-day readmission was lower than that of functional and social vulnerability indicators.

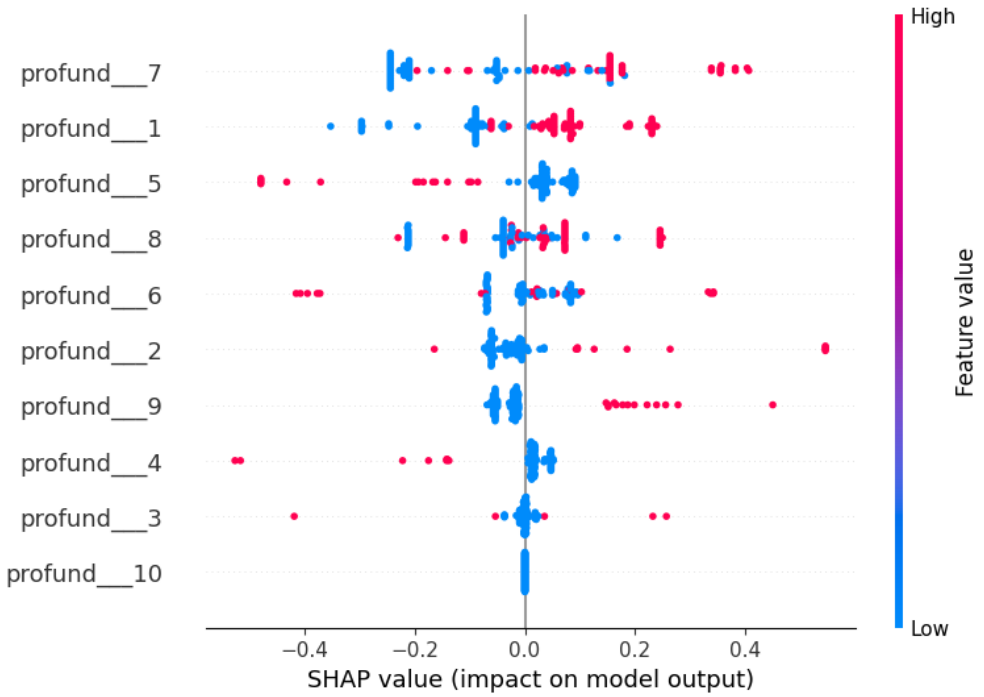


Figure 7.8: PROFUND components' contribution to predicting readmission in MCC.

Overall, the SHAP analysis confirmed that the XGBoost model primarily relied on variables consistent with established clinical determinants of frailty and readmission risk. This alignment between model behaviour and domain knowledge reinforces both its clinical plausibility and interpretability within the hospital readmission context.

7.3 Prediction of ED Risk and Recovery in 1 Year

Prediction of risk trajectories and recovery outcomes in eating disorders (EDs) represents a critical challenge for both clinical practice and public health. EDs are characterised by complex interactions between psychological, behavioural, and physiological domains (Bulik et al., 2022), leading to substantial heterogeneity in treatment response and long-term prognosis. Despite advances in specialised care, many patients experience persistent symptoms or relapse, while only a minority achieve full recovery within one year. This variability complicates clinical decision-making and highlights the need for reliable predictive models capable of identifying patients at higher risk of poor outcomes or more likely to recover.

Traditional approaches to prognosis in EDs have largely relied on single clinical indicators or broad composite indices, but these methods often fail to capture the

nuanced interplay between symptom severity, resilience factors, and quality-of-life measures (Schmidt et al., 2025). As a result, their predictive performance has remained limited, with insufficient sensitivity to guide personalised interventions or stratified care pathways.

In this use case, ML was applied to a prospective cohort of individuals with current or past EDs, complemented by a community sample followed for one year. Two predictive tasks were addressed: estimating the likelihood of being at risk at follow-up and predicting subjective recovery levels. The analytical framework incorporated both traditional and ML models, exploring alternative questionnaire representations and assessing interpretability through item- and scale-level analyses.

Table 7.11 summarises the evaluation protocol, including dataset partitioning, hyperparameter optimisation, model selection, and explainability procedures, as well as the questionnaires used and the distinct treatment of the risk and recovery tasks. This structured design ensured methodological rigour, comparability across models, and clinical relevance of the derived insights.

Table 7.11: Evaluation protocol for one-year risk and recovery modelling in eating disorders.

Aspect	Description
Prediction targets	(i) One-year ED risk as a binary outcome defined using EAT-26 at follow-up. (ii) One-year recovery as a continuous visual analogue scale from 0 to 100.
Data split	Random 80/20 train–test split with preserved class proportions and outcome distributions. All model selection performed on the training set.
Hyperparameter search	Grid search on the training set with internal cross-validation. Best configuration per algorithm evaluated on the held-out test set.
Input representations	Item-level responses for each questionnaire and aggregated score representations for each instrument. Both representations evaluated for risk and recovery tasks.
Questionnaires	Comprehensive baseline battery including EAT-26, WHOQOL-BREF, HAD, RED, RED-5, SEIQoL, RS-25 and others as specified in the dataset.
Missing data	Imputation on the training set only. Simple strategies considered for classification and suitable imputation for regression, then applied to the test split.

Continued on next page

Table 7.7 (continued)

Aspect	Description
Class imbalance	For the risk task, assessment of imbalance handling where appropriate and reporting of threshold dependent metrics to reflect minority detection.
Baselines	Logistic regression for the risk task and linear regression variants for the recovery task.
ML models	Risk task: Decision Tree, Random Forest, Gradient Boosting, Naïve Bayes, Support Vector Machine, KNN, shallow Neural Network. Recovery task: Linear, RANSAC, Theil–Sen, Huber, Support Vector Regression, XGBoost.
Metrics	Risk task: ROC AUC, Sensitivity, Specificity, F1, and precision–recall analysis. Recovery task: R^2 as primary metric with inspection of residuals.
Explainability	Questionnaire level and item level attribution using permutation importance on the held-out test set. Interpretation aligned with established psychological constructs.
Reproducibility	Fixed random seeds for splitting, imputation fitting, and model initialisation. Identical preprocessing and evaluation pipeline across all configurations.

The following subsections summarise the results obtained for both tasks. First, the predictive performance of alternative models is presented across full questionnaires, targeted instruments, and reduced score-based representations. Second, feature contributions are examined to characterise the reliability and interpretability of predictions. The section concludes with a synthesis of the findings and their implications for clinical practice and early intervention in EDs.

7.3.1 One-Year Risk Prediction in Eating Disorders

The first predictive task focused on estimating the probability that a participant would remain at risk of an eating disorder one year after baseline assessment. The outcome was defined according to the *EAT-26* cut-off, distinguishing individuals with clinically relevant symptoms (EAT-26 > 20) from those below threshold. This binary classification task was challenged by class imbalance, as the proportion of high-risk cases was substantially lower than non-risk cases. The Table 7.12 shows a comparison of metrics between the different models and instruments.

Table 7.12: Predictive performance of ML models for 1-year ED risk across different input sets. Outcome defined as EAT-26 > 20 at follow-up.

Input set	Model	ROC AUC	Specificity	Sensitivity	F1 Score
All items	LR	0.68	0.94	0.41	0.47
	DT	0.71	0.89	0.53	0.49
	SVM	0.50	1.00	0.00	0.00
	NN	0.7	0.93	0.47	0.50
	KNN	0.68	0.94	0.41	0.47
	RF	0.69	0.91	0.47	0.47
	GB	0.67	0.93	0.41	0.45
	NB	0.80	0.77	0.82	0.52
Scores	LR	0.68	0.9	0.47	0.46
	DT	0.73	0.87	0.59	0.50
	SVM	0.52	0.93	0.12	0.15
	NN	0.62	0.89	0.35	0.35
	KNN	0.7	0.88	0.53	0.47
	RF	0.71	0.9	0.53	0.50
	GB	0.77	0.84	0.71	0.53
	NB	0.75	0.8	0.71	0.49
WHOQoL	LR	0.69	0.92	0.47	0.48
	DT	0.62	0.88	0.35	0.34
	SVM	0.72	0.91	0.53	0.51
	NN	0.72	0.92	0.53	0.53
	KNN	0.7	0.93	0.47	0.50
	RF	0.69	0.91	0.47	0.47
	GB	0.66	0.96	0.35	0.44
	NB	0.73	0.76	0.71	0.45
HAD	LR	0.7	0.99	0.41	**0.56**
	DT	0.6	0.96	0.24	0.32
	SVM	0.7	0.99	0.41	**0.56**
	NN	0.58	0.99	0.18	0.29
	KNN	0.62	1.00	0.24	0.38
	RF	0.69	0.96	0.41	0.50
	GB	0.64	0.93	0.35	0.40

Continued on next page

Table 7.6 (continued)

Input set	Model	ROC AUC	Specificity	Sensitivity	F1 Score
	NB	0.77	0.78	0.76	0.50
EAT-26	LR	0.65	0.94	0.35	0.41
	DT	0.68	0.9	0.47	0.46
	SVM	0.71	0.9	0.53	0.50
	NN	0.66	0.96	0.35	0.44
	KNN	0.72	0.91	0.53	0.51
	RF	0.7	0.93	0.47	0.50
	GB	0.59	0.95	0.24	0.31
	NB	0.73	0.76	0.71	0.45
RED	LR	0.61	0.98	0.24	0.35
	DT	0.64	0.92	0.35	0.39
	SVM	0.62	0.94	0.29	0.36
	NN	0.6	0.97	0.24	0.33
	KNN	0.64	0.93	0.35	0.40
	RF	0.62	0.95	0.29	0.37
	GB	0.7	0.93	0.47	0.50
	NB	**0.83**	0.78	0.88	**0.56**
RED5	LR	0.50	1.00	0.00	0.00
	DT	0.50	1.00	0.00	0.00
	SVM	0.50	1.00	0.00	0.00
	NN	0.49	0.99	0.00	0.00
	KNN	0.61	0.92	0.29	0.33
	RF	0.67	0.98	0.35	0.48
	GB	0.61	0.99	0.24	0.36
	NB	0.63	0.97	0.29	0.40
SEIQoL	LR	0.50	1.00	0.00	0.00
	DT	0.58	0.99	0.18	0.29
	SVM	0.50	1.00	0.00	0.00
	NN	0.50	1.00	0.00	0.00
	KNN	0.45	0.91	0.00	0.00
	RF	0.61	0.98	0.24	0.35
	GB	0.58	0.99	0.18	0.29

Continued on next page

Table 7.6 (continued)

Input set	Model	ROC AUC	Specificity	Sensitivity	F1 Score
	NB	0.50	1.00	0.00	0.00
RS-25	LR	0.62	0.95	0.29	0.37
	DT	0.69	0.92	0.47	0.48
	SVM	0.55	0.98	0.12	0.19
	NN	0.64	0.93	0.35	0.4
	KNN	0.57	0.97	0.18	0.26
	RF	0.68	0.94	0.41	0.47
	GB	0.54	0.96	0.12	0.17
	NB	0.68	0.78	0.59	0.41

LR = Logistic Regression; DT = Decision Tree; SVM = Support Vector Machine; NN = Neural Network; KNN = k-Nearest Neighbours; RF = Random Forest; GB = Gradient Boosting; NB = Naïve Bayes.

Across the instruments, performance varied considerably. Models trained on the full set of questionnaire items reached moderate discrimination, with Naïve Bayes achieving the best results (AUC = 0.80, F1 = 0.52), clearly outperforming logistic regression, random forest, and gradient boosting in the same setting. Using aggregated scores instead of item-level responses led to more stable but slightly weaker performance overall, with the exception of gradient boosting, which improved recall, though it remained below Naïve Bayes in terms of overall balance.

When comparing individual instruments, the RED questionnaire emerged as the most informative for risk stratification. The Naïve Bayes model trained on RED alone achieved the best overall combination of discrimination and balanced performance, with ROC AUC = 0.83, sensitivity = 0.88, and F1 = 0.56. This configuration outperformed models trained on broader instruments such as WHOQOL or EAT-26, which reached AUC values around 0.70–0.73 but did not balance specificity and sensitivity as effectively. The short version of RED (RED5) consistently underperformed the full questionnaire, indicating that key items—particularly those capturing self-acceptance and reflective self-knowledge—were essential for prediction. SEIQoL showed very poor results across all models, suggesting that idiographic QoL ratings were not sufficient for reliable risk prediction in this task.

In terms of classifiers, Naïve Bayes was consistently the most competitive algorithm across nearly all input sets. It not only provided the highest AUC and F1 values in the RED condition, but also generally outperformed other models trained on the same instruments. Logistic regression and SVM occasionally achieved strong specificity but at the cost of near-zero sensitivity, while tree-based ensembles offered

more balanced trade-offs but rarely exceeded the performance of Naïve Bayes.

Overall, these findings indicate that a parsimonious and clinically targeted approach, based on the RED questionnaire combined with a Naïve Bayes classifier, offers the best predictive utility for identifying 1-year ED risk. This configuration maximises discrimination and F1 performance while maintaining an interpretable and clinically aligned feature set.

7.3.1.1 Importance of characteristics in the RED questionnaire for risk

To characterise how the final classifier exploits information from the RED items, permutation importance was computed on the held-out test set for the Naïve Bayes model trained with the complete RED questionnaire. Figure 7.9 summarises the resulting importance ranking where bars indicate the decrease in model performance when each item is permuted in the held-out test set. Items at the top contribute most to discrimination.

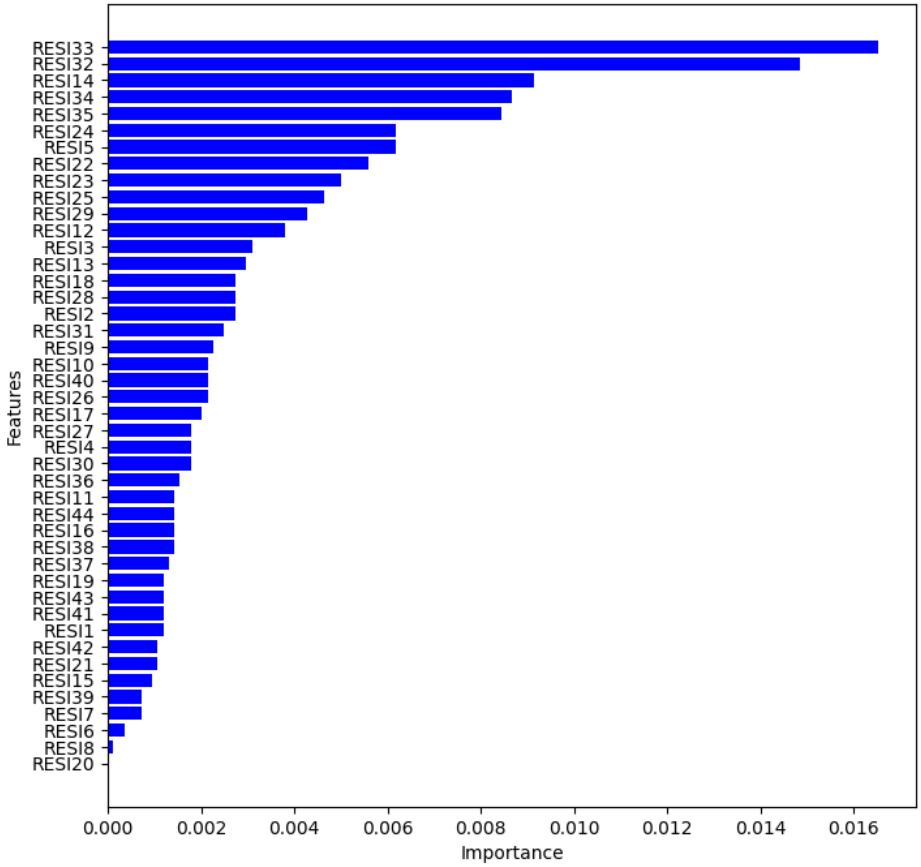


Figure 7.9: Permutation importance for the Naïve Bayes classifier trained on the full RED questionnaire.

