Postgraduate Medical Journal, 2025, **101**, 1202, 1351–1360
https://doi.org/10.1093/postmj/qgaf102
Advance access publication date 6 July 2025
Original Research

Machine learning approaches for predicting heart failure readmissions

Amaia Pikatza-Huerga[1,*], Aitor Almeida[1], Raul Quiros[2], Nere Larrea[3], Mari Jose Legarreta[3], Unai Zulaika[2], Rodrigo Garcia[4], Susana Garcia[4]

[1]Faculty of Engineering, University of Deusto, Av. de las Universidades, 24, Deusto, E-48007 Bilbao, Bizkaia, Spain
[2]Internal Medicine Department, Hospital Costa del Sol, 29603 Marbella, Málaga Spain
[3]Kronikgune, Galdakao, Spain
[4]Faculty of Health Sciences, University of Deusto, Av. de las Universidades, 24, Deusto, E-48007 Bilbao, Bizkaia, Spain

[*]Corresponding author. Postal mail: Faculty of Engineering, University of Deusto, Av. de las Universidades, 24, Deusto, E-48007 Bilbao, Bizkaia, Spain.
E-mail: a.pikatza@deusto.es

Abstract

Purpose: This study aims to develop and evaluate machine learning (ML) models to predict the likelihood of hospital readmission within 30 days after discharge for patients with heart failure (HF). The goal is to compare the predictive accuracy of ML models with traditional methods such as those based on Cox proportional hazards and logistic regression, to improve clinical outcomes and reduce hospital costs.
Methods: We conducted a prospective cohort study of patients discharged from five hospitals following admission for HF. Data were collected on variables including sociodemographic characteristics, medical history, admission details, patient-reported outcomes, and clinical parameters. ML techniques were employed to analyse the data and predict readmission risk, incorporating strategies to handle class imbalance and missing data. Model performance was assessed based on accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and F1 score.
Results: Ensemble methods with Synthetic Minority Over-sampling Technique balancing and bagging improved the predictive performance of ML models compared with traditional models. The best-performing ensemble model, using decision trees, Gaussian Naïve Bayes, and neural networks, achieved an AUC of 0.81. In contrast, Cox and logistic regression models showed significantly poorer performance (AUC of 0.58 and 0.50, respectively). SHapley Additive exPlanations analysis revealed that frailty, anxiety, and depression were critical in predicting readmission.
Conclusion: ML models, particularly those using ensemble methods, significantly outperform traditional models in predicting short-term readmission for patients with HF. These findings highlight the potential of ML to improve clinical decision-making and resource allocation in HF management.

Key messages:

What is already known on this topic?

- Traditional models like Cox regression and logistic regression have limited accuracy in predicting 30-day readmissions for AHF patients. ML shows potential but lacks clinical validation.

What this study adds?

- Ensemble ML models (decision trees, Naïve Bayes, and neural networks) achieved superior prediction accuracy (AUC 0.81) over Cox (0.58) and logistic regression (0.50). SHAP analysis identified frailty, anxiety, and depression as key predictors.

How this study might affect research, practice, or policy?

- ML models can improve early risk stratification, enabling targeted interventions to reduce readmissions and optimize healthcare resources. Integration into clinical workflows could enhance decision-making and policy development.

Keywords: machine learning; acute heart failure; explainability; bagging; readmission prediction

Received 25 February 2025; **revised** 26 May 2025; **accepted** 08 June 2025

Introduction

Acute heart failure (AHF) is a serious medical condition that often requires hospitalization. Further, even after receiving treatment and being discharged, many patients are readmitted to hospital within a short time due to the recurrence of symptoms or related complications, so-called post-discharge syndrome [1].

The readmission index is a quantitative measure indicating the percentage of AHF patients who are rehospitalized within a specific period, most commonly within 30 days after discharge. It serves as a key indicator of healthcare quality, disease progression, and patient prognosis. High readmission rates are linked to poorer long-term outcomes, including increased mortality, and often reflect clinical instability or inadequate care. As such, the readmission index is widely used to assess hospital performance, guide quality improvement efforts, and shape reimbursement policies [2]. Reducing readmissions is a major goal in heart failure (HF) management, as it improves patient outcomes and lowers healthcare costs.

As the population ages, the prevalence of HF increases, as does the incidence of readmissions. In Spain, Bonilla-Palomas *et al.* [3] measured the incidence of HF hospitalization in 2021 and showed an increase of 61% from 2003 to 2015, mainly in the elderly. In patients older than 75 years of age, there is a 40% chance of being readmitted in the next year after an acute episode of HF [4].

Hospitalization has consequences for the patient and the health care system. In terms of health outcomes, readmissions are associated with thromboembolic events and short- and long-term mortality [5]. The economic costs associated with HF decreased over time since diagnosis, but remain significant in part due to survival bias. Escobar *et al.* [6] showed that hospital admissions accounted for 63.3% of HF-related costs and 38.2% in the fourth year, with HF being the first cause of admission in Spain.

Post-discharge syndrome refers to a critical period of vulnerability after discharge and requires a multifaceted and coordinated approach to reduce readmissions and improve outcomes in patients with HF. Risk stratification tools have been developed to improve resource allocation by prioritizing high-risk cases. However, current prediction models are inaccurate, partly due to the heterogeneity of the data collected and the lack of external validation in different populations [7].

Artificial intelligence (AI) refers to the capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. Machine learning (ML), a subset of AI, involves the development of algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed for each task. [8]

In 2016, Mortarazi *et al.* [9] reported that ML methods improved on the prediction of readmission after hospitalization for HF obtained using logistic regression and provided the greatest predictive range for observed readmission rates. Suliman *et al.* [10] are currently performing a systematic review on the predictive performance of ML compared with statistical methods in time-to-event analysis of cardiovascular disease (CVD). In recent years, ML has emerged as a transformative tool for predicting CVD, leveraging advanced algorithms to analyse complex health data. Techniques such as deep learning have demonstrated superior performance in identifying risk factors and classifying patient health states from datasets, improving early detection and treatment outcomes [11, 12]. Various studies highlight the potential of ML methods like support vector machines, random forest, and decision trees, which have shown promising accuracy in predicting heart disease [13, 14]. Overall, the integration of ML into cardiovascular medicine not only enhances diagnostic precision but also optimizes patient management and resource utilization in healthcare systems [12, 15]. Historically, the incidence of readmission has been very difficult to model and there is great hope that ML may improve on the predictive performance of Cox and logistic regression models [16].