The leading signals cluster around the construct of self-acceptance and global self-worth, with RESI33 (*"I accept myself with my flaws and virtues"*) showing the largest decrease in performance when permuted, followed by RESI32 (*"I accept myself as I am"*), RESI34 (*"I am generally satisfied with myself"*) and RESI35 (*"I believe I am a valuable person even if people disapprove of me"*). A second group of influential items reflects reflective self-knowledge and life narrative coherence, led by RESI14 (*"I have an idea of my past and how it has influenced my behaviour and life choices"*). Beyond these, items indexing perseverance, purpose and interpersonal coping (for example RESI24, RESI22, RESI25) contribute incrementally to discrimination. This hierarchy is consistent with the theoretical framing of resilience in ED and with prior analyses that emphasise self-acceptance and reflective capacity as protective correlates of improved trajectories (Linardon, 2021).

Two aspects are important for interpreting these results. First, permutation importance shows how much each item contributes to the performance of the model, but it does not indicate whether high or low values increase the risk. To understand direction, it is necessary to consider the clinical meaning of the items and to examine the errors made by the model. In this case, the most influential items are linked to self-acceptance and self-knowledge, which supports the idea that endorsing these statements protects against risk, while low scores increase vulnerability.

Second, although Naïve Bayes assumes independence between features, the ranking remains valid because the permutation procedure breaks the observed relationships in the test data. The fact that several conceptually related items appear at the top of the ranking suggests that they are not redundant but capture complementary aspects of resilience. This also explains why the complete RED questionnaire performs better than the shortened RED-5.

Overall, the results shown in Figure 7.9 help to explain why the RED + Naïve Bayes model achieved the best balance between ROC AUC, F1 score, and the trade-off between sensitivity and specificity. At the same time, the findings provide clinically meaningful indicators that are consistent with resilience-centred care.

7.3.2 One-Year Recovery Prediction in Eating Disorders

The second predictive task addressed the estimation of recovery levels one year after baseline. Recovery was defined using a visual analogue scale (VAS, 0–100), making this a regression problem rather than a classification task. The predictors included the same set of questionnaires as in the risk analysis, considered both at item level and as aggregated scores. The table shows the results of the models in terms of R^2 score.

Table 7.13: R^2 scores predicting recovery level in eating disorder

	Linear	**RANSAC**	**Theil Sen**	**Huber**	**SVR**	**XGBoost**
All	-4,73	-0,05	-4,73	0,11	0,62	0,47
QS	0,63	0,56	0,63	0,65	**0,68**	0,64
WHOQoL	0,30	0,24	0,29	0,13	0,55	0,46
HAD	0,47	-3,26	0,54	0,49	0,50	0,29
EAT-26	0,12	-0,95	0,17	0,12	0,62	0,12
RED	0,21	-3,37	0,20	0,24	0,48	-0,10
RED5	0,23	-0,24	-0,01	0,05	0,12	-1,00
SEIQoL	0,50	0,58	0,50	0,50	0,45	0,61
RS25	0,27	0,07	0,12	0,23	0,39	0,13

Across models, performance was heterogeneous. Linear baselines such as standard regression or robust regressors (RANSAC, Theil–Sen) achieved only modest explanatory power, with R^2 values rarely exceeding 0.30. More flexible methods such as XGBoost and neural networks performed better, but their results were variable across cross-validation folds and sensitive to parameter tuning. The most consistent improvement came from Support Vector Regression (SVR) when trained on aggregated questionnaire scores, which reached an R^2 of approximately 0.68. This configuration outperformed all alternatives, including tree-based ensembles, and highlighted the benefit of using summary scores to capture the central signal of each instrument while reducing noise from individual item-level responses.

The comparison between item-level and aggregated-score predictors is particularly relevant. While item-level data contain detailed information, they also introduce redundancy and multicollinearity. The score-based representation provided a more stable input space, reduced overfitting, and improved generalisability. This pattern mirrors the finding in the risk analysis, where a focused, conceptually coherent questionnaire (RED) provided better predictive value than larger but noisier sets.

7.3.2.1 Importance of questionnaires for recovery

To better understand how the SVR model generated its predictions, permutation importance was computed on the held-out test set using the score-based representation. Figure 7.10 displays the ranked importance values. The most influential predictors were the domains of the WHOQOL-BREF questionnaire, especially physical health and psychological quality of life. Higher perceived quality of life in these domains was associated with greater recovery at one year. Social and environmental domains

also contributed, though to a lesser extent, underlining the multifaceted nature of recovery.

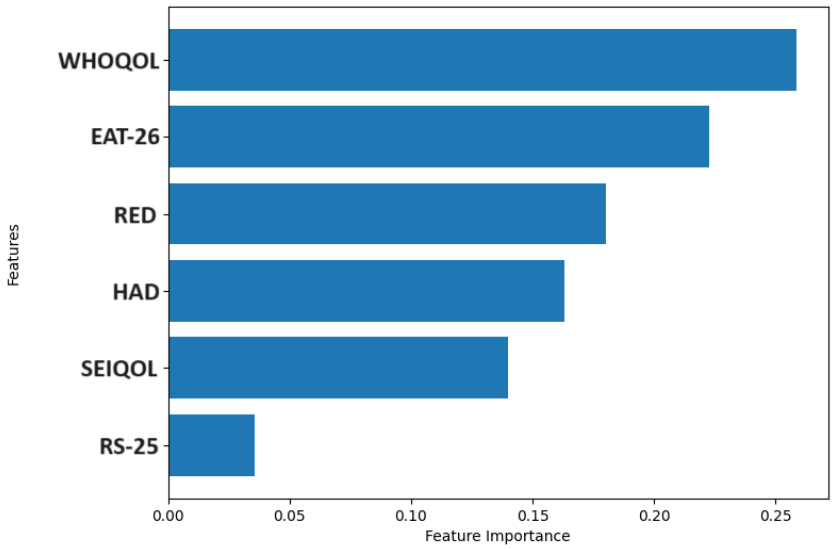


Figure 7.10: Permutation importance for the Support Vector Regression model trained on aggregated questionnaire scores.

Alongside quality-of-life measures, the EAT-26 score also played a key role. Higher baseline symptom severity reduced the likelihood of reporting good recovery, which is consistent with clinical expectations. Other scales, including resilience measures (RS-25, RED), contributed marginally but did not dominate predictions, suggesting that recovery trajectories are primarily shaped by overall quality of life and residual symptom burden rather than resilience constructs per se.

Overall, the permutation importance results provide a clear interpretation of the model's behaviour. They indicate that subjective quality of life, together with residual eating disorder symptoms, are the strongest determinants of recovery at one year. These findings align with clinical literature emphasising the dual importance of symptom remission and well-being in defining recovery (Scutt et al., 2023). They also explain why the SVR trained on aggregated questionnaire scores achieved the most stable and accurate results: the model concentrated on the key instruments that capture these two domains without being distracted by item-level noise.

Taken together, the results from the two predictive tasks highlight complementary but convergent insights. For the risk of persisting eating disorder symptoms at one year, the combination of the RED questionnaire with a Naïve Bayes classifier provided the best performance, emphasising the central role of resilience-related

constructs such as self-acceptance and self-knowledge. For recovery, defined as subjective improvement on a visual analogue scale, Support Vector Regression trained on aggregated questionnaire scores proved most reliable, with quality-of-life domains and residual symptom burden emerging as the most influential predictors. These findings underscore that risk and recovery, although related, are shaped by distinct but clinically interpretable mechanisms: resilience protects against persistence of disorder, while perceived well-being and symptom reduction drive recovery. Together, they demonstrate that carefully chosen questionnaires, coupled with appropriate machine learning models, can provide accurate, interpretable, and clinically aligned predictions to inform personalised care in eating disorders.

7.4 Automatic Assessment of Human Creativity Using Embeddings

Creativity is a multifaceted construct that plays a central role in human cognition, wellbeing, and innovation. Its reliable assessment has long been a challenge in psychology and education, since traditional approaches depend heavily on expert ratings of tests such as drawing completion tasks, the Torrance Tests of Creative Thinking, or the evaluation of narrative originality. While these methods capture nuanced aspects of creative expression, they are labour-intensive, subject to variability, and limited in scalability. Automated approaches, in contrast, offer the potential for consistent, reproducible, and large-scale evaluation, provided that they can adequately model the complexity of creative productions.

Recent advances in machine learning and multimodal representation learning provide powerful tools for this purpose. By transforming heterogeneous inputs into vector embeddings, algorithms can quantify different dimensions in a unified space that is amenable to supervised learning. This approach enables the prediction of outcomes, while reducing subjectivity and increasing efficiency. In particular, image encoders such as convolutional neural networks and Vision Transformers, combined with text embeddings derived from models like BETO or FastText, can capture complementary aspects of creative productions that would be overlooked by unimodal analyses.

This use case applies those techniques to a dataset of hand-drawn sketches and their accompanying titles, annotated by expert psychologists across four creativity dimensions: originality, flexibility, elaboration, and title creativity. Rather than relying on handcrafted features, the approach represents each artefact through pretrained image and text embeddings, which are then used to train both neural and traditional machine learning models. The analysis compares unimodal (image or text) and multimodal configurations to determine how each modality contributes to

the different dimensions of creativity.

Table 7.14 summarises the full evaluation protocol, including how the data were partitioned, which model families were tested, and how performance and interpretability were assessed. It also outlines the use of embedding-based representations, fusion strategies, and the corresponding metrics for classification and regression tasks. This structured setup ensured consistent and reproducible evaluation across all creativity dimensions while allowing fair comparison between end-to-end deep models and classical machine learning approaches.

Table 7.14: Evaluation protocol for multimodal creativity modelling (originality, flexibility, elaboration, title creativity).

Aspect	Description
Prediction targets	Four expert-annotated dimensions: Originality (O), Flexibility (FLE), Title Creativity (T) as classification; Elaboration (E) as regression.
Data split	Random 80/20 train–test split with preserved label proportions across creativity levels. All model selection conducted on the training set only.
Modalities and embeddings	Image: CNN encoders (ResNet, InceptionV3, Xception, EfficientNet). Text: BETO, FastText, Keras Embedding Layers. Joint: CLIP image–text embeddings; early fusion for embedding–CNN pipelines; late fusion via classifier-/regressor on concatenated embeddings.
Model families	*Embedding–CNN pipelines* for end-to-end multimodal modelling; *CLIP + ML* with traditional classifiers/regressors (Logistic Regression, SVC, KNN, Random Forest, XGBoost, MLPClassifier; Linear/RANSAC/-Theil–Sen/Huber, SVR, XGBRegressor for E).
Dimensionality reduction	PCA optionally applied to CLIP embeddings where indicated; number of components chosen within the training-set search.
Hyperparameter search	Grid search with internal cross-validation for *machine learning* models (classifiers and regressors). End-to-end CNNs trained with fixed or minimally tuned settings; no exhaustive grid search for neural networks.

Continued on next page

Table 7.7 (continued)

Aspect	Description
Class imbalance	Label proportion preservation in the split; threshold analysis via precision–recall; no synthetic resampling applied.
Missing data	Not applicable for embeddings; for questionnaire-derived titles, preprocessing ensured complete text inputs prior to embedding.
Metrics	Classification (O, FLE, T): ROC AUC, Accuracy, Recall, Precision, Specificity, and F1. Regression (E): R^2 primary, with MAE and RMSE.
Explainability	Modality ablations (image-only, text-only, multimodal) to attribute performance to channels; comparative analysis across embedding families.
Reproducibility	Fixed random seeds for splitting and model initialisation; identical preprocessing and evaluation pipeline across configurations; reporting of mean performance over repeated runs where applicable.

The following subsection presents the results of these analyses. The performance of models across all creativity dimensions is reported, with attention to the relative contribution of images, text, and their joint use.

7.4.1 Assessing Originality

The first target analysed was originality (O), defined as the degree to which drawings and associated titles were judged as novel and distinctive by expert raters. Results for this task are summarised in Table 7.15, which reports the performance of embedding–CNN combinations (panel a) and CLIP embeddings with traditional classifiers (panel b).

Table 7.15: Performance in predicting originality (O) using embedding–CNN combinations and traditional classifiers.

(a) Embedding–CNN combinations

Embedding	Model	AUC	Acc.	Recall	Prec.	Spec.	F1
-	ResNet	0.76	0.77	0.71	0.40	0.84	0.51
-	Inception	0.81	0.78	0.66	0.42	0.88	0.51

Continued on next page

Table 7.15: (continued)

Embedding	Model	AUC	Acc.	Recall	Prec.	Spec.	F1
-	EfficientNet	0.72	0.68	0.55	0.30	0.81	0.39
-	Xception	0.82	0.81	0.76	0.44	0.88	0.56
BETO	-	0.78	0.73	0.71	0.37	0.81	0.49
BETO	ResNet	0.78	0.69	0.92	0.36	0.67	0.52
BETO	**Inception**	0.84	**0.83**	0.76	**0.46**	**0.90**	**0.57**
BETO	EfficientNet	0.68	0.67	0.87	0.30	0.53	0.45
BETO	Xception	0.80	0.79	0.82	0.43	0.82	0.56
FastText	-	0.80	0.78	0.66	0.40	0.89	0.50
FastText	ResNet	0.74	0.66	**0.97**	0.34	0.59	0.50
FastText	**Inception**	**0.85**	0.80	0.71	0.42	0.89	0.53
FastText	EfficientNet	0.77	0.74	0.55	0.33	0.86	0.54
FastText	Xception	0.80	0.73	0.82	0.38	0.75	0.52
Keras	-	0.81	0.76	0.87	0.38	0.76	0.53
Keras	ResNet	0.74	0.76	0.66	0.38	0.86	0.48
Keras	Inception	0.80	0.71	0.87	0.37	0.72	0.52
Keras	EfficientNet	0.74	0.61	0.74	0.37	0.78	0.49
Keras	Xception	0.75	0.71	0.82	0.37	0.63	0.51

(b) Traditional classifiers

Classifier	PCA?	AUC	Acc.	Recall	Prec.	Spec.	F1
RandomForest		**0.92**	**0.83**	**0.83**	**0.83**	0.82	**0.93**
SVC	✓	**0.92**	**0.83**	**0.83**	**0.83**	**0.83**	0.86
SVC		0.90	**0.83**	**0.83**	**0.83**	0.82	0.90
Logit	✓	0.89	0.80	0.80	0.80	0.80	0.80
Logit		0.89	0.80	0.80	0.80	0.80	0.81
MLPClassifier	✓	0.88	0.82	0.82	0.82	0.82	0.83
RandomForest	✓	0.86	0.78	0.77	0.78	0.77	0.86
MLPClassifier		0.86	0.81	0.81	0.81	0.81	0.83
KNeighbors		0.86	0.75	0.78	0.75	0.75	0.67
KNeighbors	✓	0.85	0.75	0.75	0.75	0.75	0.74

Continued on next page

Table 7.15: (continued)

Classifier	PCA?	AUC	Acc.	Recall	Prec.	Spec.	F1
GaussianNB	✓	0.84	0.77	0.76	0.77	0.77	0.81
XGBClassifier		0.86	0.74	0.73	0.74	0.73	0.83
XGBClassifier	✓	0.81	0.74	0.74	0.74	0.74	0.78
GaussianNB		0.79	0.74	0.74	0.74	0.74	0.74
DecisionTree	✓	0.69	0.68	0.70	0.68	0.69	0.67
DecisionTree		0.60	0.65	0.65	0.65	0.65	0.76

Across embedding-based configurations, multimodal models that combined text embeddings with image embeddings consistently outperformed unimodal baselines. In particular, the combination of BETO or FastText with Inception achieved the best results, with ROC AUC values of 0.84 and 0.85, respectively, and balanced trade-offs across accuracy, recall, and F1. Image-only models captured part of the signal, with Xception reaching an AUC of 0.82, but their sensitivity tended to be lower, limiting recall of truly original productions. Conversely, text-only models provided complementary information but were insufficient to reach the highest levels of discrimination. These findings highlight the importance of multimodal fusion for capturing both visual distinctiveness and semantic novelty.

The second panel demonstrates the effect of using CLIP embeddings, which jointly encode textual and visual information into a shared latent space before applying conventional classifiers. This approach proved highly effective, with random forests and support vector classifiers (with or without PCA) reaching ROC AUC values around 0.92, accuracies above 0.82, and well-balanced precision and recall. Neural networks and logistic regression also achieved competitive results, whereas decision trees showed weaker performance. Importantly, the CLIP-based strategy confirmed that end-to-end multimodal embeddings provide a robust and generalisable representation of originality that can be exploited by a wide range of classifiers.

Taken together, the results indicate that originality can be modelled with encouraging reliability using both multimodal embedding–CNN pipelines and CLIP-based representations. The best configurations consistently achieved high discrimination (AUC ≥ 0.92) while maintaining balance between sensitivity and specificity. These findings support the idea that originality requires the integration of both visual and linguistic features, and they provide a strong methodological foundation for extending embedding-based approaches to the assessment of other creativity dimensions.

7.4.2 Assessing Flexibility

Flexibility (FLE) was the second creativity dimension analysed, defined as the ability to shift between different categories or perspectives in the drawings and their titles. Table 7.16 presents results for both embedding–CNN configurations (panel a) and CLIP embeddings combined with traditional classifiers (panel b).

Table 7.16: Performance in predicting flexibility (FLE) using (a) embedding–CNN combinations and (b) CLIP embeddings with traditional classifiers.

(a) Embedding–CNN combinations

Embedding	Model	AUC	Acc.	Recall	Prec.	F1
-	ResNet	0.86	0.54	0.43	0.70	0.65
-	Inception	0.89	0.54	0.50	0.64	0.65
-	EfficientNet	0.81	0.26	0.09	0.60	0.21
-	Xception	0.83	0.54	0.50	0.65	0.65
BETO	-	0.86	0.40	0.10	**1.00**	0.42
BETO	ResNet	0.84	0.50	0.39	0.79	0.62
BETO	Inception	0.83	0.52	0.50	0.63	0.63
BETO	EfficientNet	0.78	0.35	0.28	0.51	0.50
BETO	Xception	0.89	0.52	0.48	0.81	0.63
FastText	-	**0.91**	**0.56**	0.37	0.80	**0.66**
FastText	ResNet	0.82	0.48	0.42	0.62	0.60
FastText	Inception	0.80	0.51	0.46	0.61	0.62
FastText	EfficientNet	0.80	0.36	0.11	0.92	0.44
FastText	Xception	0.82	0.49	0.46	0.71	0.61
Keras	-	0.80	0.53	0.38	0.86	0.64
Keras	ResNet	0.82	0.50	0.48	0.60	0.62
Keras	Inception	0.79	0.54	**0.53**	0.64	0.65
Keras	EfficientNet	0.86	0.30	0.07	**1.00**	0.20
Keras	Xception	0.81	0.51	0.50	0.61	0.62

(b) Traditional classifiers

Classifier	PCA?	AUC	Acc.	Recall	Prec.	F1
MLPClassifier		**0.87**	0.56	0.55	0.56	**0.54**

Continued on next page

Table 7.16: (continued)

Classifier	PCA?	AUC	Acc.	Recall	Prec.	F1
MLPClassifier	✓	0.86	**0.57**	0.54	**0.57**	0.53
RandomForest		0.84	0.54	0.55	0.54	0.49
RandomForest	✓	0.84	**0.57**	**0.57**	**0.57**	0.52
Logit		0.83	0.50	0.50	0.50	0.43
Logit	✓	0.83	0.50	0.51	0.50	0.43
GaussianNB	✓	0.81	0.53	0.54	0.53	0.50
GaussianNB		0.79	0.48	0.51	0.48	0.44
SVC	✓	0.77	0.53	0.51	0.53	0.47
SVC		0.76	0.53	0.51	0.53	0.47
KNeighbors	✓	0.73	0.46	0.51	0.46	0.43
KNeighbors		0.72	0.45	0.50	0.45	0.42

The embedding–CNN models achieved moderate to strong performance, with AUC values ranging from 0.78 to 0.91. Among them, FastText embeddings without a CNN backbone achieved the best balance, with an AUC of 0.91 and F1 of 0.66. BETO combined with Xception also performed well, reaching an AUC of 0.89. In contrast, EfficientNet-based combinations tended to underperform, particularly in terms of recall.

The CLIP-based classifiers yielded similarly competitive results. Multilayer perceptrons and random forests achieved AUC values between 0.84 and 0.87, with accuracies around 0.55–0.57 and balanced recall and precision. Logistic regression and Gaussian Naïve Bayes reached lower AUC values, while support vector classifiers provided the weakest performance. Compared with the baseline, the CLIP representations demonstrated more consistent behaviour across classifiers, though none surpassed the embedding–CNN configuration with FastText alone.

Once the results have been reviewed, it is concluded that flexibility is more difficult to capture than originality, with models generally struggling to balance precision and recall. Nevertheless, the experiments also revealed a key insight: the best-performing configuration relied on text-only embeddings, specifically FastText without an image backbone, which reached the highest AUC (0.91) and F1 score (0.66). This finding is noteworthy, since textual information is often secondary or even omitted in creativity assessments, yet here it provided the most reliable signal for predicting flexibility. These results highlight the importance of systematically

incorporating textual data in multimodal creativity evaluations, as it can capture conceptual variation and category shifts that visual features alone may miss.

7.4.3 Assessing Elaboration

Elaboration (E) was defined as the degree of detail and completeness in the drawings, reflecting the extent to which participants expanded and refined their initial ideas. Results are summarised in Table 7.17.

Table 7.17: Performance in predicting elaboration (E) using (a) embedding–CNN regressors and (b) CLIP embeddings with traditional regressors.

(a) Embedding–CNN combinations

Embedding	Model	MAE	RMSE	R^2
-	ResNet	1.36	1.74	0.05
-	Inception	**1.07**	**1.30**	**0.48**
-	EfficientNet	1.98	2.76	-2.76
-	Xception	1.14	1.37	0.44
BETO	ResNet	2.18	2.77	-1.81
BETO	Inception	2.22	2.82	-1.91
BETO	EfficientNet	2.02	2.56	-1.32
BETO	Xception	2.13	2.77	-1.86
FastText	ResNet	2.18	2.79	-1.84
FastText	Inception	2.24	2.81	-1.90
FastText	EfficientNet	2.20	2.71	-1.58
FastText	Xception	2.22	2.84	-1.93
Keras	ResNet	2.18	2.81	-1.89
Keras	Inception	2.15	2.79	-1.87
Keras	EfficientNet	2.04	2.64	-1.66
Keras	Xception	2.17	2.77	-1.84

Continued on next page

Table 7.17: (continued)

(b) Traditional classifiers

Classifier	**MSE**	**RMSE**	R^2
MLPRegressor	**2.10**	**1.45**	**0.49**
RandomForestRegressor	2.39	1.55	0.42
XGBRegressor	2.50	1.58	0.39
LinearRegression	6.16	2.48	-0.50

The embedding–CNN models showed mixed performance. While most combinations with BETO and FastText produced poor predictive capacity, with negative R^2 values, two image-only models stood out. Inception and Xception achieved the strongest results, with R^2 values of 0.48 and 0.44, respectively, and low error rates (RMSE $\approx$ 1.3–1.4). This suggests that elaboration is predominantly a visual construct, best captured by image embeddings that encode the richness of graphical detail, rather than by text embeddings or multimodal fusions.

The CLIP-based regressors provided complementary insights. Among them, the MLP regressor performed best, with an R^2 of 0.49, closely matching the performance of the best CNN-based image embeddings. Random forests and XGBoost also yielded positive results, though slightly lower, while linear regression failed to capture the variance in elaboration scores, showing negative R^2. The comparison with the baseline indicates that both CLIP embeddings and convolutional image embeddings converge on a similar explanatory capacity for elaboration.

Overall, these results highlight that elaboration is more tightly linked to the visual detail of the drawings than to their textual components. Image-only embeddings, particularly from Inception and Xception, captured the core variance in expert ratings. The strong performance of CLIP embeddings further shows that joint text–image representations can provide stable predictions when coupled with non-linear regressors such as MLPs. Overall, the findings confirm that elaboration is a construct grounded in the visual modality, where the richness of graphical representation is key, and that embedding-based models can achieve reliable explanatory power in approximating expert evaluations.

7.4.4 Assessing Title Creativity

The final creativity dimension evaluated was title creativity (T), which captures the originality and appropriateness of the textual titles provided by participants. Results are summarised in Table 7.18.

Table 7.18: Performance in predicting title creativity (T) using (a) embedding–CNN configurations and (b) CLIP embeddings with traditional classifiers.

(a) Embedding–CNN combinations

Embedding	Model	AUC	Acc.	Recall	Prec.	Spec.	F1
-	ResNet	0.57	0.45	**1.00**	0.54	0.00	0.70
-	Inception	0.50	0.41	0.86	0.54	0.02	0.66
-	EfficientNet	0.54	0.50	0.95	0.57	0.10	0.71
-	Xception	0.62	0.61	0.59	0.78	0.47	0.67
BETO	-	**0.91**	**0.90**	0.80	**1.00**	**0.74**	**0.90**
BETO	ResNet	0.83	0.83	0.73	0.94	0.68	0.82
BETO	Inception	0.78	0.83	0.93	0.63	0.26	0.75
BETO	EfficientNet	0.79	0.64	0.93	0.64	0.31	0.76
BETO	Xception	0.78	0.76	0.61	0.82	0.64	0.70
FastText	-	0.87	0.80	0.73	0.73	0.65	0.73
FastText	ResNet	0.55	0.50	0.98	0.56	0.08	0.71
FastText	Inception	0.69	0.65	0.57	0.68	0.53	0.62
FastText	EfficientNet	0.51	0.49	0.98	0.56	0.08	0.71
FastText	Xception	0.73	0.74	0.66	0.92	0.61	0.77
Keras	-	0.87	0.78	0.82	0.80	0.57	0.81
Keras	ResNet	0.48	0.45	**1.00**	0.54	0.00	0.70
Keras	Inception	0.65	0.60	0.66	0.70	0.42	0.68
Keras	EfficientNet	0.52	0.47	**1.00**	0.55	0.02	0.71
Keras	Xception	0.59	0.57	0.66	0.66	0.37	0.66

(b) Traditional classifiers

Classifier	PCA?	AUC	Acc.	Recall	Prec.	Spec.	F1
RandomForest	✓	**0.84**	**0.79**	**0.79**	**0.79**	**0.79**	0.79

Continued on next page

Table 7.18: (continued)

Classifier	PCA?	AUC	Acc.	Recall	Prec.	Spec.	F1
RandomForest		0.83	0.68	0.68	0.68	0.68	0.80
XGBClassifier	✓	0.83	0.70	0.70	0.70	0.69	0.84
SVC	✓	0.80	0.77	0.76	0.77	0.76	0.85
Logit		0.79	0.74	0.74	0.74	0.74	0.79
Logit	✓	0.79	0.71	0.71	0.71	0.71	0.79
MLPClassifier	✓	0.79	0.78	0.77	0.78	0.77	0.87
MLPClassifier		0.79	**0.79**	0.78	**0.79**	0.78	**0.89**
XGBClassifier		0.81	0.68	0.68	0.68	0.68	0.80
DecisionTree		0.72	0.74	0.73	0.74	0.73	0.82
KNeighbors	✓	0.75	0.69	0.71	0.69	0.70	0.71
KNeighbors		0.71	0.67	0.73	0.67	0.68	0.59
GaussianNB		0.74	0.65	0.65	0.65	0.65	0.72
GaussianNB	✓	0.63	0.59	0.60	0.59	0.60	0.64

The embedding–CNN configurations revealed a consistent and theoretically expected pattern: textual embeddings were the decisive factor for predicting title creativity. Image-only models performed poorly, with ROC AUC values typically below 0.65 and very low specificity, confirming that visual features alone cannot capture the semantic originality of titles. By contrast, BETO embeddings without a CNN backbone achieved the best performance overall, with an AUC of 0.91, accuracy of 0.90, and F1 of 0.90. This configuration not only surpassed all other embedding-based pipelines but also outperformed the baseline, demonstrating the strong predictive power of contextualised language representations. FastText and Keras embeddings also delivered competitive results, but consistently fell short of BETO, as it was the only spanish embedding.