This study aims to use ML techniques to predict the likelihood of a patient with AHF being readmitted to hospital within 1 month of their initial admission and to compare these techniques with Cox proportional hazards and logistic regression models. Early readmission risk prediction could enable preventive interventions, thereby improving clinical outcomes and reducing hospital costs.

Methods

Design

The data collection process was conducted as part of the Predictive Models of Readmission in Heart Failure-REIC study (ClinicalTrials.gov NCT03300791), a prospective cohort study. This is a non-probabilistic convenience sample of patients recruited consecutively in each participating centre for 12 effective months.

Participants

Patients with HF discharged from five participating hospitals after admission for acute decompensated heart failure (ADHF) were recruited and followed up for 1 year. The study included patients over 18 years of age admitted for AHF syndrome, which encompasses both acute (*de novo*) HF and acute decompensated HF. This category includes patients with previously diagnosed HF who experienced a progressive or rapid worsening of signs and symptoms requiring urgent intervention (ICD-9-CM codes: 428.x; and some from 402.x). Patients were excluded if they experienced an episode of ADHF during admission for another reason, were transferred from other hospitals, had had myocardial infarction or stroke within 4 weeks before admission, and/or had a life expectancy of <1 year, as well as if they were unable to complete questionnaires due to sensory impairment, dementia, or language barriers. Patients with missing follow-up data were excluded from the final analysis.

Trained interviewers used a standardized guide for interviewing patients and were directly supervised by clinical researchers. The interviews were conducted just before patients left the hospital. After discharge, patients were interviewed by telephone the day after the index emergency department (ED) visit and 7 days later. Clinical records were reviewed up to 2 months after the ED visit.

Outcome

The primary outcome measured was whether readmission occurred within 30 days after hospital discharge.

Predictors

Data were collected at discharge on the following variables:

- Sociodemographic characteristics: age, sex, and ethnicity.
- Medical history: '*De novo*' or chronic decompensated HF; time since diagnosis; cardiovascular risk factors; aetiology; previous echocardiographic parameters; number of previous admissions and ED visits; baseline functional status (New

York Heart Association [NYHA] class); comorbidities; previous treatments and procedures; and previous vaccinations (pneumococcus, influenza).

- Admission details: precipitating factors; symptoms; signs; ancillary test results; treatments and procedures; functional status at discharge (NYHA class); decompensated comorbidities and any other cardiac complications: acute myocardial infarction, ventricular fibrillation, cardiogenic shock, or cardiac arrest; other complications: renal, hepatic, and/or thromboembolic conditions, stroke, nosocomial infection, adverse effects or drug interactions; and length of stay.
- Patient-reported outcome measures:
  Dyspnoea at discharge was measured on a Likert scale ranging from 'absent' to 'the worst possible' and an analogue visual scale ranging from 0 (no dyspnoea) to 10 (the worst dyspnoea imaginable) [17].
  Minnesota Living with Heart Failure Questionnaire (MLHFQ): this questionnaire assesses the perceived impact of HF and its treatment on patients' lives. It consists of 21 items covering the physical, psychological, and social limitations associated with HF. Each item is rated on a scale ranging from zero (not at all) to five (very much). The total MLHFQ score is obtained by summing the scores for all 21 items (range: 0–105); the higher the score, the poorer the health-related quality of life. Scores can also be calculated for physical and emotional dimensions (based on eight and five items respectively). The MLHFQ has been validated for use in Spain [18].
  Barthel Index: the Barthel Index is a scale used to measure a person's level of disability in performing daily activities such as feeding, bathing, grooming, dressing, and mobility. It consists of 10 items rated based on how much assistance the person needs for each activity. While the scale typically uses scores of 0, 5, 10, and 15 for each item, some activities are rated with response options of only 0 or 5; or 0, 5 or 10. Item scores are summed to give a total score that ranges from 0 (complete dependence) to 100 (complete independence), with different levels of dependence categorized as severe (0–50), moderate (51–75), or mild independence (76–100). This method has been widely used in similar research to assess disability [19].
  Tilburg Frailty Indicator (TFI): this indicator is composed of two parts. The first part (Part A) describes various determinants of frailty based on sociodemographic characteristics and health-related issues. The second part (Part B) contains 15 items that measure three frailty domains: physical (eight items), psychological (four items) and social (three items). The response options are yes/no/sometimes for four items, and yes/no for the others. All items are then dichotomized into 0 (present) or 1 (absent). As well as a total score, domain scores are calculated for Part B. Scores are obtained by summing the scores on the corresponding items. A total score ≥ 5 points indicates frailty. [20, 21]
  Pre-episode mood was measured by the Hospital Anxiety and Depression Scale (HADS), which is a self-administered questionnaire consisting of two seven-item subscales, one for anxiety and one for depression. The items of the anxiety subscale were selected based on the analysis of the Hamilton Rating Scale for Anxiety, avoiding the inclusion of physical symptoms that could be confused by patients with symptoms of their physical illness. The items of the depression subscale focus on the area of anhedonia. [22]

Sample size

The sample size was determined using the rule of thumb of 10–20 outcome events per predictor variable in multivariable models. Assuming a 15% 30-day readmission rate across four centres over 12 months, ∼1100 patients (660 in the development cohort and 440 in the validation cohort) are expected to be enrolled, yielding over 100 readmission events in the development sample—thereby exceeding the minimum requirement for reliable estimation. Additionally, Haoqing Wang *et al.*'s [23] recommendation that a sample of 250 patients offers a good balance between model performance and data collection cost was taken into account.

Statistical analysis

Descriptive statistics

Only quantitative data were analysed in this study. First, descriptive statistics were generated and calculated: frequency tables were used for categorical variables, and means with standard deviations were reported for continuous variables. To compare sociodemographic characteristics, HF history, and questionnaire responses between cases and controls at different time points, we used the Chi-square test or Fisher exact test for categorical data, as appropriate. For continuous variables, normality was assessed using the Shapiro–Wilk test and homogeneity of variances with Levene test. When data were normally distributed and variances were homogeneous, comparisons were performed using Student t-test; if these assumptions were violated, the non-parametric Wilcoxon rank-sum test was applied. A P-value $< .05$ was considered statistically significant.

Preprocessing

Missing data Two approaches were used to address the missing values in the dataset collected for this study. In the first approach, we used ML models capable of handling missing values directly, without prior imputation. In the second approach, we imputed missing values using three different methods, namely, global constant (0), mean, and median, and then applied a broader set of models. The algorithms used are detailed in Supplementary material 1 (see Additional file 1).