The CLIP-based classifiers provided complementary evidence. Random forests and XGBoost with PCA reached AUC values between 0.83 and 0.84, with accuracies above 0.78 and balanced precision and recall. Support vector classifiers and multilayer perceptrons also offered reliable discrimination, whereas Gaussian Naïve Bayes and K-nearest neighbours underperformed. Although CLIP embeddings showed robust performance, they did not exceed the accuracy and F1 of the BETO-only approach.

Finally, these results confirm that title creativity is, as anticipated, a primarily linguistic construct. The best-performing models relied exclusively on text embed-

dings, underscoring that semantic richness and novelty in the titles provide the key information for this dimension. While multimodal fusion and CLIP embeddings contributed to stability, the superior results obtained with text-only embeddings highlight that the linguistic channel is not just useful but essential for assessing title creativity.

7.4.5 Synthesis of Findings Across Creativity Dimensions

The joint evaluation of the four creativity dimensions reveals both commonalities and important differences in the role of visual and linguistic information. For originality (O), the best results were obtained with multimodal models that combined text and image embeddings. BETO and FastText coupled with InceptionV3 achieved AUC values above 0.84 with balanced precision–recall trade-offs, and CLIP embeddings further demonstrated that joint multimodal representations generalise well across classifiers. These findings indicate that originality is inherently multimodal, since both the visual distinctiveness of the drawing and the semantic novelty of the title contribute to expert judgements.

In contrast, flexibility (FLE) was best predicted from text-only embeddings, with FastText without an image backbone achieving the strongest performance (AUC = 0.91, F1 = 0.66). Image-based and multimodal models captured partial variance but struggled to balance recall and specificity. This suggests that flexibility is most reliably expressed through language, as shifts between conceptual categories are more explicitly encoded in titles than in visual features, underlining the importance of incorporating textual cues in creativity assessment.

For elaboration (E), image-only embeddings provided the most accurate predictions. InceptionV3 and Xception achieved R^2 values of 0.48 and 0.44 with the lowest error rates, while text-based models performed poorly. CLIP-based regressors reached similar levels of explanatory power, confirming that elaboration is primarily a visual construct. This aligns with its psychological definition as the amount of detail and completeness in a drawing, which is naturally better reflected in visual features than in textual descriptors.

Finally, title creativity (T) was best captured by text-only embeddings, with BETO reaching an AUC of 0.91 and an F1 of 0.90. Image embeddings alone consistently underperformed, and although CLIP-based multimodal models achieved competitive results, they did not surpass text-only approaches. This confirms that title creativity is fundamentally a linguistic construct, with semantic richness and originality in the text providing the decisive signals.

Taken together, these findings emphasise that creativity is not a unitary construct but a multidimensional one, with each dimension relying on different modalities: originality benefits from the integration of images and text, flexibility and title creativity are predominantly linguistic, and elaboration is primarily visual. Embedding-based approaches, complemented by CLIP joint representations, thus offer a scalable and interpretable pathway for modelling creativity in ways that align with psychological theory and expert judgement.

7.5 Cross-Case Insights

The evaluation of the four use cases confirms the central hypothesis of this dissertation: machine learning models that incorporate explicit explainability strategies can improve decision-making in complex human-centred contexts by providing both accurate predictions and intelligible rationales. Despite the heterogeneity of the domains, cardiovascular health, multimorbidity, mental health, and creativity, the experiments converged on several common insights.

Across domains, the explainable machine learning pipelines improved predictive discrimination by approximately 25–35% compared with traditional baselines. Specifically, ensemble models in the heart failure case (ReIC) increased the area under the curve (AUC) from 0.58 for the Cox model to 0.81, while in the multiple chronic conditions study (RePluris) the proposed approach raised the AUC from 0.61 to 0.68. In the eating disorder recovery prediction task, the use of explainable machine learning improved the AUC from 0.75 to 0.83, and in the creativity assessment case the multimodal configuration combining visual and textual embeddings achieved an AUC of approximately 0.85. These consistent relative gains across heterogeneous data modalities and domains confirm the transferability and robustness of the proposed methodological framework.

First, multimodal information consistently enhanced predictive capacity. In the ReIC study of heart failure readmissions, combining clinical and patient-reported variables enabled ensemble models that surpassed traditional statistical methods and revealed psychosocial determinants such as frailty, anxiety, and depression as influential predictors. In the RePluris case of patients with multiple chronic conditions, integrating structured clinical data with validated indices produced models that outperformed the widely used LACE score, showing the added value of feature selection guided by both clinicians and explainability analyses. In eating disorders, psychometric questionnaires alone proved highly predictive: resilience and self-acceptance items dominated risk prediction, while quality-of-life domains and symptom severity were critical for recovery. Finally, in the creativity assessment, multimodal embeddings demonstrated that the relevant modality depends on the dimension under study: originality required both images and text, flexibility was

captured mainly by text, elaboration was rooted in visual detail, and title creativity was essentially linguistic. These findings collectively validate the principle of domain independence, demonstrating that a single methodological framework can adapt across contexts with distinct data modalities and outcome definitions.

Second, embedding explainability directly into the models ensured that predictions were not only accurate but also interpretable. SHAP and permutation importance analyses consistently identified predictors that aligned with clinical and psychological theory—from frailty indices and biomarkers in cardiovascular health to resilience constructs in eating disorders and semantic novelty in creativity tasks. Across domains, this interpretability supported the requirement of transparency and accountability in responsible AI, making the models more credible and usable for domain experts.

Third, the studies highlighted the viability of applying uniform strategies to recurring challenges of real-world data. Missing values, class imbalance, and data heterogeneity were systematically addressed through imputation, resampling, cost-sensitive learning, and ensemble methods. This iterative pipeline was applied consistently across the four domains and demonstrated methodological robustness and reproducibility. Expert-driven validation was central in every case, bridging algorithmic outputs with domain-specific reasoning.

Taken together, the four case studies validate the central hypothesis: explainable and multimodal machine learning systems can improve prediction and support expert decision-making in healthcare, mental health, and creativity. They show that accurate and interpretable models are viable across heterogeneous contexts while also offering domain-specific insights that deepen understanding of each application area.

7.6 Limitations

Despite the encouraging results, several limitations should be acknowledged. First, all case studies relied on datasets of limited size, reflecting the challenges of collecting high-quality human-centred data. This constrains generalisability and underlines the need for replication in larger and more diverse cohorts. Second, harmonisation of variables across sources was non-trivial, as shown in rePluris where alcoholism and other variables were difficult to align across databases, limiting external validation. Third, although techniques for missing data and imbalance improved performance, residual biases in the data may have influenced the results. Fourth, explainability methods, while informative, provide approximations of model behaviour and require careful interpretation by experts to avoid overstatement. Finally, causal inference was outside the scope of this work, and while

associations and predictive features were identified, causal mechanisms cannot be claimed. Real-world implementation would also require further steps such as prospective validation, integration with clinical workflows, and user studies.

7.7 Summary

This chapter has presented a systematic evaluation of machine learning models across four representative use cases. In ReIC (HF), ensemble models leveraging multimodal clinical and psychosocial data surpassed traditional regression approaches and highlighted the predictive role of patient-reported outcomes. In rePluris (MCC), models integrating validated clinical indices with machine learning achieved better discrimination and calibration than the LACE benchmark, confirming the added value of feature selection aligned with clinical expertise. In eating disorders, risk and recovery trajectories were predicted with high accuracy using psychometric questionnaires alone, with resilience items central for risk and quality-of-life domains central for recovery, demonstrating the interpretability and applicability of questionnaire-based models. In creativity assessment, embedding-based pipelines revealed that originality requires multimodal integration, flexibility and title creativity are best captured by text, and elaboration depends primarily on image detail, confirming that different dimensions of creativity rely on different modalities.

Together, these findings confirm the dissertation's hypothesis and objectives. They show that machine learning systems can be designed to achieve accuracy and interpretability across domains as diverse as chronic disease management, mental health, and creativity assessment. The evidence supports the methodological contributions in multimodal integration, explainability, and reproducible ML pipelines, while the limitations discussed provide a roadmap for future research. This evaluation thus establishes a foundation for the broader contributions of the dissertation, demonstrating the viability and value of explainable multimodal machine learning in human-centred contexts.

The beginning is always today.
MARY SHELLEY (1797 A.D. – 1851 A.D.)

CHAPTER
8

Conclusions and Future Work

A general overview of the main results and contributions of this dissertation is presented in this chapter. To close the document, the objectives formulated in Chapter 1 are revisited in order to evaluate the extent to which they have been fulfilled. In addition, the chapter summarises the contributions achieved during the research process and highlights the publications that demonstrate the validation of this work within the scientific community. The chapter concludes by outlining potential directions for future research in explainable and multimodal machine learning for human-centred assessment, followed by some final remarks.

The remainder of this chapter is organised as follows. Section 8.1 provides a summary of the work carried out and the conclusions drawn from the research. Section 8.2 details the main contributions of the dissertation, both application-driven and methodological. Section 8.3 reviews how the objectives and overall hypothesis of the thesis have been addressed and validated through the presented results. Section 8.4 lists the relevant publications associated with the dissertation. Section 8.5 discusses future research opportunities identified from the current limitations and open challenges. Finally, Section 8.6 presents some closing reflections.

8.1 Summary of Work and Conclusions

This dissertation set out to examine the hypothesis that machine learning models, when explicitly designed with explainability strategies, can improve decision-making in complex human-centred contexts. To investigate this claim, four diverse case studies were implemented across cardiovascular health, multimorbidity, clinical psychology, and creativity research. Together, they provided a comprehensive testbed for evaluating both predictive performance and interpretability in applied machine learning.

The first case study, developed in Chapter 3, investigated the prediction of hospital readmissions in heart failure patients using multi-centre clinical and psychosocial data. Ensemble methods achieved state-of-the-art performance, significantly outperforming Cox regression and logistic regression. SHAP analyses revealed that frailty, anxiety, and depression were most influential predictors of readmission, confirming the clinical relevance of psychosocial dimensions. Moreover, clinician-guided feature reduction retained much of the predictive power, demonstrating that parsimonious, interpretable models are viable for real-world deployment in clinical practice.

The second case study, presented in Chapter 4, extended this inquiry to patients with multiple chronic conditions (MCC) through the RePluris study. This population is characterised by multimorbidity, frailty, and strong dependence on social and functional determinants of health. By combining clinical indicators with patient- and caregiver-reported outcomes, the study demonstrated that predictive models can achieve robust risk stratification while remaining interpretable. The integration of survival models and classification approaches, complemented by SHAP-based explanations, underscored the value of incorporating social determinants into predictive frameworks. This contribution illustrates the role of explainable ML in managing highly complex patient groups where clinical and social vulnerabilities intersect.

The third case study, reported in Chapter 5, focused on predicting outcomes and recovery trajectories in eating disorders using validated questionnaires. While questionnaires are traditionally employed as screening tools, this dissertation showed that they can also serve as predictors of treatment adherence and long-term recovery. Machine learning models consistently outperformed traditional regression approaches in predicting risk and recovery trajectories. Feature attribution analyses revealed resilience, self-acceptance, and quality of life as central determinants of recovery, thereby underscoring the importance of psychosocial resources in addition to symptom severity. This reframes the role of standardised instruments in clinical

psychology from mere assessment to tools for personalised prognosis.

The fourth case study, presented in Chapter 6, addressed creativity assessment through multimodal models that integrate sketches with textual descriptions. This study demonstrated that multimodal architectures outperform unimodal baselines, particularly in capturing abstract constructs such as originality and flexibility. Importantly, the models operationalised key dimensions of creativity theory (fluency, flexibility, originality) through computational metrics, thereby extending psychometric frameworks into algorithmic domains. This contribution highlights the potential of multimodal AI to complement and enrich psychological assessment methods.

Taken together, the case studies confirm the dissertation's central hypothesis. Machine learning models, when paired with explainability techniques, not only outperform traditional statistical methods in predictive accuracy but also provide understandable basis that align with domain expertise. Across domains as diverse as cardiology, multimorbidity, eating disorders, and creativity, the combination of predictive power and transparency proved essential for practical applicability.

Beyond the individual results, several broader insights emerge. First, the added value of machine learning does not lie solely in incremental improvements in accuracy but in its capacity to integrate heterogeneous data and generate explanations that support trust and decision-making. Second, the extension of standard instruments, such as creativity tests or clinical questionnaires, into predictive tools demonstrates how existing resources can be re-purposed for personalised and scalable assessments. Third, the iterative methodological pipeline illustrates how systematic refinement, validation, and clinician-in-the-loop feedback can contribute to the applicability of prediction models.

Ethical considerations are central to this research. The thesis contributes to more informed and accountable decision-making by promoting transparency through explainable machine learning models. By providing interpretable predictions, these systems can enhance professional judgment rather than replace it, supporting fairer and more evidence-based decisions in clinical and psychological contexts. However, the work also recognises that predictive models are inherently constrained by the data on which they are trained. Biases related to demographic imbalance, cultural representation, or measurement design may propagate through the modelling process, potentially influencing the fairness of predictions. Future applications must therefore incorporate systematic bias assessment, representative data collection, and stakeholder engagement to ensure that explainable artificial intelligence remains both trustworthy and equitable.

At the same time, the research highlights unresolved challenges. Data scarcity and demographic bias limit generalisability. Post-hoc explanation methods, while useful, often lack stability and usability for end-users. Finally, the translation of predictive models into clinical workflows remains an open challenge, requiring interdisciplinary collaboration and attention to ethical, clinical, and social considerations.

8.2 Contributions

A summary of the contributions explained in this dissertation is presented in this section. They are grouped into two categories: (i) application-driven contributions, which focus on the four use cases explored in this dissertation, and (ii) methodological contributions, which represent cross-cutting advances in multimodal, explainable machine learning.

8.2.1 Application-Driven Contributions

This group of contributions relates to the practical advances obtained in the four case studies analysed in this dissertation: prediction of hospital readmissions in heart failure, prediction of hospital readmissions in patients with multiple chronic conditions, eating disorder risk and recovery modelling, and creativity assessment. They represent the clinical and applied dimension of the research, showing how machine learning models can achieve higher predictive accuracy than traditional approaches, while also providing insights of direct value to domain experts. These contributions demonstrate that explainable machine learning can help address real-world challenges in healthcare and psychology.

- **Heart failure readmission prediction (Chapter 3).** This contribution addresses objective 1 and part of objective 2.
  - Developed ensemble models that integrated clinical and patient-reported variables, combining decision trees, Gaussian Naïve Bayes, and neural networks.
  - Achieved superior performance (AUC = 0.81) compared to Cox regression (0.58) and logistic regression (0.50).
  - SHAP analysis revealed frailty, anxiety, and depression as key predictors of readmission risk.
  - Clinician-guided feature selection reduced 96% of the variables while still retaining strong predictive ability (AUC = 0.70), improving practical applicability.

- **Prediction of readmission in MCC (Chapter 4).** This contribution addresses objective 1 and part of objective 2.
  - Designed predictive models for pluripathological patients using the RePluris dataset, integrating clinical, functional, and social determinants of health.
  - Evaluated both survival analysis methods (Cox, Random Survival Forest, DeepSurv) and classification models, showing that ML frameworks outperformed traditional baselines while maintaining interpretability.
  - Demonstrated that patient-reported and caregiver-reported outcomes, alongside social factors, significantly improved the prediction of 30-day readmission.
  - Provided a direct comparison with the LACE index, widely used in clinical practice as a readmission risk score. The models developed in this dissertation consistently outperformed LACE, thereby establishing their superiority as tools for anticipatory care planning in multimorbidity contexts.
  - Demonstrated that explainable ML can provide actionable insights in the management of multimorbidity, where medical and social vulnerabilities intersect.
- **Prediction of eating disorder outcomes and recovery (Chapter 5).** This contribution addresses objective 1 and part of objective 2.
  - Designed predictive models of ED risk and recovery trajectories based on validated questionnaires.
  - Showed that machine learning approaches outperform traditional regression methods in both risk classification and recovery prediction.
  - Feature importance analysis highlighted resilience, self-acceptance, and quality of life as the strongest predictors of long-term recovery.
- **Creativity assessment with multimodal machine learning (Chapter 4).** This contribution addresses objectives 1 and 3.
  - Designed multimodal models integrating text embeddings (BETO, FastText, Keras Embedding) with image encoders (InceptionV3, Xception, EfficientNetB0).
  - Text-based models reached a recall = 0.97 for originality, while image models excelled in elaboration prediction.
  - Multimodal fusion balanced performance across metrics, with ROC AUC values up to 0.85.

– Demonstrated that textual self-descriptions provide complementary value for capturing abstract creative constructs.

8.2.2 Methodological Contributions

In addition to case-specific results, the dissertation provides cross-cutting methodological advances. These contributions focus on explainability, reproducibility, and the integration of multimodal data, establishing frameworks that can be applied beyond the three use cases presented here. They highlight how research can move from technical improvements to systematic strategies that ensure interpretability, robustness, and clinical applicability.

- **Integration of explainability methods.** This addresses objective 3. Post-hoc methods were applied across domains to identify global and local feature contributions. Their systematic application provided both domain-relevant insights and a critical evaluation of their reliability, highlighting the need for user-centred and clinician-in-the-loop explainability. In the ReIC and RePluris studies, SHAP-derived feature rankings were reviewed and validated by clinical experts, ensuring that the explanations were both faithful to model behaviour and clinically meaningful. This human-in-the-loop validation confirms that interpretability was not post-hoc but embedded in the model-development cycle. The deliberate selection of SHAP and Permutation Importance was motivated by their model-agnostic nature and their ability to provide consistent explanatory outputs across heterogeneous predictive frameworks. While alternative deep learning-specific approaches were considered, maintaining a unified interpretability strategy was prioritised to support methodological coherence and domain comparability. This design choice reinforces the thesis's commitment to human-centred explainability rather than architecture-specific introspection.

- **Iterative methodological pipeline.** This addresses objective 4. An iterative workflow was designed. The pipeline integrates preprocessing, modelling, explainability, evaluation, and validation, ensuring methodological robustness and reproducibility.

- **Multimodal integration framework.** This framework effectively processes and integrates different types of data, demonstrating that multimodal models outperform unimodal models in complex human-centred learning situations.

- **Context-aware algorithm selection strategy.** Across the four case studies, heterogeneous model families were deliberately evaluated, including regression-based methods, tree-based ensembles, support vector machines, and deep neural networks. This diversity was not exploratory in nature but

grounded in the specific characteristics of each prediction task and dataset. By analysing performance, calibration, robustness to imbalance, and explainability across model types, the dissertation establishes a structured decision framework for selecting appropriate algorithms in human-centred contexts. This contribution reinforces the thesis that methodological coherence does not require algorithmic uniformity, but rather principled and transparent justification of modelling choices.

- **Critical assessment of questionnaires in predictive modelling.** This thesis provides empirical evidence on questionnaire-based prediction in eating disorders, demonstrating both their usefulness as scalable instruments and their predictive power, as well as the current limitations of bias and generalisation.

Taken together, these methodological elements constitute a generalisable framework for explainable and multimodal machine learning in human-centred contexts. The contribution of this dissertation is therefore not limited to improvements within specific prediction tasks, but extends to the formalisation of a design philosophy in which interpretability, robustness, and contextual adaptability are treated as primary objectives rather than secondary considerations. This positioning situates the work at the intersection of applied machine learning and responsible AI research.

8.3 Hypothesis and Objectives Validation

In the beginning of this dissertation, concretely in Section 1.2, a hypothesis was posed, which stated the following:

> **Hypothesis.** *Machine learning models that incorporate explicit explainability strategies can improve decision-making in complex human-centred contexts by providing both accurate predictions and understandable basis.*

In order to be able to validate this hypothesis, a goal was also defined, which is also shown below for convenience:

> **Goal.** *To design and implement explainable and multimodal machine learning systems that predict relevant events in health and behavioural domains, evaluate their predictive performance, and demonstrate their interpretability and applicability in real-world contexts.*

To achieve such a goal, some specific and measurable objectives were identified in Section 1.2. It is the moment now to go through all those objectives and see how they have been addressed in this dissertation.

1. *To develop machine learning models for representative use cases.* This objective was fulfilled across four domains. Chapter 3 presented predictive models for heart failure readmissions, integrating clinical and psychosocial variables. Chapter 4 addressed hospital readmission prediction in patients with multiple chronic conditions, combining clinical, functional, and social determinants of health. Chapter 5 introduced predictive models for eating disorder outcomes and recovery trajectories based on validated questionnaires. Finally, Chapter 6 designed multimodal models for creativity assessment, integrating image and text modalities to evaluate originality, elaboration, flexibility, and title creativity.

2. *To analyse the performance of the developed models by comparing them with traditional approaches.* Each use case incorporated comparative experiments, with results presented in Chapter 7. In heart failure, ensemble models outperformed Cox and logistic regression. In multiple chronic conditions, machine learning models surpassed the LACE index, a widely used clinical baseline. In eating disorders, naïve Bayes and support vector regression achieved superior predictive performance compared to regression-based approaches. In creativity assessment, multimodal architectures consistently outperformed unimodal baselines. These comparisons demonstrate that while traditional methods remain informative, machine learning frameworks provide clear advantages in complex, heterogeneous domains.

3. *To integrate explainability techniques into the models to provide human-understandable explanations of predictions.* This objective was achieved in all case studies. In Chapter 3, SHAP analysis was applied to interpret hospital readmission predictors, highlighting frailty, anxiety, and depression. In Chapter 4, explanations revealed the importance of functional dependency, caregiver burden, and social support. In Chapter 5, feature attribution analyses identified resilience, self-acceptance, and quality of life as central predictors of recovery. In Chapter 6, explainability analyses clarified the complementary contributions of textual and visual features in creativity evaluation. These examples underscore the potential and the limitations of current post-hoc methods, reinforcing the importance of user-centred explainability.

4. *To assess the applicability of the models within their respective domains.* Applicability was systematically evaluated in Chapter 7. In creativity, multimodal models demonstrated potential for use in educational and psychological contexts. In eating disorders, validated questionnaires were shown to be repurposable as predictive tools, though challenges remain in generalisability and sample diversity. In heart failure, clinician-guided feature reduction produced parsimonious models suitable for integration into clinical workflows. In multiple chronic conditions, the inclusion of social determinants of

health improved predictive accuracy while retaining interpretability, thereby enhancing the potential for deployment in anticipatory care planning.

For greater clarity and transparency, Figure 8.1 maps objectives to contributions, providing a visual summary of the dissertation's validation.

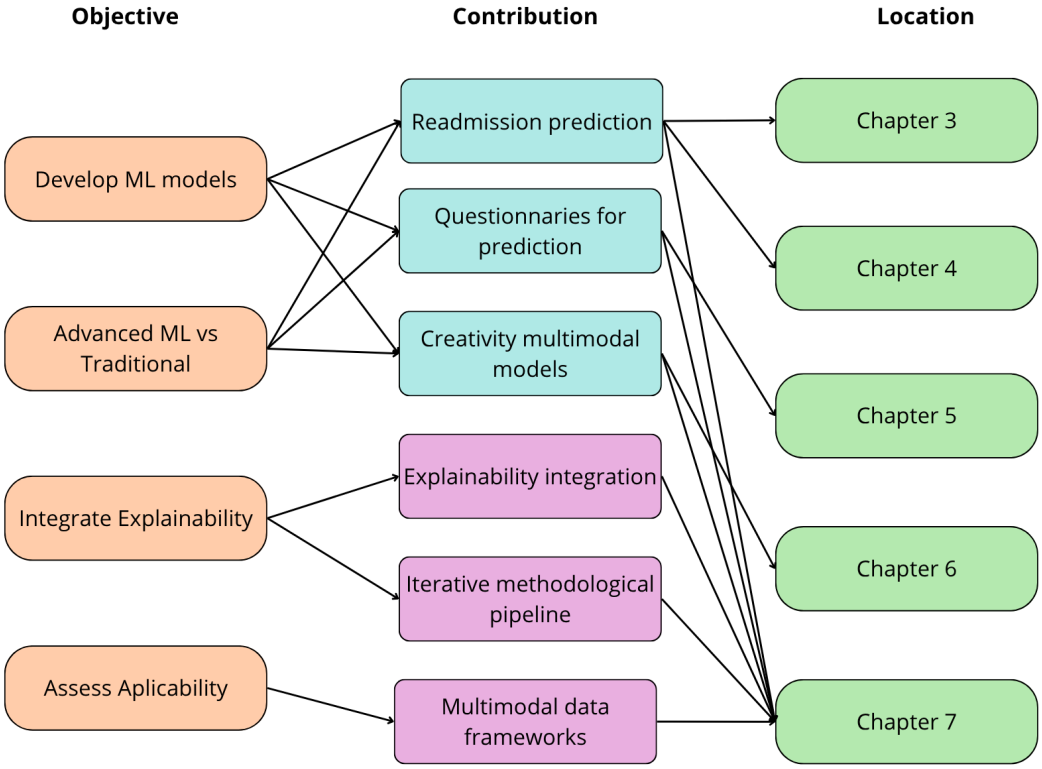


Figure 8.1: The relation of the objectives defined, the contributions and their location in this dissertation.

It can be claimed that all the objectives set in Chapter 1 Section 1.2 have been accomplished in this dissertation. As a consequence, the overall goal of the work has been achieved. The representative use cases developed in this thesis, prediction of heart failure readmissions (Chapter 3), prediction of readmissions in patients with multiple chronic conditions (Chapter 4), prediction of eating disorder outcomes and recovery trajectories (Chapter 5), and creativity assessment (Chapter 6), have been implemented, tested, and evaluated. Each case study demonstrated that explainable machine learning models improved predictive accuracy and yielded interpretable insights relevant to each domain.

The integration of explanation methods across all studies provided global and local interpretations of model decisions, revealing clinically, socially, and psychologically meaningful predictors. For example, Chapter 7 showed that resilience,

self-acceptance, and quality of life were decisive factors in recovery trajectories; that frailty, anxiety, and depression consistently shaped readmission risk in heart failure; that functional dependency and caregiver burden strongly influenced outcomes in multimorbidity; and that multimodal modelling enhanced the assessment of originality and flexibility in creativity research.

Finally, through the accomplishment of the objectives and the overall goal, the hypothesis of this dissertation has been validated: *machine learning models, if designed with strategies for explainability, can improve decision-making in complex human-centred contexts*. The results obtained across Chapters 3-7 prove that it is possible to achieve both predictive performance superior to traditional approaches and interpretable outputs that can be meaningfully integrated into clinical settings. The evaluations presented further confirm that these models are not only technically accurate but also practically applicable, demonstrating the viability of explainable multimodal learning for human-centred assessment.

8.4 Relevant Publications

The contribution of this dissertation has been validated through peer-reviewed publications in journals and conferences:

8.4.1 International JCR Journals

The study on 30-day readmissions in patients with heart problems was published in the following journal:

- Pikatza-Huerga, A., Almeida, A., Quiros, R., Larrea, N., Legarreta, M.J., Zulaika, U., Garcia, R., Garcia, S., Machine learning approaches for predicting heart failure readmissions, Postgraduate Medical Journal, 2025;, qgaf102, `https://doi.org/10.1093/postmj/qgaf102`, JCR Impact Factor (2024): Q1.

The scale for resilience in eating disorders and the research on the use of questionnaires for predicting outcomes in patients with eating disorders was published in the following journals:

- Carlota las Hayas, Odin Hjemdal, Pedro-José Muñoz, Jesús-Ángel Padierna, Luís Beato, Andrés Gómez-del-Barrio, Diana M. Pérez-Valencia, Amaia Pikatza, Aitor Almeida. Development and Validation of the Resilience in Eating Disorders Scale (RED-5). Actas Espanolas de Psiquiatria, 2025. JCR Impact factor (2023): Q4.

- Pikatza-Huerga, A., Las Hayas, C., Zulaika, U., Almeida, A. Predictive assessment of eating disorder risk and recovery: Uncovering the effectiveness of questionnaires and influencing characteristics. Computational and Structural Biotechnology Journal, vol. 28, pages 118–127, 2025. ISSN 2001-0370. JCR Impact factor (2023): Q1.

8.4.2 International Conferences

The multimodal model for assessing human creativity was presented at the following international conference:

- Pikatza-Huerga, A.; Matanzas de Luis, P.; Fernandez-de-Retana Uribe, M.; Lasa, J. P.; Zulaika, U. and Almeida, A. (2025). Analysing the Impact of Images and Text for Predicting Human Creativity Through Encoders. In Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health, ISBN 978-989-758-743-6, ISSN 2184-4984, pages 15-24. DOI: 10.5220/0013203600003938.

These publications confirm both the novelty of the contributions and their acceptance by the research community.

8.5 Future Work

While this dissertation has advanced interpretable and multimodal machine learning for healthcare and psychology, several avenues remain open for future exploration. These can be grouped into three strategic themes: data and validation, methodological advances, and translation into practice.

8.5.1 Data and Validation

A persistent limitation across domains is the scarcity of large, representative, and standardised datasets. Future work should prioritise the collection and integration of data from multiple centres, languages, and populations to ensure fair and generalisable models. This includes extending research beyond Western and predominantly female cohorts, particularly in eating disorder prediction, and addressing under-represented regions such as Africa and the Middle East in heart failure studies. Establishing benchmark datasets for creativity assessment, incorporating both visual and textual artefacts, would provide a valuable resource for reproducibility and comparison across methods.