Outliers Extreme or outlier values in the dataset were identified using both visual and statistical methods. Initially, box plots and histograms were constructed to visually inspect the distribution of each continuous variable and highlight potential outliers. Once identified, outliers were carefully reviewed. If an outlier was due to a data entry error or was not plausible based on clinical or experimental context, it was excluded from the analysis. Otherwise, if the value was plausible but extreme, sensitivity analyses were conducted both including and excluding these data points to assess their impact on the results.

Data scaling To avoid introducing bias in certain ML models, particularly those sensitive to feature scale, data preprocessing included standardization using a Standard Scaler.

Class imbalance

There was substantial imbalance in the dataset, with 84% of patients not being readmitted versus only 16% being readmitted. This imbalance could bias predictive models towards the majority class, diminishing their ability to correctly identify patients at risk of readmission. To mitigate class imbalance, we applied three methods: Synthetic Minority Over-sampling Technique (SMOTE), balanced subsampling during training, and class weight balancing in models that allow this type of adjustment. SMOTE generates synthetic minority class samples by interpolating between existing instances, enhancing the representation of the minority class. Balanced subsampling reduces the number of majority class samples to match the minority class, preventing model bias towards the majority class. Class weight balancing assigns higher weights to the minority class during training, penalizing misclassifications more heavily and improving model sensitivity to the minority class. Further details on these techniques are provided in Supplementary material 1 (see Additional file 1). It was decided to impute missing values and apply class imbalance management techniques, such as SMOTE, subsampling and weight balancing, to improve the detection of the minority class and, consequently, the model's overall performance in terms of F1 score.

Feature selection

The feature selection process was carried out in several stages. Initially, all available features were included to train the different ML models. Once the best-performing model was identified, SHapley Additive exPlanations (SHAP) values were used to rank the top 50 most influential features based on their contribution to the model's predictions. And, to ensure the model's applicability in real-world settings, these 50 features were reviewed in collaboration with two clinicians highly experienced in HF, who provided valuable input on the clinical relevance and practicality of these variables. Specifically, the clinicians assessed each variable for its importance and the feasibility of measuring it in routine medical practice. Their expertise was crucial in identifying which features would be realistic and beneficial to use in a real-world clinical environment. From this detailed evaluation, a refined subset of key features was selected.

Model output

The algorithms included in the bagging process were selected based on their performance in terms of sensitivity and specificity during the phase of applying various class imbalance handling methods. Only those algorithms that demonstrated strong performance in either sensitivity or specificity, without neglecting the other, were chosen for inclusion in the bagging ensemble. The aim was to balance these two metrics effectively, leveraging the power of bagging and soft voting to combine the strengths of each algorithm and enhance overall model performance. Next, the bagging process implemented in the study is described:

1. Data partitioning: initial dataset D is bifurcated into training (Dtrain) and testing (Dtest) subsets.
2. Training data balancing: subsampling is applied to Dtrain to generate n balanced training sets (T1, T2,..., Tn), ensuring equal representation of majority and minority classes.
3. Ensemble model training: each balanced set Ti serves as training data for m distinct ML models, resulting in n×m trained models. For each of these models, Grid Search was employed to optimize hyperparameters, ensuring optimal performance for each configuration. [24]
4. Prediction: the trained models generate predictions on Dtest, which is subjected to balancing procedures similar to those applied to the training data.
5. Ensemble decision-making: predictions are aggregated through two alternative voting mechanisms:

   Hard voting: prediction of the likelihood of readmission within 30 days based on the majority vote.
   - Soft voting: prediction of the likelihood of readmission within 30 days based on mean probability distributions.

Supplementary material Figure 2 visually represents the bagging and ensemble methods used in the final model to predict hospital readmission in patients with AHF. Supplementary material 1 explains the model output process (see Additional file 1).

Model evaluation

The performance and validity of the model were assessed using a separate test set that was not seen by the model at any stage during training. This ensured an unbiased evaluation of its generalization capacity. Several performance metrics were calculated on the test set, including accuracy, ROC AUC, sensitivity, specificity, and F1 score, providing a comprehensive overview of classification performance. Further details on the metric definitions can be found in Supplementary material 1: Model Evaluation (see Additional file 1).

For the final selected model, calibration was also evaluated to assess how well predicted probabilities reflected actual outcomes. This was done both graphically, using a calibration plot, and quantitatively, through the Brier score, which measures the mean squared difference between predicted probabilities and the actual outcomes. Lower Brier scores indicate better calibrated probability estimates.

Model explainability

In the same way that SHAP values were used for feature selection, SHAP values were also used to interpret the predictive model, identifying the variables most influential in predicting readmission risk, such as patient quality of life, social support, treatment adherence, and other comorbidities. SHAP values broke down the individual impact of each variable on the predictions, providing a quantitative explanation of their contribution to the model results. More details on the metrics for assessing the model and the methodology for interpreting the model can be found in Supplementary material 1 (see Additional file 1).

SHAP values provide insights into how each feature contributes to the model's predictions, thereby allowing a deeper understanding of the ensemble's decision-making process. Feature values are indicated by a spectrum with blue representing the lowest value and red the highest. A positive SHAP value represents an increase in the risk of readmission.

Software and libraries

Statistical analyses were performed using SAS for Windows, version 9.4 (SAS Institute, Cary, NC, USA). The remaining analyses were conducted using Python within a virtual environment configured for the project. The core ML tasks were implemented using scikit-learn (version 1.0.2). To address class imbalance in the dataset, the imbalanced-learn library (version 0.8.1) was employed. For time-to-event prediction and survival analysis, scikit-survival (version 0.17.2) was used. Additionally, XGBoost (version 1.6.2) was included for gradient boosting-based models. All packages were installed and managed in a local virtual environment, and dependencies such as NumPy, SciPy, and joblib were also utilized as required by the main libraries.

Results

Participants

A total of 1092 patients were recruited at the index admission; 96% of them were included in the analysis, with 843 patients assigned to the training cohort and 210 to the test cohort. Of the total sample, 15% of patients were readmitted during the follow-up

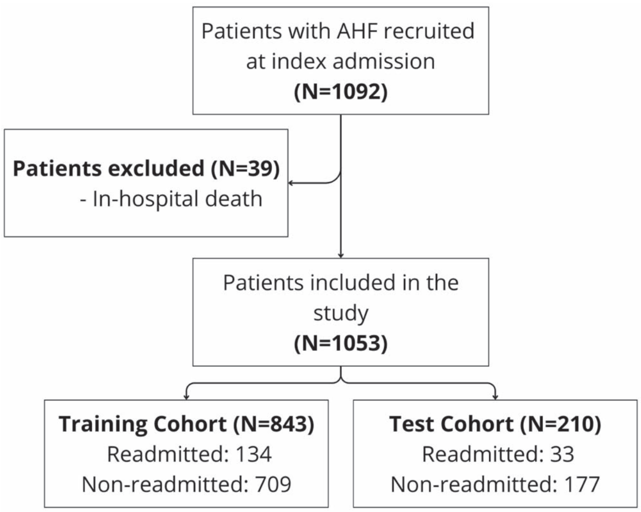

Figure 1 Flow chart of the data

period, 134 from the training cohort and 33 from the test cohort. (See Fig. 1)

There were slight statistically significant differences between the cohorts, but differences in the percentages of patients with mild chronic obstructive pulmonary disease and those admitted to the intensive care unit (ICU) were not clinically important. Table 1 presents a comparative descriptive analysis between the training and test sets, both overall and stratified by 30-day readmission status, including corresponding *P*-values.

Class imbalance

Different data balancing techniques influence model performance in predicting 30-day readmissions. Subsampling methods, particularly Subsample I and II, achieve the highest F1 scores, while SMOTE and Weight Balance perform worse. The variability in scores across techniques highlights the importance of selecting an appropriate balancing method. (See Supplementary Figure S1.) The superior performance of subsampling methods may be due to the presence of patients whose readmission risk is easier to predict.

ML modelling

Table 2 compares the performance of different ensemble models that use hard and soft voting techniques for predicting hospital readmission in patients with AHF. The model with the best predictive ability was 5 Decision Tree +5 Gaussian Naïve Bayes +5 Neural Network, with an AUC (95% CI) of 0.81 (0.73-0.88). The commonly used models yielded the following results: an AUC of 0.58 (0.47, 0.69), F1 score of 0.27, sensitivity of 0.41 and specificity of 0.72 with the Cox regression model; and an AUC of 0.50 (0.41-0.61), F1 score of 0.18, sensitivity of 0.31, and specificity of 0.62 with the logistic regression model.

Model applicability

SHAP values of the 50 main variables of the best initial model are shown in Figure S3 (see Additional file 1). The clinicians selected 32 of these, and as can be seen in the AUC values plotted in Fig. 2, the performance of some models decreases slightly when using this refined subset of features; however, it still outperforms both the Cox and logistic regression models. The best model achieved an AUC of 0.70 (0.61-0.80) and a Brier score of 0.24 (see Fig. 3 for the calibration plot).

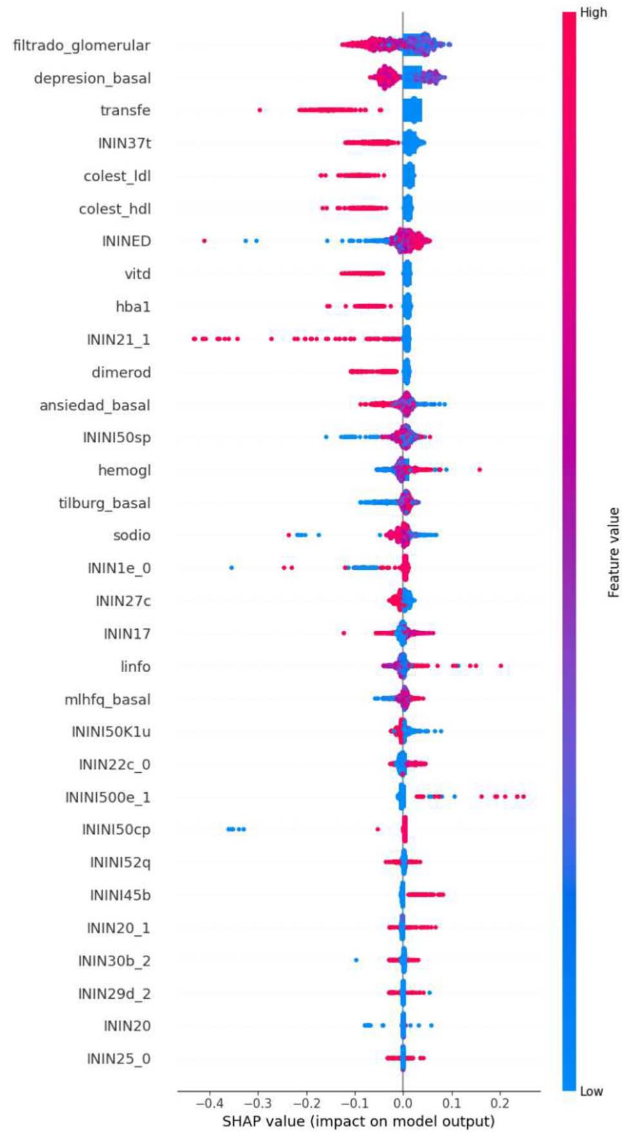

Figure 2 ROC AUC curve of models with refined subset of features

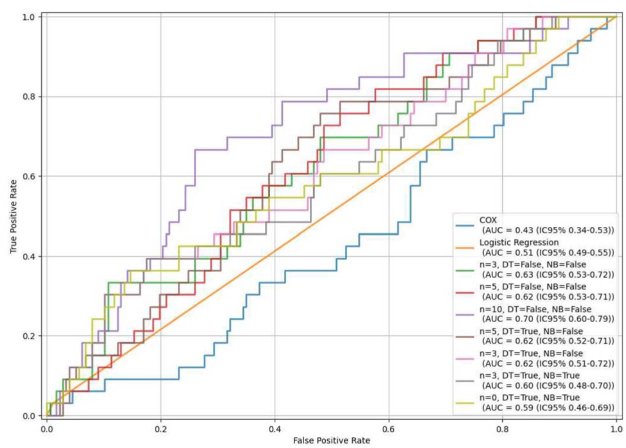

Figure 3 Calibration plot of the final selected model

Figure 4 shows the SHAP values of the features selected by the clinical collaborators. In these plots, each point represents a SHAP value for a single prediction; the position on the x-axis

Table 1. Comparative descriptive analysis between the training and test sets, overall and by 30-day readmission status.