8.5.2 Methodological Advances

Future research should also explore new methodological directions. In particular, inherently interpretable models, such as prototype-based learning or symbolic approaches, could complement post-hoc explanation frameworks and strengthen trust. Advances in multimodal fusion strategies are equally critical, with hybrid and sequential pipelines offering opportunities to capture temporal and cross-modal dynamics more effectively. Moreover, embedding explainability directly into predictive pipelines, rather than as an add-on, remains an important challenge. The extension of multimodal approaches to domains beyond creativity, such as clinical psychometrics or cognitive assessment in XR environments, also offers promising directions.

8.5.3 Translation into Practice

Finally, the transition from research to real-world application requires systematic attention to usability, ethics, and workflow integration. Co-design studies with clinicians, psychologists, and educators are essential to ensure that models are not only technically sound but also practically relevant. Regulatory and ethical frameworks must be refined to address issues of data privacy, accountability, and bias in multimodal models. Implementation studies, embedding predictive tools into clinical or educational platforms and assessing their impact on decision-making, are crucial for translating methodological innovation into tangible benefits for patients, learners, and practitioners.

8.5.4 Closing Perspective

Together, these directions outline a roadmap toward the next generation of human-centred machine learning: models that are rigorously validated, inherently interpretable, seamlessly multimodal, and responsibly integrated into real-world settings. Pursuing these lines of work will extend the contributions of this dissertation and bring closer the goal of transparent, trustworthy, and actionable machine learning systems in healthcare and psychology.

8.6 Final Remarks

This dissertation has addressed the pressing need for machine learning systems in healthcare and psychology that are simultaneously validated, interpretable, and multimodal. Beginning from the observation that existing models often suffer from poor generalisability, limited interpretability, and narrow reliance on unimodal data, the work has proposed novel frameworks that confront these challenges directly.

The contributions can be summarised in four dimensions. First, in cardiovascular health, the integration of psychosocial and clinical variables in heart failure readmission models demonstrated that explainable ensemble methods outperform traditional approaches while preserving practical applicability through clinician-guided feature reduction. Second, in multimorbidity, the RePluris case study showed that machine learning models can meaningfully incorporate social and functional determinants of health, achieving predictive performance superior to the widely used LACE index while remaining interpretable for clinicians and policymakers. Third, in clinical psychology, the use of psychometric-only models demonstrated that validated questionnaires can be repurposed as predictive instruments, providing clinically relevant insights into risk and recovery trajectories without the need for invasive data collection. Fourth, in the domain of creativity, multimodal frameworks integrating images and text established a foundation for scalable, objective, and theoretically grounded assessment in a domain where traditional methods are subjective and resource-intensive.

Across these domains, the dissertation has shown that the integration of explainability into the predictive process enhances trust and usability, aligning algorithmic outputs with the reasoning practices of clinicians, psychologists, and educators. It has further demonstrated that multimodal integration, whether of biomedical, psychosocial, or perceptual data, enables richer and more accurate modelling of complex human-centred constructs.
By connecting methodological innovation with applied impact, this dissertation contributes to the broader aim of designing machine learning systems that are not only technically accurate but also ethically, clinically, and socially meaningful. The findings respond directly to the research gaps identified in Chapter 2, demonstrating how advances in validation, interpretability, and multimodality can be combined into coherent, human-centred approaches.

In closing, the work positions itself as a step toward the next generation of machine learning in healthcare and psychology: systems that integrate heterogeneous data, respect psychometric theory, and provide explanations that clinicians and patients alike can understand and trust. While many challenges remain, the frameworks developed here illustrate how machine learning can move beyond proof-of-concept studies to deliver transparent and actionable insights that support human decision-making.

In this sense, the dissertation contributes at two complementary levels. At the applied level, it delivers validated predictive systems across cardiovascular health, multimorbidity, eating disorders, and creativity assessment. At the conceptual level, it demonstrates how heterogeneous algorithms, multimodal data sources, and explainability strategies can be integrated into a unified human-centred modelling

paradigm. This dual contribution reinforces the general thesis that machine learning systems can be both technically advanced and epistemologically aligned with domain expertise.

The dissertation thus contributes not only to applied case studies but also to the methodological foundations of human-centred machine learning, offering lessons for future research across domains. Ultimately, this work demonstrates how human-centred AI can evolve from research prototypes into tools that meaningfully enhance decision-making and patient understanding.

Bibliography

Abdullah, T. A. A., Zahid, M. S. M. y Ali, W. A review of interpretable ml in healthcare: Taxonomy, applications, challenges, and future directions. *Symmetry*, vol. 13(12), 2021. ISSN 2073-8994. (ref. pág. 5)

Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F. y Alotaiq, N. The use of machine learning for analyzing real-world data in disease prediction and management: Systematic review. *JMIR Med Inform*, vol. 13, página e68898, 2025. ISSN 2291-9694. (ref. pág. 16)

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S. y Albekairy, A. M. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, vol. 23(1), 2023. ISSN 1472-6920. (ref. pág. 16)

Alvanpour, A., Acun, C., Spurlock, K., Robinson, C. K., Das, S. K., Popa, D. O. y Nasraoui, O. Comparative analysis of post hoc explainable methods for robotic grasp failure prediction. *Electronics*, vol. 14(9), página 1868, 2025. ISSN 2079-9292. (ref. pág. 33, 34)

Arjunan, G. Implementing explainable ai in healthcare: Techniques for interpretable machine learning models in clinical decision-making. *International Journal of Scientific Research and Management (IJSRM)*, vol. 9(05), página 597–603, 2021. ISSN 2321-3418. (ref. pág. 32, 33)

Aryal, Y., Maag, A. y Gunasekera, N. Application of machine learning algorithms in diagnosis and detection of psychological disorders. En *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, página 1–10. IEEE, 2020. (ref. pág. 20)

Babu, A. R., Rajavenkatanarayanan, A., Brady, J. R. y Makedon, F. Multimodal approach for cognitive task performance prediction from body postures, facial expressions and eeg signal. En *Proceedings of the Workshop on Modeling*

Cognitive Processes from Multimodal Data, ICMI '18, página 1–7. ACM, 2018. (ref. pág. 27)

BANAPURAM, C., NAIK, A. C., VANTERU, M. K., V, S. K. y VAIGANDLA, K. K. A comprehensive survey of machine learning in healthcare: Predicting heart and liver disease, tuberculosis detection in chest x-ray images. *International Journal of Electronics and Communication Engineering*, vol. 11(5), página 155–169, 2024. ISSN 2348-8549. (ref. pág. 17, 20)

BARENHOLTZ, E., FITZGERALD, N. D. y HAHN, W. E. Machine-learning approaches to substance-abuse research: emerging trends and their implications. *Current Opinion in Psychiatry*, vol. 33(4), página 334–342, 2020. ISSN 1473-6578. (ref. pág. 21)

BULIK, C. M., COLEMAN, J. R. I., HARDAWAY, J. A., BREITHAUPT, L., WATSON, H. J., BRYANT, C. D. y BREEN, G. Genetics and neurobiology of eating disorders. *Nature Neuroscience*, vol. 25(5), página 543–554, 2022. ISSN 1546-1726. (ref. pág. 119)

BŁAZIAK, M., URBAN, S., WIETRZYK, W., JURA, M., IWANEK, G., STAŃCZYKIEWICZ, B., KULICZKOWSKI, W., ZYMLIŃSKI, R., PONDEL, M., BERKA, P., DANEL, D., BIEGUS, J. y SIENNICKA, A. An artificial intelligence approach to guiding the management of heart failure patients using predictive models: A systematic review. *Biomedicines*, vol. 10(9), página 2188, 2022. ISSN 2227-9059. (ref. pág. 19)

CARVALHO, D. y CRUZ, R. Big data and machine learning in health. *European Journal of Public Health*, vol. 30, página ckaa040.030, 2020. ISSN 1101-1262. (ref. pág. 17)

CHADDAD, A., PENG, J., XU, J. y BOURIDANE, A. Survey of explainable ai techniques in healthcare. *Sensors*, vol. 23(2), página 634, 2023. ISSN 1424-8220. (ref. pág. 32, 33)

CHANDLER, C., FOLTZ, P. W., COHEN, A. S., HOLMLUND, T. B., CHENG, J., BERNSTEIN, J. C., ROSENFELD, E. P. y ELVEVÅG, B. Machine learning for ambulatory applications of neuropsychological testing. *Intelligence-Based Medicine*, vol. 1–2, página 100006, 2020. ISSN 2666-5212. (ref. pág. 21)

CHAPA, D. A. N., HAGAN, K. E., FORBUSH, K. T., CLARK, K. E., TREGARTEN, J. P. y ARGUE, S. Longitudinal trajectories of behavior change in a national sample of patients seeking eating-disorder treatment. *The International journal of eating disorders*, 2020. (ref. pág. 22, 23)

CHEKROUD, A. M., BONDAR, J., DELGADILLO, J., DOHERTY, G., WASIL, A., FOKKEMA, M., COHEN, Z., BELGRAVE, D., DERUBEIS, R., INIESTA, R., DWYER, D. y CHOI, K. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, vol. 20(2), página 154–170, 2021. ISSN 2051-5545. (ref. pág. 21, 22)

CHEN, H., LUNDBERG, S. M. y LEE, S.-I. Explaining a series of models by propagating shapley values. *Nature Communications*, vol. 13(1), 2022. ISSN 2041-1723. (ref. pág. 34, 35)

CHEN, Y.-M., HSIAO, T.-H., LIN, C.-H. y FANN, Y. C. Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. *Journal of Biomedical Science*, vol. 32(1), 2025. ISSN 1423-0127. (ref. pág. 1, 2)

COHN, C., DAVALOS, E., VATRAL, C., FONTELES, J. H., WANG, H. D., MA, M. y BISWAS, G. Multimodal methods for analyzing learning and training environments: A systematic literature review. 2024. (ref. pág. 27, 28)

CROPLEY, D. H., THEURER, C., MATHIJSSEN, A. C. S. y MARRONE, R. L. Fit-for-purpose creativity assessment: Automatic scoring of the test of creative thinking – drawing production (tct-dp). *Creativity Research Journal*, página 1–16, 2024. ISSN 1532-6934. (ref. pág. 29)

DARA, S., DHAMERCHERLA, S., JADAV, S. S., BABU, C. M. y AHSAN, M. J. Machine learning in drug discovery: A review. *Artificial Intelligence Review*, vol. 55(3), página 1947–1999, 2021. ISSN 1573-7462. (ref. pág. 16)

DHUDUM, R., GANESHPURKAR, A. y PAWAR, A. Revolutionizing drug discovery: A comprehensive review of ai applications. *Drugs and Drug Candidates*, vol. 3(1), páginas 148–171, 2024. ISSN 2813-2998. (ref. pág. 16)

DONG, C., JI, Y., FU, Z., QI, Y., YI, T., YANG, Y., SUN, Y. y SUN, H. Precision management in chronic disease: An ai empowered perspective on medicine-engineering crossover. *iScience*, vol. 28(3), página 112044, 2025. ISSN 2589-0042. (ref. pág. 2)

AAFJES-VAN DOORN, K., KAMSTEEG, C., BATE, J. y AAFJES, M. A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, vol. 31(1), página 92–116, 2020. ISSN 1468-4381. (ref. pág. 21)

DOW, G. T. Defining creativity. *Creativity and Innovation Theory, Research, and Practice*, 2022. (ref. pág. 85)

DYKXHOORN, J. y KIRKBRIDE, J. B. The epidemiological burden of major psychiatric disorders. *The Oxford textbook of public mental health*, páginas 73–84, 2018. (ref. pág. 1)

ESPEL-HUYNH, H. M., ZHANG, F., BOSWELL, J. F., THOMAS, J. G., THOMPSON-BRENNER, H., JUARASCIO, A. S. y LOWE, M. R. Latent trajectories of eating disorder treatment response among female patients in residential care. *Int J Eat Disord*, vol. 53(10), páginas 1647–1656, 2020. (ref. pág. 22, 23)

ESPEL-HUYNH, H., ZHANG, F., THOMAS, J. G., BOSWELL, J. F., THOMPSON-BRENNER, H., JUARASCIO, A. S. y LOWE, M. R. Prediction of eating disorder treatment response trajectories via machine learning does not improve performance versus a simpler regression approach. *International Journal of Eating Disorders*, vol. 54(7), página 1250–1259, 2021. ISSN 1098-108X. (ref. pág. 22)

FAHIM, Y. A., HASANI, I. W., KABBA, S. y RAGAB, W. M. Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. *European Journal of Medical Research*, vol. 30(1), 2025. ISSN 2047-783X. (ref. pág. 1, 2)

FALVO, F. R. y CANNATARO, M. Explainability techniques for artificial intelligence models in medical diagnostic. En *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, página 6907–6913. IEEE, 2024. (ref. pág. 32)

FARDOULY, J., CROSBY, R. D. y SUKUNESAN, S. Potential benefits and limitations of machine learning in the field of eating disorders: current research and future directions. *Journal of Eating Disorders*, vol. 10(1), 2022. ISSN 2050-2974. (ref. pág. 20, 21)

GARCIA-GUTIERREZ, S. Re-pluris project: Identification of multimorbid patients with a higher probability of readmission and mortality. 2025. (ref. pág. 56)

GARCIA-GUTIERREZ, S., VILLANUEVA, A., LAFUENTE, I., RODRIGUEZ, I., LOZANO-BAHAMONDE, A., MURGA, N., ORUS, J., CAMACHO, E. R., QUINTANA, J. M., QUIROS, R. y REIC-REDISSEC WORKING GROUP. Factors related to early readmissions after acute heart failure: a nested case-control study. *BMC Cardiovasc Disord*, vol. 23(1), página 17, 2023. (ref. pág. 40, 48)

GASPAR, D., SILVA, P. y SILVA, C. Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron. *IEEE Access*, vol. 12, página 30164–30175, 2024. ISSN 2169-3536. (ref. pág. 34)

GHASSEMI, M., NAUMANN, T., SCHULAM, P. F., BEAM, A. L., CHEN, I. Y. y RANGANATH, R. A review of challenges and opportunities in machine learning for

health. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2020, páginas 191–200, 2018. (ref. pág. 19)

GONZALEZ-ERENA, P. V., FERNANDEZ-GUINEA, S. y KOURTESIS, P. Cognitive assessment and training in extended reality: Multimodal systems, clinical utility, and current challenges. *Encyclopedia*, vol. 5(1), página 8, 2025. ISSN 2673-8392. (ref. pág. 27)

GORRELL, S., HAIL, L. y REILLY, E. E. Predictors of treatment outcome in eating disorders: A roadmap to inform future research efforts. *Current Psychiatry Reports*, vol. 25, páginas 213–222, 2023. (ref. pág. 22)

GUERRERO-SOSA, J. D. T., ROMERO, F. P., MENENDEZ-DOMINGUEZ, V. H., SERRANO-GUERRERO, J., MONTORO-MONTARROSO, A. y OLIVAS, J. A. A comprehensive review of multimodal analysis in education. *Applied Sciences*, vol. 15(11), página 5896, 2025. ISSN 2076-3417. (ref. pág. 27)