Variables	TRAIN DATASET			P-value	TEST DATASET			P-value	P-value
	Total (N=843)	No (N=709)	Yes (N=134)		Total (N=210)	No (N=177)	Yes (N=33)		
Sex (Female)	398 (47.2)	334 (47.1)	64 (47.8)	0.8896	86 (41.0)	72 (40.7)	14 (42.4)	0.8514	0.1034
Age	78.2 ± 24.3	78.0 ± 26.1	79.2 ± 10.0	0.1378	77.1 ± 11.8	76.9 ± 12.4	78.7 ± 7.7	0.7703	0.3527
Known heart failure (HF)	385 (45.7)	327 (46.1)	58 (43.3)	0.5453	84 (40.0)	79 (44.6)	5 (15.2)	0.0015	0.1391
Years of evolution of HF	0.0 (0.0–3.0)	0.0 (0.0–3.0)	0.0 (0.0–3.0)	0.4582	0.0 (0.0–3.0)	0.0 (0.0–3.0)	3.0 (1.0–4.0)	0.0018	0.1521
Cardiovascular risk factors	10 (1.2)	10 (1.4)	0 (0.0)	0.1667	2 (1.0)	2 (1.1)	0 (0.0)	0.5395	0.751
Obesity	245 (29.1)	208 (29.3)	37 (27.6)	0.6867	70 (33.3)	57 (32.2)	13 (39.4)	0.4211	0.2266
Smoking				0.4745				0.3003	0.0602
Never smoker	535 (63.5)	455 (64.2)	80 (59.7)		115 (54.8)	101 (57.1)	14 (42.4)		
Ex smoker	234 (27.8)	191 (26.9)	43 (32.1)		70 (33.3)	56 (31.6)	14 (42.4)		
Smoker	74 (8.8)	63 (8.9)	11 (8.2)		25 (11.9)	20 (11.3)	5 (15.2)		
History of alcoholism	87 (10.3)	74 (10.4)	13 (9.7)	0.7974	25 (11.9)	19 (10.7)	6 (18.2)	0.2252	0.5052
Dislipemia	442 (52.4)	366 (51.6)	76 (56.7)	0.2788	109 (51.9)	84 (47.5)	25 (75.8)	0.0028	0.8912
Arterial blood pressure	689 (81.7)	578 (81.5)	111 (82.8)	0.7184	167 (79.5)	136 (76.8)	31 (93.9)	0.0254	0.4629
Hyperuricemia	114 (13.5)	90 (12.7)	24 (17.9)	0.1054	27 (12.9)	19 (10.7)	8 (24.2)	0.0333	0.7998
Anaemia	177 (21.0)	141 (19.9)	36 (26.9)	0.0689	46 (21.9)	36 (20.3)	10 (30.3)	0.2039	0.7731
Previous echocardio-graphic parameters	648 (76.9)	536 (75.6)	112 (83.6)	0.0445	167 (79.5)	134 (75.7)	33 (100)	0.0015	0.4104
Ischemic disease	37 (4.4)	31 (4.4)	6 (4.5)	0.9565	9 (4.3)	7 (4.0)	2 (6.1)	0.5835	0.9477
Angina	86 (10.2)	70 (9.9)	16 (11.9)	0.4684	17 (8.1)	14 (7.9)	3 (9.1)	0.8193	0.3579
Arrhythmia	440 (52.2)	369 (52.0)	71 (53.0)	0.8417	114 (54.3)	92 (52.0)	22 (66.7)	0.1199	0.5871
Valve disease	238 (28.2)	198 (27.9)	40 (29.9)	0.6500	60 (28.6)	49 (27.7)	11 (33.3)	0.5095	0.9223
Congestive HF	345 (40.9)	283 (39.9)	62 (46.3)	0.1702	88 (41.9)	66 (37.3)	22 (66.7)	0.0017	0.7963
Cerebrovascular disease	150 (17.8)	122 (17.2)	28 (20.9)	0.3060	27 (12.9)	21 (11.9)	6 (18.2)	0.3196	0.0870
Peripheral vascular disease	114 (13.5)	90 (12.7)	24 (17.9)	0.1054	26 (12.4)	21 (11.9)	5 (15.2)	0.5986	0.6627
Arterial blood pressure	665 (78.9)	556 (78.4)	109 (81.3)	0.4471	160 (76.2)	130 (73.4)	30 (90.9)	0.0306	0.3963
Diabetes Mellitus (DM)	289 (34.3)	245 (34.6)	44 (32.8)	0.7005	63 (30.0)	53 (29.9)	10 (30.3)	0.9670	0.2392
Mild chronic pulmonary	89 (10.6)	67 (9.4)	22 (16.4)	0.0161	33 (15.7)	25 (14.1)	8 (24.2)	0.1426	0.0367
Moderate-severe chronic pulmonary	108 (12.8)	90 (12.7)	18 (13.4)	0.8144	23 (11.0)	18 (10.2)	5 (15.2)	0.4002	0.4652
Mild kidney failure	60 (7.1)	53 (7.5)	7 (5.2)	0.3526	13 (6.2)	10 (5.6)	3 (9.1)	0.4514	0.6361
Moderate or severe kidney failure	70 (8.3)	61 (8.6)	9 (6.7)	0.4678	18 (8.6)	14 (7.9)	4 (12.1)	0.4275	0.9002
Nurse mild hepatic	23 (2.7)	18 (2.5)	5 (3.7)	0.4371	8 (3.8)	7 (4.0)	1 (3.0)	0.7990	0.4069
Nurse moderate or severe hepatic	15 (1.8)	13 (1.8)	2 (1.5)	0.7842	3 (1.4)	1 (0.6)	2 (6.1)	0.0146	0.7257
Tumour	83 (9.8)	70 (9.9)	13 (9.7)	0.9513	19 (9.0)	16 (9.0)	3 (9.1)	0.9925	0.7264
Lymphoma	7 (0.8)	5 (0.7)	2 (1.5)	0.3570	0 (0.0)	0 (0.0)	0 (0.0)	-	0.1852
Leukaemia	10 (1.2)	9 (1.3)	1 (0.7)	0.6080	2 (1.0)	2 (1.1)	0 (0.0)	0.5395	0.7751
Metastatic cancer	6 (0.7)	5 (0.7)	1 (0.7)	0.9587	3 (1.4)	3 (1.7)	0 (0.0)	0.4513	0.3127
Admission days	6.0 (4.0–9.0)	6.0 (4.0–9.0)	5.0 (3.0–8.0)	0.0455	6.5 (4.0–10.0)	6.0 (4.0–10.0)	7.0 (4.0–9.0)	0.7292	0.0047
Admission to ICU	9 (1.1)	8 (1.1)	1 (0.7)	0.6931	8 (3.8)	6 (3.4)	2 (6.1)	0.4618	0.0048
Minnesota score	58.5 ± 26.2	58.2 ± 26.0	60.1 ± 27.2	0.4467	59.9 ± 25.9	58.6 ± 25.7	67.2 ± 26.1	0.0710	0.5070
Barthel score	61.6 ± 39.4	62.0 ± 39.3	59.1 ± 40.0	0.3410	65.5 ± 37.9	67.9 ± 37.2	52.9 ± 39.8	0.0113	0.1902
Tilburg score	7.3 ± 3.1	7.1 ± 3.2	7.8 ± 3.0	0.0297	7.3 ± 2.8	7.0 ± 2.9	8.5 ± 2.3	0.0061	0.7293
HAD-depression score	9.8 ± 4.2	9.9 ± 4.3	9.4 ± 4.0	0.1613	9.7 ± 4.3	9.7 ± 4.3	10.0 ± 4.0	0.9913	0.8151
HAD-anxiety score	10.4 ± 4.5	10.5 ± 4.5	9.8 ± 4.4	0.1317	10.2 ± 4.5	10.2 ± 4.6	10.1 ± 3.9	0.7355	0.4895

indicates the impact of the feature on the model output: positive values push the prediction towards readmission, while negative values push it towards non-readmission. The colour represents the feature value, with high values in pink and low values in blue. Features are ordered vertically from most important at the top to least important at the bottom. As can be seen, glomerular filtration rate is the most important feature for the prediction of readmission in the final model, and lower levels indicate a higher probability of readmission. Higher baseline depression values decrease the likelihood that the model predicts patient readmission. This feature has a mixed impact, with some values having a negative effect and others having a positive effect on the model output. On the other hand, variables from tricuspid regurgitation velocity to the end are associated with a minor improvement in the model's predictions, but these features do not have a strong influence on the overall model output compared with some of the other variables shown.

Discussion

We developed and internally validated an ML-based model that achieves excellent discriminative power. To our knowledge, this performance has not been achieved with classical predictive

Table 2. Performance of different ensemble models before feature selection.

	Accuracy (mean ± SD)	ROC AUC (mean ± SD)	Sensitivity (mean ± SD)	Specificity (mean ± SD)	F1 Score (mean ± SD)
CoX	0.67 ± 0.03	0.58 ± 0.06 95% CI: (0.47, 0.69)	0.41 ± 0.09	0.72 ± 0.03	0.27 ± 0.06
Logistic Regression	0.57 ± 0.03	0.50 ± 0.05			
95% CI: (0.41, 0.61)	0.31 ± 0.08	0.62 ± 0.04	0.18 ± 0.05		
Hard Voting	**Accuracy (mean ± SD)**	**ROC AUC (mean ± SD)**	**Sensitivity (mean ± SD)**	**Specificity (mean ± SD)**	**F1 Score (mean ± SD)**
3 DecisionTree + 3 GaussianNB	0.72 ± 0.05	0.73 ± 0.05 95% CI: (0.63, 0.83)	0.90 ± 0.08	0.56 ± 0.13	0.76 ± 0.04
5 DecisionTree + 5 GaussianNB	0.74 ± 0.06	0.74 ± 0.06 95% CI: (0.62, 0.86)	0.90 ± 0.08	0.58 ± 0.14	0.77 ± 0.04
10 DecisionTree + 10 GaussianNB	0.79 ± 0.04	0.79 ± 0.04 95% CI: (0.71, 0.87)	0.91 ± 0.06	0.67 ± 0.10	0.81 ± 0.03
5 DecisionTree + 5 GaussianNB + 5 NeuralNetwork	0.79 ± 0.05	0.80 ± 0.05 95% CI: (0.70, 0.90)	0.93 ± 0.03	0.66 ± 0.10	0.82 ± 0.03
Soft Voting	**Accuracy (mean ± SD)**	**ROC AUC (mean ± SD)**	**Sensitivity (mean ± SD)**	**Specificity (mean ± SD)**	**F1 Score (mean ± SD)**
3 DecisionTree + 3 GaussianNB	0.76 ± 0.05	0.76 ± 0.05 95% CI: (0.66, 0.86)	0.84 ± 0.10	0.68 ± 0.11	0.77 ± 0.05
5 DecisionTree + 5 GaussianNB	0.77 ± 0.05	0.77 ± 0.0 5 95% CI: (0.67, 0.87)	0.85 ± 0.11	0.70 ± 0.11	0.78 ± 0.05
10 DecisionTree + 10 GaussianNB	0.79 ± 0.04	0.79 ± 0.04 95% CI: (0.71, 0.87)	0.89 ± 0.07	0.70 ± 0.10	0.81 ± 0.03
5 DecisionTree + 5 GaussianNB + 5 NeuralNetwork	0.81 ± 0.04	0.81 ± 0.04 95% CI: (0.73, 0.89)	0.93 ± 0.03	0.69 ± 0.09	0.83 ± 0.03

SD, standard deviation; NB, Naïve Bayes; CI, confidence interval.

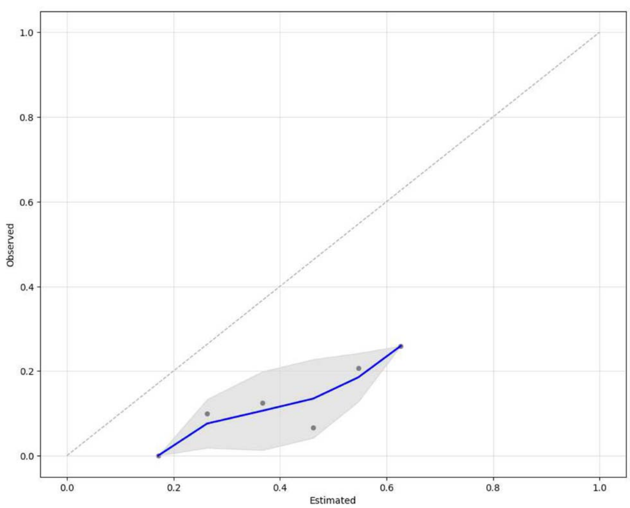

Figure 4 SHAP values of the final selected model. *(The glossary of variables can be found in Supplementary Material 1 (see Additional file 1))*

models. The delay in the implementation of ML models in electronic health records and the obstacles to their use by physicians and policymakers in the decision-making process motivated us to develop other more readily interpretable and easy-to-use models that achieve good predictive ability.

Our main hypothesis was that patient-reported information beyond known clinical variables could play a role in the likelihood of being readmitted in the short term [17]. Mood has been identified as a key predictor, confirming the findings of Veskovic *et al.* [25]. On the other hand, frailty is another key predictor identified by our model, as recent literature suggests [26].

Except for the subsampling technique, the F1 scores were quite low for all the techniques, not exceeding 0.2. On the other hand, the subsamples showed notably higher mean F1 scores, with subsample II being the best overall performer. The differences between the subsamples suggest the existence of distinct groups of non-readmitted patients, some of whom are more easily identifiable.