GUO, A., PASQUE, M., LOH, F., MANN, D. L. y PAYNE, P. R. O. Heart failure diagnosis, readmission, and mortality prediction using machine learning and artificial intelligence models. *Current Epidemiology Reports*, vol. 7(4), página 212–219, 2020. ISSN 2196-2995. (ref. pág. 18)

HAJISHAH, H., KAZEMI, D., SAFAEE, E., AMINI, M. J., PEISEPAR, M., TANHAPOUR, M. M. y TAVASOL, A. Evaluation of machine learning methods for prediction of heart failure mortality and readmission: meta-analysis. *BMC Cardiovascular Disorders*, vol. 25(1), 2025. ISSN 1471-2261. (ref. pág. 18)

HAYNOS, A. F., WANG, S. B., LIPSON, S., PETERSON, C. B., MITCHELL, J. E., HALMI, K. A., AGRAS, W. S. y CROW, S. J. Machine learning enhances prediction of illness course: a longitudinal study in eating disorders. *Psychological Medicine*, vol. 51(8), página 1392–1402, 2020. ISSN 1469-8978. (ref. pág. 22, 23)

HIDAYATURROHMAN, Q. A. y HANADA, E. Predictive analytics in heart failure risk, readmission, and mortality prediction: A review. *Cureus*, 2024. ISSN 2168-8184. (ref. pág. 18, 19)

HILDT, E. What is the role of explainability in medical artificial intelligence? a case-based approach. *Bioengineering*, vol. 12(4), página 375, 2025. ISSN 2306-5354. (ref. pág. 2, 3)

HOBBACH, A. J., FELD, J., LINKE, W. A., SINDERMANN, J. R., DRÖGE, P., RUHNKE, T., GÜNSTER, C. y REINECKE, H. Bmi-stratified exploration of the 'obesity paradox': Heart failure perspectives from a large german insurance database. *Journal of Clinical Medicine*, vol. 13(7), página 2086, 2024. ISSN 2077-0383. (ref. pág. 110)

HOOSHYAR, D. y YANG, Y. Problems with shap and lime in interpretable ai for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*, vol. 12, página 137472–137490, 2024. ISSN 2169-3536. (ref. pág. 34)

HOUSE, E. T., LISTER, N. B., SEIDLER, A. L., LI, H., ONG, W. Y., MCMASTER, C. M., PAXTON, S. J. y JEBEILE, H. Identifying eating disorders in adolescents and adults with overweight or obesity: A systematic review of screening questionnaires. *International Journal of Eating Disorders*, vol. 55(9), página 1171–1193, 2022. ISSN 1098-108X. (ref. pág. 24)

HWANG, M., ZHENG, Y., CHO, Y. y JIANG, Y. Ai applications for chronic condition self-management: Scoping review. *J Med Internet Res*, vol. 27, página e59632, 2025. ISSN 1438-8871. (ref. pág. 2)

JOYCE, D. W., KORMILITZIN, A., SMITH, K. A. y CIPRIANI, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, vol. 6(1), 2023. ISSN 2398-6352. (ref. pág. 5)

KANT, S., DEEPIKA y ROY, S. Artificial intelligence in drug discovery and development: transforming challenges into opportunities. *Discover Pharmaceutical Sciences*, vol. 1(1), 2025. ISSN 3005-1835. (ref. pág. 16)

KARAKO, K. y TANG, W. Applications of and issues with machine learning in medicine: Bridging the gap with explainable ai. *BioScience Trends*, vol. 18(6), página 497–504, 2024. ISSN 1881-7823. (ref. pág. 32)

KHAN, S. M. *Multimodal Behavioral Analytics in Intelligent Learning and Assessment Systems*, página 173–184. Springer International Publishing, 2017. ISBN 9783319332611. (ref. pág. 28)

KLINE, A., WANG, H., LI, Y., DENNIS, S., HUTCH, M., XU, Z., WANG, F., CHENG, F. y LUO, Y. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, vol. 5(1), 2022. ISSN 2398-6352. (ref. pág. 18)

KOVALKOV, A., PAASEN, B., SEGAL, A., PINKWART, N. y GAL, K. Automatic creativity measurement in scratch programs across modalities. *IEEE Transactions on Learning Technologies*, vol. 14(6), página 740–753, 2021. ISSN 2372-0050. (ref. pág. 29)

KREUZBERGER, D., KÜHL, N. y HIRSCHL, S. Machine learning operations (mlops): Overview, definition, and architecture. 2022. (ref. pág. 8)

KRONES, F., MARIKKAR, U., PARSONS, G., SZMUL, A. y MAHDI, A. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, vol. 114, página 102690, 2025. ISSN 1566-2535. (ref. pág. 17)

KRUG, I., LINARDON, J., GREENWOOD, C., YOUSSEF, G., TREASURE, J., FERNANDEZ-ARANDA, F., KARWAUTZ, A., WAGNER, G., COLLIER, D., ANDERLUH, M., TCHANTURIA, K., RICCA, V., SORBI, S., NACMIAS, B., BELLODI, L. y FULLER-TYSZKIEWICZ, M. A proof-of-concept study applying machine learning methods to putative risk factors for eating disorders: results from the multi-centre european project on healthy eating. *Psychological Medicine*, vol. 53(7), página 2913–2922, 2021. ISSN 1469-8978. (ref. pág. 25)

KRZESIŃSKI, P., SIEBERT, J., JANKOWSKA, E. A., GALAS, A., PIOTROWICZ, K., STAŃCZYK, A., SIWOŁOWSKI, P., GUTKNECHT, P., CHROM, P., MURAWSKI, P., WALCZAK, A., SZALEWSKA, D., BANASIAK, W., PONIKOWSKI, P. y GIELERAK, G. Nurse-led ambulatory care supported by non-invasive haemodynamic assessment after acute heart failure decompensation. *ESC Heart Failure*, vol. 8(2), página 1018–1026, 2021. ISSN 2055-5822. (ref. pág. 110)

KUMAR, A., VEERAIAH, V., GONGADA, T. N., AHAMAD, S., KHAN, H. y GUPTA, A. Explainable machine learning models for clinical decision support systems. En *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, página 1–6. IEEE, 2024. (ref. pág. 32, 33)

LAS-HAYAS, C., HJEMDAL, O., MUNOZ, P.-J., PADIERNA, J.-A., BEATO, L. y GOMEZ-DEL BARRIO, A. Resilience in eating disorders (red-5) scale. 2025. (ref. pág. 70)

LEE, E. E., TOROUS, J., DE CHOUDHURY, M., DEPP, C. A., GRAHAM, S. A., KIM, H.-C., PAULUS, M. P., KRYSTAL, J. H. y JESTE, D. V. Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 6(9), página 856–864, 2021. ISSN 2451-9022. (ref. pág. 2, 5, 6)

LI, Y.-H., LI, Y.-L., WEI, M.-Y. y LI, G.-Y. Innovation and challenges of artificial intelligence technology in personalized healthcare. *Scientific Reports*, vol. 14(1), 2024. ISSN 2045-2322. (ref. pág. 1)

LINARDON, J. Positive body image, intuitive eating, and self-compassion protect against the onset of the core symptoms of eating disorders: A prospective study. *International Journal of Eating Disorders*, vol. 54(11), página 1967–1977, 2021. ISSN 1098-108X. (ref. pág. 126)

LINARDON, J., FULLER-TYSZKIEWICZ, M., SHATTE, A. y GREENWOOD, C. J. An exploratory application of machine learning methods to optimize prediction of

responsiveness to digital interventions for eating disorder symptoms. *International Journal of Eating Disorders*, vol. 55(6), página 845–850, 2022. ISSN 1098-108X. (ref. pág. 24)

LIU, Y., LIU, C., ZHENG, J., XU, C. y WANG, D. Improving explainability and integrability of medical ai to promote health care professional acceptance and use: Mixed systematic review. *J Med Internet Res*, vol. 27, página e73374, 2025. ISSN 1438-8871. (ref. pág. 2)

MACEACHERN, S. J. y FORKERT, N. D. Machine learning for precision medicine. *Genome*, vol. 64(4), página 416–425, 2021. ISSN 1480-3321. (ref. pág. 3)

MIOTTO, R., WANG, F., WANG, S., JIANG, X. y DUDLEY, J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, vol. 19(6), página 1236–1246, 2017. ISSN 1477-4054. (ref. pág. 16)

MPANYA, D., CELIK, T., KLUG, E. y NTSINJANA, H. Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review. *IJC Heart & Vasculature*, vol. 34, página 100773, 2021. ISSN 2352-9067. (ref. pág. 18, 19)

NAJJAR, R. Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*, vol. 13(17), página 2760, 2023. ISSN 2075-4418. (ref. pág. 16)

NARIMANI, M. y NAEIM, M. Artificial intelligence in mental health: integrating opportunities and challenges of multimodal deep learning for mental disorder prevention and treatment. *Annals of Medicine & Surgery*, vol. 87(9), página 5757–5761, 2025. ISSN 2049-0801. (ref. pág. 2, 3)

NIEMANN, M., PRANGE, A. y SONNTAG, D. Towards a multimodal multisensory cognitive assessment framework. En *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, página 24–29. IEEE, 2018. (ref. pág. 28)

OCANA, A., PANDIELLA, A., PRIVAT, C., BRAVO, I., LUENGO-OROZ, M., AMIR, E. y GYORFFY, B. Integrating artificial intelligence in drug discovery and early drug development: a transformative approach. *Biomarker Research*, vol. 13(1), 2025. ISSN 2050-7771. (ref. pág. 16)

OZCELIK, R., VAN TILBORG, D., JIMENEZ-LUNA, J. y GRISONI, F. Structure-based drug discovery with deep learning. 2022. (ref. pág. 16, 17)

PAN, M., LI, R., WEI, J., PENG, H., HU, Z., XIONG, Y., LI, N., GUO, Y., GU, W. y LIU, H. Application of artificial intelligence in the health management of chronic

disease: bibliometric analysis. *Frontiers in Medicine*, vol. Volume 11 - 2024, 2025. ISSN 2296-858X. (ref. pág. 1)

PETCH, J., DI, S. y NELSON, W. Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology*, vol. 38(2), página 204–213, 2022. ISSN 0828-282X. (ref. pág. 32)

PIKATZA-HUERGA, A., ALMEIDA, A., QUIROS, R., LARREA, N., JOSE LEGARRETA, M., ZULAIKA, U., GARCIA, R. y GARCIA, S. Machine learning approaches for predicting heart failure readmissions. *Postgraduate Medical Journal*, página qgaf102, 2025a. ISSN 0032-5473. (ref. pág. 53, 97)

PIKATZA-HUERGA, A., LAS HAYAS, C., ZULAIKA, U. y ALMEIDA, A. Predictive assessment of eating disorder risk and recovery: Uncovering the effectiveness of questionnaires and influencing characteristics. *Computational and Structural Biotechnology Journal*, vol. 28, páginas 118–127, 2025b. ISSN 2001-0370. (ref. pág. 80)

RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J., KRUEGER, G. y SUTSKEVER, I. Learning transferable visual models from natural language supervision. 2021. (ref. pág. xvii, 90)

RAHMAN, E. U., CHOBUFO, M. D., FARAH, F., MOHAMED, T., ELHAMDANI, M., RUEDA, C., ARONOW, W. S., FONAROW, G. C. y THOMPSON, E. Prevalence and temporal trends of anemia in patients with heart failure. *QJM: An International Journal of Medicine*, vol. 115(7), página 437–441, 2021. ISSN 1460-2393. (ref. pág. 107)

RAMANARAYANAN, V. Multimodal technologies for remote assessment of neurological and mental health. *Journal of Speech, Language, and Hearing Research*, vol. 67(11), página 4233–4245, 2024. ISSN 1558-9102. (ref. pág. 28, 30)

RASHEED, K., QAYYUM, A., GHALY, M., AL-FUQAHA, A., RAZI, A. y QADIR, J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Institute of Electrical and Electronics Engineers (IEEE)*, 2021. (ref. pág. 32)

ROSENBACKE, R., MELHUS, A., MCKEE, M. y STUCKLER, D. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: Systematic review. *JMIR AI*, vol. 3, página e53207, 2024. ISSN 2817-1705. (ref. pág. 2, 3)

SALIH, A. M., RAISI-ESTABRAGH, Z., GALAZZO, I. B., RADEVA, P., PETERSEN, S. E., LEKADIR, K. y MENEGAZ, G. A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, vol. 7(1), 2024. ISSN 2640-4567. (ref. pág. 33, 34)

SCHAPIN, N., MAJEWSKI, M., VARELA, A., ARRONIZ, C. y FABRITIIS, G. D. Machine learning small molecule properties in drug discovery. 2023. (ref. pág. 16)

SCHMIDT, U. H., CLAUDINO, A., FERNÁNDEZ-ARANDA, F., GIEL, K. E., GRIFFITHS, J., HAY, P. J., KIM, Y., MARSHALL, J., MICALI, N., MONTELEONE, A. M., NAKAZATO, M., STEINGLASS, J., WADE, T. D., WONDERLICH, S., ZIPFEL, S., ALLEN, K. L. y SHARPE, H. The current clinical approach to feeding and eating disorders aimed to increase personalization of management. *World Psychiatry*, vol. 24(1), página 4–31, 2025. ISSN 2051-5545. (ref. pág. 120)

SCUTT, K., ALI, K., RIEGER, E., MONAGHAN, C., FORD, R., FABRY, E. y FASSNACHT, D. An investigation of the dual continua model of mental health in the context of eating disorder symptomatology using latent profile analysis. *British Journal of Clinical Psychology*, vol. 62(4), página 782–799, 2023. ISSN 2044-8260. (ref. pág. 128)

SIDEY-GIBBONS, J. A. M. y SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, vol. 19(1), 2019. ISSN 1471-2288. (ref. pág. 4, 6)

SONG, B., MILLER, S. y AHMED, F. Attention-enhanced multimodal learning for conceptual design evaluations. *Journal of Mechanical Design*, vol. 145(4), 2023. ISSN 1528-9001. (ref. pág. 29)

SPEE, B. T. M., LEDER, H., MIKUNI, J., SCHARNOWSKI, F., PELOWSKI, M. y STEYRL, D. Using machine learning to predict judgments on western visual art along content-representational and formal-perceptual attributes. *PLOS ONE*, vol. 19(9), página e0304285, 2024. ISSN 1932-6203. (ref. pág. 5)

SPEE, B. T. M., MIKUNI, J., LEDER, H., SCHARNOWSKI, F., PELOWSKI, M. y STEYRL, D. Machine learning revealed symbolism, emotionality, and imaginativeness as primary predictors of creativity evaluations of western art paintings. *Scientific Reports*, vol. 13(1), 2023. ISSN 2045-2322. (ref. pág. 29)