Based on the success of the subsampling technique, ensemble and bagging methods have been applied to further improve the results. By using bagging with multiple balanced subsets, this approach trains numerous models on diverse, balanced samples, mitigating class imbalance by ensuring equal representation of each class in every subset. Bagging reduces bias by averaging predictions across the models, increasing the likelihood that minority class patterns are captured. The ensemble method, which combines predictions through majority voting, enhances performance on imbalanced data by integrating outputs from a variety of balanced subsamples, resulting in a more robust model that is less sensitive to class imbalance and better at generalizing across classes.

The table displaying the performance of the different ensemble methods compares hard voting with soft voting, highlighting that soft voting achieves higher specificity and overall improved performance metrics. Its ability to accurately identify true negatives while maintaining high sensitivity and other key metrics makes it particularly effective in contexts where minimizing false positives is crucial.

When only the subset of key features is used, after removing 96% of the original variables, the best model achieves an AUC of 0.70. This corresponds to a 13.6% decrease compared with the AUC of 0.81 obtained using the full feature set, but the drop is not substantial considering the significant reduction in input complexity. This finding demonstrates that the subset selected by the clinicians retains much of the model's predictive ability while being more practical for real-world applications.

In addition, the results show that the new ensemble methodology with subsampling outperforms traditional models, such as those based on Cox and logistic regression. This approach improved the sensitivity and specificity metrics and handled class imbalance more effectively. Though more computationally intensive, the ensemble model showed superior performance in terms of the AUC and overall predictive ability. This methodology offers better results in scenarios with unbalanced classes than do traditional models.

SHAP values are the way we show the interpretability of the ML model. In order to reduce confusion and dependency between variables, two clinical collaborators decided, based on the whole model, which variables could have a clinical interpretation. In relation to this, it is important to note that SHAP values should not be interpreted as in regression models, where weights are assigned to each variable, and reflect the magnitude and direction of risk associated with that variable. In ML models, SHAP values should be interpreted as the influence of each variable on the model prediction. In particular, high baseline levels of depression and anxiety appear to be associated with a lower likelihood of readmission, but the SHAP analysis indicates that higher baseline depression decreases the model's prediction, making readmission less likely, while lower depression increases it, suggesting higher readmission risk. In fact, in the descriptive statistics table, there are no clinically significant differences in depression and anxiety levels when the outcome is present or absent.

Limitations

This study presents several limitations that should be considered when interpreting the results. First, the response rate among eligible patients may have introduced selection bias, as those who responded to the questionnaires had to be in relatively good health to complete them. This limits the generalizability of the results to the broader HF population.

One of the most important factors in predicting early readmission among HF patients is cognitive impairment, typically assessed using standardized tools such as the Mini-Mental State Examination or the Mini-Cog [27–29]. In our study, it was not possible to adequately evaluate this aspect due to the real-world clinical context in which the research was conducted. As a result, a single question from the TFI was used as an indirect measure, which was not statistically significant. We recommend that future studies include a formal assessment of cognitive function [30].

Finally, a key limitation of ML-based models is the need for their outputs to be interpretable in order to support integration into clinical practice. These models also need to be incorporated into automated systems such as electronic health records to maximize their utility. Although they demonstrate acceptable predictive capacity, often higher than that of traditional regression models, their real-world impact will only be known once they are validated and implemented in clinical settings. Additionally, the ensemble methods used, such as bagging and soft voting, have their own limitations. Bagging can increase model complexity and risk overfitting if not properly controlled, while soft voting may be affected by poorly calibrated models or by assumptions in methods like Naive Bayes. These considerations should be taken into account when interpreting the model's performance. Further studies are required to explore their clinical applicability and added value in decision-making.

Conclusion

The ML-based model developed demonstrated strong discriminative power, outperforming traditional models such as those based on Cox and logistic regression. Key predictors, including mood and frailty, were identified as significant factors in predicting short-term readmission, consistent with findings from previous studies.

SHAP values were used to enhance the interpretability of the model, showing the influence of variables like depression and anxiety on the model's predictions. Although higher levels of depression were associated with a lower likelihood of readmission, no significant clinical differences were found between readmitted and non-readmitted patients.

The model's handling of class imbalance through subsampling and ensemble methods improved its performance. Additionally, using a small set of key features maintained much of the model's predictive ability, highlighting its potential for real-world application.

Despite these promising results, integration into clinical practice remains a significant challenge. Future research should focus on external validation using datasets from different populations and healthcare settings to assess generalizability. Developing a prospective dataset and testing the model in real-time clinical workflows will be crucial for evaluating its feasibility and clinical utility. In addition, incorporating formal cognitive assessments and structured social determinants of health could help address some of the data limitations identified in this study. Research into optimizing model calibration and explainability, particularly in the context of ensemble methods, will also be essential to support clinical decision-making and foster trust among healthcare providers.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions. The authors are also grateful for the support of the Deustek5 Research group; the participating hospitals, and in particular, ED physicians, other clinicians and staff members of the various clinical, research, quality, and medical units of these hospitals; and the patients who participated in the study.

Author contributions

A.P.-H. (Data curation, Methodology, Writing – original draft), A.A. (Methodology), R.Q. (Writing – original draft), N.L. (Formal analysis), M.J.L. (Data curation, Formal analysis), U.Z. (Methodology), S.G. (Data curation, Methodology, Writing – original draft), and R.G. (Conceptualization, Writing – review & editing). Data collection was carried out by the ReIC working group, and all authors reviewed and approved the final version of the manuscript.

Supplementary data

Supplementary data is available at *Postgraduate Medical* Journal online.

Competing interests

No competing interest is declared.

Ethics approval and consent to participate

The research was conducted in accordance with the principles outlined in the Declaration of Helsinki (Br Med J 1964; ii: 177). The study was approved by the Basque Country Ethics Committee (Reference Number: PI2015151). Written informed consent was obtained from the patients or their legal representatives.

Funding

This research has been funded by the Carlos III Health Institute through project 'PI15/01343' (co-funded by the European Regional Development Fund/European Social Fund 'Investing in your future'), the Basque Government Health Department (2015111003), and the Research Committee of Galdakao-Usansolo Hospital (OSIBG18/203).

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Krumholz HM. Post-hospital syndrome—an acquired, transient condition of generalized risk. *N Engl J Med* 2013; **368**:100–2. https://doi.org/10.1056/NEJMp1212324
2. Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following heart failure (HF) hospitalization. *Partnership for Quality Measurement* —p4qm.org. https://p4qm.org/measures/0330. (20 May 2025, date last accessed).
3. Bonilla-Palomas JL, Anguita-Sánchez MP, Elola-Somoza FJ. et al. Thirteen-year trends in hospitalization and outcomes of patients with heart failure in Spain. *Eur J Clin Investig* 2021; **51**:e13606. https://doi.org/10.1111/eci.13606
4. Esteban Fernandez A, Anguita M, Bonilla JL. et al. 1-year hospital readmissions due to cardiovascular causes after a heart failure episode in elderly patients in Spain. *Eur Heart J* 2022; **43**:ehac544.1044.
5. Chaudhury P, Alvarez P, Michael M. et al. Incidence and prognostic implications of readmissions caused by thrombotic events after a heart failure hospitalization. *J Am Heart Assoc* 2022; **11**:e025342. https://doi.org/10.1161/JAHA.122.025342
6. Escobar C, Palacios B, Gonzalez V. et al. Evolution of economic burden of heart failure by ejection fraction in newly diagnosed patients in Spain. *BMC Health Serv Res* 2023; **23**:1340. https://doi.org/10.1186/s12913-023-10376-z
7. Kansagara D, Englander H, Salanitro A. et al. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011; **306**:1688–98. https://doi.org/10.1001/jama.2011.1515
8. Yuan H. Current perspective on artificial intelligence, machine learning and deep learning. *Appl Comput Eng* 2023; **19**:116–22. https://doi.org/10.54254/2755-2721/19/20231019
9. Mortazavi BJ, Downing NS, Bucholz EM. et al. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes* 2016; **9**:629–40. https://doi.org/10.1161/CIRCOUTCOMES.116.003039
10. Suliman A, Masud M, Serhani MA. et al. Predictive performance of machine learning compared to statistical methods in time-to-event analysis of cardiovascular disease: a systematic review protocol. *BMJ Open* 2024; **14**:e082654. https://doi.org/10.1136/bmjopen-2023-082654
11. Krittanawong C, Virk HUH, Bangalore S. et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* 2020; **10**:16057. https://doi.org/10.1038/s41598-020-72685-1
12. Chowdhury MA, Rodrigue Rizk J, Chiu C. et al. The heart of transformation: exploring artificial intelligence in cardiovascular disease. *Biomedicines* 2025; **13(2), 427**. https://doi.org/10.3390/biomedicines13020427
13. Paithane P. *Heart Disease Prediction Using Multiple Machine Learning Algorithms*, Adv Rob Tec 2024; Vol. **2**, 2024, 1–5. https://doi.org/10.23880/art-16000114
14. Bhuvana P, Rohith B, Swathi BM. et al. Prediction of cardiovascular disorders using machine learning. Educational Administration: Theory and Practice 2024, **30**:1206-1215.
15. Mathur P, Srivastava S, Xiaowei X. et al. Artificial intelligence, machine learning, and cardiovascular disease. Clinical Medicine Insights (2020). Cardiology, 14, 1179546820927404. https://doi.org/10.1177/1179546820927404
16. Rajkomar A, Oren E, Chen K. et al. Scalable and accurate deep learning with electronic health records. *npj Digit Med* 2018; **1**, 18.
17. Garcia-Gutierrez S, Villanueva A, Lafuente I. et al. Factors related to early readmissions after acute heart failure: a nested case–control study. *BMC Cardiovasc Disord* 2023; **23**:17. https://doi.org/10.1186/s12872-022-03029-2
18. Zuluaga MC, Guallar-Castillón P, López-García E. et al. Generic and disease-specific quality of life as a predictor of long-term mortality in heart failure. *Eur J Heart Fail* 2010; **12**:1372–8. https://doi.org/10.1093/eurjhf/hfq163
19. González N, Bilbao A, Forjaz MJ. et al. Psychometric characteristics of the Spanish version of the Barthel index. *Aging Clin Exp Res* 2017; **30**:489–97. https://doi.org/10.1007/s40520-017-0809-5
20. Gobbens RJJ, van Assen MALM, Luijkx KG. et al. The Tilburg Frailty Indicator: psychometric properties. *J Am Med Dir Assoc* 2010; **11**:344–55. https://doi.org/10.1016/j.jamda.2009.11.003
21. Vrotsou K, Machón M, Rivas-Ruíz F. et al. Psychometric properties of the Tilburg Frailty Indicator in older Spanish people. *Arch Gerontol Geriatr* 2018; **78**:203–12. https://doi.org/10.1016/j.archger.2018.05.024
22. Quintana JM, Padierna A, Esteban C. et al. Evaluation of the psychometric characteristics of the Spanish version of the hospital anxiety and depression scale. *Acta Psychiatr Scand* 2003; **107**:216–21. https://doi.org/10.1034/j.1600-0447.2003.00062.x
23. Wang H, Yi W, Liu Y. An innovative approach of determining the sample data size for machine learning models: a case study on health and safety management for infrastructure workers. *Electronic Research Archive* 2022; **30**:3452–62. https://doi.org/10.3934/era.2022176
24. GridSearchCV — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. (20 May 2025, date last accessed).
25. Veskovic J, Cvetkovic M, Tahirovic E. et al. Depression, anxiety, and quality of life as predictors of rehospitalization in patients with chronic heart failure. *BMC Cardiovasc Disord* 2023; **23**:525. https://doi.org/10.1186/s12872-023-03500-8
26. Lai H-Y, Huang S-T, Anker SD. et al. The burden of frailty in heart failure: prevalence, impacts on clinical outcomes and the role of heart failure medications. *J Cachexia Sarcopenia Muscle* 2024; **15**:660–70. https://doi.org/10.1002/jcsm.13412
27. Agarwal KS, Kazim R, Jiaqiong X. et al. Unrecognized cognitive impairment and its effect on heart failure readmissions of

elderly adults. *J Am Geriatr Soc* 2016; **64**:2296–301. https://doi.org/10.1111/jgs.14471
28. Ketterer MW, Draus C, McCord J. et al. Behavioral factors and hospital admissions/readmissions in patients with chf. *Psychosomatics* 2014; **55**:45–50. https://doi.org/10.1016/j.psym.2013.06.019
29. Patel A, Parikh R, Howell EH. et al. Mini-cog performance. *Circulation. Heart Failure* 2015; **8**:8–16. https://doi.org/10.1161/CIRCHEARTFAILURE.114.001438
30. Ketterer MW, Peltzer J, Mossallam U. et al. Cognitive impairment and reduced early readmissions in congestive heart failure. *Am J Manag Care* 2016; **4**:e1–6.

Postgraduate Medical Journal, 2025, **101**, 1202, 1351–1360
https://doi.org/10.1093/postmj/qgaf102
Original Research