STANKIĆ, D., MILIĆ, J. y STEFANOVIĆ, D. The possibilities for detection of eating disorders and potential health risks in untreated patients. *Glasnik javnog zdravlja*, vol. 97(2), página 188–199, 2023. ISSN 2812-8842. (ref. pág. 24, 25)

THIEME, A., BELGRAVE, D. y DOHERTY, G. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction*, vol. 27(5), página 1–53, 2020. ISSN 1557-7325. (ref. pág. 21)

TUENA, C., CHIAPPINI, M., REPETTO, C. y RIVA, G. *Artificial Intelligence in Clinical Psychology*, página 10–27. Elsevier, 2022. ISBN 9780128222324. (ref. pág. 20)

VALENTOVA, M., ANKER, S. D. y VON HAEHLING, S. Cardiac cachexia revisited. *Cardiology Clinics*, vol. 40(2), página 199–207, 2022. ISSN 0733-8651. (ref. pág. 110)

VANI, M. S., SUDHAKAR, R. V., MAHENDAR, A., LEDALLA, S., RADHA, M. y SUNITHA, M. Personalized health monitoring using explainable ai: bridging trust in predictive healthcare. *Scientific Reports*, vol. 15(1), 2025. ISSN 2045-2322. (ref. pág. 3, 4, 6)

VIMBI, V., SHAFFI, N. y MAHMUD, M. Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection. *Brain Informatics*, vol. 11(1), 2024. ISSN 2198-4026. (ref. pág. 33, 34, 35)

DE VOS, J. A., RADSTAAK, M., BOHLMEIJER, E. T. y WESTERHOF, G. J. Modelling trajectories of change in psychopathology and well-being during eating disorder outpatient treatment. *Psychotherapy Research*, vol. 33(4), páginas 415–427, 2023. PMID: 36330764. (ref. pág. 22, 23)

YANG, H.-H., LI, F.-R., CHEN, Z.-K., ZHOU, M.-G., XIE, L.-F., JIN, Y., LI, Z. y CHEN, G.-C. Duration of diabetes, glycemic control, and risk of heart failure among adults with diabetes: A cohort study. *The Journal of Clinical Endocrinology & Metabolism*, vol. 108, 2022. (ref. pág. 108)

ZHANG, H., DONG, H., WANG, Y., ZHANG, X., YU, F., REN, B. y XU, J. Automated graphic divergent thinking assessment: A multimodal machine learning approach. *Journal of Intelligence*, vol. 13(4), 2025. ISSN 2079-3200. (ref. pág. 5, 6)

ZHANG, Z., QIAN, M., LUO, L., GAO, Q., WANG, X., SAHA, R. y SONG, X. Using a cnn model to assess paintings' creativity. En *arXiv*. 2024. (ref. pág. 29)

ZHAO, W., QIN, J., LU, G., WANG, Y., QIAO, L. y LI, Y. Association between hyponatremia and adverse clinical outcomes of heart failure: current evidence based on a systematic review and meta-analysis. *Frontiers in Cardiovascular Medicine*, vol. 10, 2023. ISSN 2297-055X. (ref. pág. 107)

ZHUANG, B., SHEN, T., LI, D., JIANG, Y., LI, G., LUO, Q., JIN, Y., SHAN, Z., CHE, L., WANG, L., ZHENG, L. y SHEN, Y. A model for the prediction of mortality and hospitalization in chinese heart failure patients. *Frontiers in Cardiovascular Medicine*, vol. 8, 2021. ISSN 2297-055X. (ref. pág. 19)

APPENDIX

A

Feature importance ranking for ReIC

In the ReIC use case, which focused on predicting 30-day readmissions in patients with heart failure, a feature importance analysis was conducted on the full set of available clinical and psychosocial variables. The ranking was computed using SHAP values derived from the best-performing ensemble model, which allowed the contribution of each predictor to the model's output to be quantified.

Figure A.1 presents the 50 most important features identified in this analysis. These variables represent the predictors with the highest marginal contribution to the discrimination between readmitted and non-readmitted patients. Importantly, this list was not intended to define the final model but rather to serve as the candidate pool for expert evaluation.

From this ranked set, clinicians and domain experts selected a clinically plausible and practically applicable subset of features. Their selection process considered both the relative importance assigned by the model and the interpretability, availability, and routine measurability of the variables in real-world settings. The final subset of predictors thus represents an expert-guided refinement of the algorithmically identified importance ranking, ensuring alignment between predictive modelling and clinical applicability.

This appendix provides the complete list of the 50 most important features as a reference for transparency and reproducibility of the feature selection process.

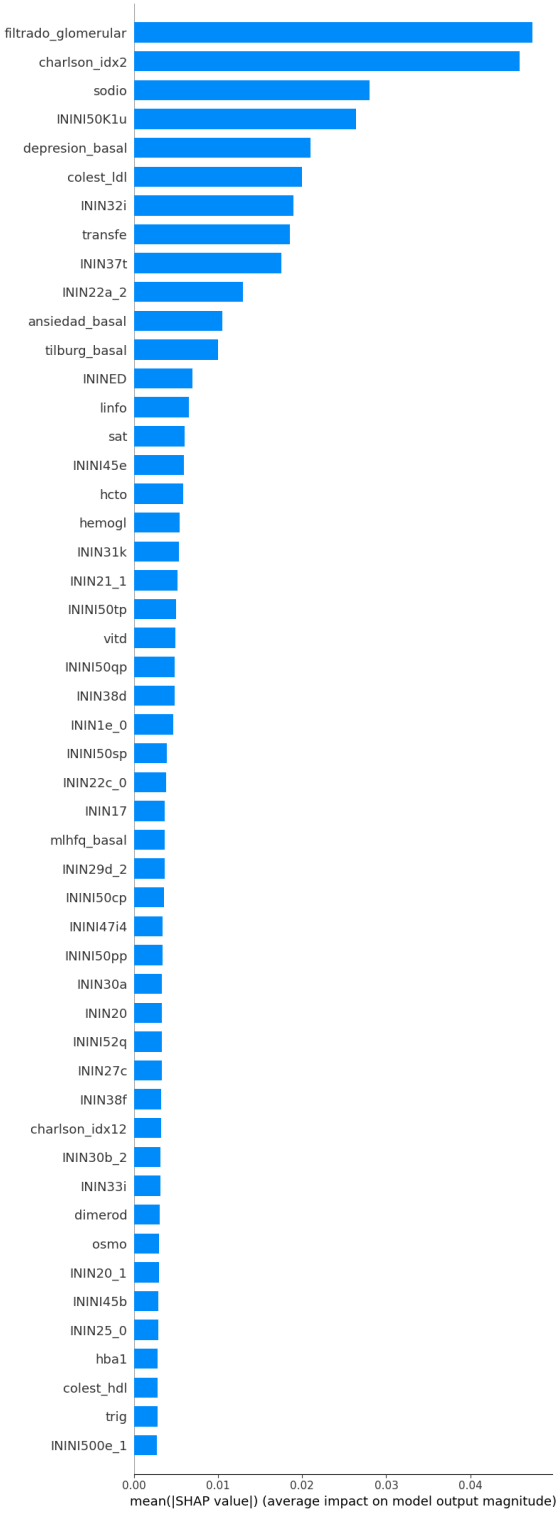


Figure A.1: The 50 most important features of the ReIC best model before clinical guided feature selection.

Table A.1: Set for clinician selection of features.

Variable	Description
filtrado_glomerular	Glomerular filtration rate (mL/min/1.73m^2)
charlson_idx2	Congestive heart failure (CHF)
sodio	Serum sodium concentration (mmol/L)
ININI50K1u	Presence of oedema up to the knee, assessed clinically
depresion_basal	HAD depression subscale score at admission (0–21)
colest_ldl	Low-density lipoprotein cholesterol (mg/dL)
ININ32i	Mild chronic pulmonary disease
transfe	Transferrin concentration (mg/dL)
ININ37t	Use of angiotensin receptor antagonists measured in the emergency department (binary)
ININ22a_2	Simpson ejection fraction
ansiedad_basal	HAD anxiety subscale score at admission (0–21)
tilburg_basal	Tilburg Frailty Indicator score
ININED	Age in years
linfo	Lymphocyte count (10^9/L)
sat	Oxygen saturation
ININI45e	Kidney failure
hcto	Haematocrit
hemogl	Haemoglobin concentration (g/dL)
ININ31k	Treatment of hyperuricaemia (oral iron)
ININ21_1	Last troponin value before admission (ng/L)
ININI50tp	Diastolic blood pressure (mmHg) at discharge
vitd	Serum vitamin D level (ng/mL)
ININI50qp	Oxygen saturation (pulse oximeter) at discharge
ININ38d	Oxygen saturation on admission
ININ1e_0	Altered acid-base type at admission (categorical)
ININI50sp	Systolic blood pressure at discharge (mmHg)
ININ22c_0	Baseline tricuspid regurgitation rate (echocardiography)
ININ17	Body mass index (kg/m^2)
mlhfq_basal	Minnesota Living with Heart Failure Questionnaire score at baseline

Continued on next page

Table 7.6 (continued)

Variable	Description
ININ29d_2	History of mitral stenosis (binary)
ININI50cp	Paroxysmal nocturnal dyspnoea at discharge (binary)
ININI47i4	Urinary Tract Infection
ININI50pp	Body temperature
ININ30a	Diagnostic coronary angiography
ININ20	NT-proBNP test conducted (Yes/No)
ININI52q	Use of calcium antagonists at discharge (binary)
ININ27c	Long-term treatment with beta-blockers at baseline (binary)
ININ38f	Heart rate monitoring
charlson_idx12	Hemiplegia or paraplegia
ININ30b_2	History of drug-eluting stent (binary)
ININ33i	Uncontrolled high blood pressure
dimerod	D-dimer test result (ng/mL)
ININ20_1	Baseline NT-proBNP level (pg/mL)
ININI45b	Pulmonary disease decompensation during admission (binary)
ININ25_0	Number of emergency department visits for HF in the last 2 years
hba1	Haemoglobin A1c (%)
colest_hdl	High-density lipoprotein cholesterol (mg/dL)
ININI500e_1	Bradyarrhythmia identified as a cause of heart failure (binary)

APPENDIX

B

Descriptions of the variables used in the complete MCC model

In the RePluris use case, which focused on predicting 30-day readmissions in patients with multiple chronic conditions, a feature importance analysis was conducted on the complete set of candidate clinical, social, and functional variables. The ranking was derived from the best-performing configuration identified during the systematic experimentation phase and reflects the relative contribution of each predictor to the discrimination between readmitted and non-readmitted patients.

The importance ranking presented in Chapter 7 includes variables encoded according to their original REDCap identifiers. Given that several of these predictors correspond to categorical indicators derived from the Ollero pluripathological classification and other structured clinical instruments, an explicit clarification of their meaning was considered necessary to enhance interpretability.

This appendix, therefore, provides a detailed description of all variables appearing in the feature importance figure, using their original variable codes alongside clinically explicit definitions. The objective is to ensure full transparency regarding the predictors included in the ranking and to facilitate reproducibility of the modelling process. By documenting the precise meaning of each coded variable, particularly those representing specific discharge categories or complexity criteria,

this appendix strengthens the methodological clarity and traceability of the feature selection procedure.

Table B.1: Description of variables included in the feature importance ranking (RePluris dataset)

Variable name	Description
olleroa_alt___1	Ollero category A1 at discharge: Heart failure with mild limitation of physical activity according to pluripathological patient criteria.
asist9	The patient receives personal assistance and help with household tasks from the same person.
comorb5	Heart failure prior to admission
asist3	The patient receives home help so that he can be cared for (personal care).
dias2	Number of days per week that someone helps with household chores.
disfuncion	Systolic dysfunction
n_complicaciones_cat	Number of complications during hospitalisation (categorical variable)
dias3	Number of days per week that a person assists them (personal care)
complicaciones_totales	Number of complications during hospitalisation
complic_medica	Medical complications
asist2	Home help for household chores
n_complicaciones	Number of complications during hospitalisation
fevi	Preserved Left Ventricular Ejection Fraction (LVEF)
cuidador___2	Prior to admission, patient care was provided by second-degree relatives.
diagcer_15	Neoplasia with some type of complication.
vig17	VigFRAIL. Cancer.
profund___1	PROFUND at discharge. Age > 85.
horas2	Number of hours per day that someone helps with household chores.
ayuda___6	If help is needed, social volunteers or home help providers are the ones who provide it.
destino_alta	Discharge destination (categorical variable)

Variable name	Description
diagcer_1	Heart failure at discharge
severvf___1	VIGFrail severity criteria at admission. Cancer.
civil	Current marital status.
cuidalt_3	Destination upon discharge. (1. Permanent social and healthcare centre, 2. Temporary social and healthcare centre, 3. Home, 4 Other acute care centres)
cuidalt_6h___3	Domestic help and personal care upon discharge. Operational.
char_4	Charlson Comorbidity Index. Ischemic heart disease at discharge
cuidador___4	Prior to admission, patient care was provided by caregivers.
procedencia_new	Origin (other social or healthcare center, home)
char_8	Charlson Comorbidity Index. Cancer at discharge.
diagpre1	Presumptive diagnosis. Heart failure.
tipo3___2	Type of home help: private.
horas3	Number of hours per day that a person assists them (personal care)
sit_clin___2	Baseline functional status recorded by the clinician: Requires assistance with complicated (instrumental) tasks.
nyha	Baseline NYHA (New York Heart Association).
profund_score_cat	PROFUND at discharge. Score categorised.
profund_score	PROFUND at discharge. Score.
asist8	Destination upon discharge: 1. To my home, 2. To a temporary residential centre, 3. Long-stay unit in a social and healthcare centre, 4. To a relative's home, 5. To a permanent residential centre, 6. Other.
complic_sindr_geriatrico	Complication due to a geriatric syndrome
edad	Age in years.
vig19	VigFRAIL. Heart disease.
vig6	VigFRAIL. Degree of cognitive impairment (absent, mild-moderate, severe-very severe).
sca___1	Can walk at discharge.
dete_psico_org	Psychological impairment.
profund___2	PROFUND at discharge. Active neoplasia at discharge

Variable name	Description
profund___8	PROFUND at discharge. No carer or carer other than spouse at discharge
ayuda___7	If help is needed, services provided by the retirement home are the ones who provide it.
cuidalt_4___4	Upon discharge from hospital, patient care will be provided by carers (not family members).
vig18	VigFRAIL. Respiratory disease.
vig4	VigFRAIL. Barthel index (independence, mild to moderate dependency, severe to very severe dependency, absolute dependency).
profund___7	PROFUND at discharge. Barthel < 60 (from questionnaire at discharge).

This thesis was completed in Bilbao on Monday 23rd February, 2026