Heart Failure Risk Stratification in an IoT Healthcare Platform using Ensemble Learning

Pietro Cassieri[+], Aiman Faiz[+], Claudio Pascarelli[#], Gianvito Mitrano[#],
Gianluca Fimiani[+], Marina Garofano[+], Mariangela Lazoi[#],
Claudio Passino[⊗], Alessia Bramanti[+], Giuseppe Scanniello[+]
[#]*University of Salento*, [⊗]*Scuola Superiore Sant'Anna di Pisa* , [+]*University of Salerno*

***Abstract*—This paper presents a predictive model founded on Machine Learning (ML) techniques to identify patients at Heart Failure (HF) risk. This model is an ensemble learning approach, a modified stacking technique, that uses three ML models leveraging clinical and echocardiographic features and then a meta-model to combine the predictions of these three models. We assessed the model on a real dataset, and our results suggest that it is better than the chosen baselines in the stratification of patients at HF risk. Specifically, we obtained high sensitivity (91%) and good accuracy (78%). To better study the value and potential of our model, we also contrasted its results with those obtained by using selected baselines. The preliminary results indicate that our predictive model outperforms these baselines that flatly consider features, *i.e.,* not grouping features in clinical and echocardiographic features.**

***Index Terms*—Healthcare, Machine Learning, Predictive Models, Artificial Intelligence, Digital Health**

I. Introduction

In modern healthcare systems, efficiently stratifying individuals[1] based on their risk profiles and predicted health trajectories is crucial for optimizing medical interventions, resource allocation, and personalized treatment strategies. The increasing availability of electronic health data, wearable sensor data, and advanced Machine Learning (ML) techniques has created new opportunities for predictive modeling in patient management. However, despite these advancements, significant challenges remain in ensuring the predictive models' accuracy. In particular, the heterogeneity of patients and the complexity of disease progression demand stratification approaches that can adapt to diverse healthcare environments [1].

A major cause of morbidity and mortality worldwide is chronic Heart Failure (HF), which is defined by the heart's inability to pump blood efficiently. Ageing of the population and lifestyle risk factors such as smoking, poor diet, and high blood pressure are associated with an increase in the incidence of heart disease. Current techniques often fail to identify heart disease, even with advances in medical diagnostics [2], [3]. In addition, cardiovascular diseases are associated with the highest number of deaths among Non-Communicable Diseases, amounting to 17.9 million deaths annually worldwide based on World Health Organization data, followed by cancers (9.3 million), chronic respiratory diseases (4.1 million), and diabetes (2 million) [4].

[1]It consists in categorizing them into different risk groups based on their health status, disease progression, or other factors.

The PrediHealth project aims to address the challenges above by integrating telemedicine, mobile health (m-health) solutions, and predictive analytics into an IoT healthcare platform. Specifically, it leverages an interoperable IoT web-based application, a telemonitoring kit equipped with medical devices and environmental sensors, and ML/AI-driven predictive models to support clinical decisions [5]. As it is easy to follow, at the core of the PrediHealth project lies the predictive model–designed for the stratification of patients at HF risk–that is presented in this paper for the first time. When a patient is considered at risk is then equipped with a telemonitoring kit. In case the PrediHealth platform, thanks to the (wearable and environmental) sensors of this kit, detects risk factors associated with HF exacerbations, a medical intervention is triggered. The developed model is based on an ensemble learning approach–a modified stacking technique–that uses three ML models leveraging clinical and echocardiographic features and then a meta-model to combine the predictions of these three bottom models. In this paper, we detail our ML-based model and its initial empirical assessment, as well as the possible implications for our results inside and outside the goal of the PrediHealth project. The initial results of our model assessment revealed high model sensitivity (91%), ensuring that it nearly identifies all HF patients. As for accuracy, we obtained 78%. It is acceptable in the PrediHealth project, given the priority of identifying patients at risk of HF to be then included in a telemonitoring program, to act promptly to let them receive medical attention by using the IoT healthcare platform developed in that project. We also compared our model against three baseline models (*e.g.,* Decision Tree and Support Vector Regression). Our outcomes suggest that the use of our model outperforms the baseline models (*e.g.,* in terms of the accuracy and the sensitivity).

II. Background and Related Work

A. The PrediHealth Project

In PrediHealth, patients are classified at HF risk (or not) using an ML predictive model, and those at risk are then equipped with wearable and environmental sensors, which are integrated into a telemonitoring system (*i.e.,* an IoT platform). For space constraints and limited relevance to the paper's main objectives, we only focus on the model for patient stratification. Additional details are available in [5].

B. Predictive Models in Digital Health

Existing research has primarily focused on developing predictive models using demographic variables to enhance early detection and risk stratification. For example, Medhi *et al.* [6] conducted a review by selecting 150 studies from databases such as PubMed, Google Scholar, and the Cochrane Library. Their synthesis detailed how AI/ML-powered models—spanning classical methods like SVMs, Random Forest, and decision trees to advanced deep learning approaches—can process large datasets from wearables, imaging, and other clinical sources to enable early detection, risk stratification, and personalized treatment of heart failure.

Traditional models, such as trajectory-based disease progression models and multivariable Cox regression, have been used to identify key predictors of mortality and morbidity, including age, diabetes, and left ventricular ejection fraction [2]. These models often emphasize heart failure with reduced ejection fraction (HFrEF), leaving a gap in predictive capabilities for preserved ejection fraction (HFpEF) [3].

Other researchers have explored ML models for HF prognosis, utilizing algorithms like logistic regression, decision trees, random forests, support vector machines (SVMs), and ensemble classifiers. Pinem *et al.* [7] demonstrated that a Voting Classifier combining Logistic Regression and SVM achieved the highest accuracy (88.04%) and ROC_AUC score (88%), with blood pressure and cholesterol emerging as key predictive factors. Similarly, Guo *et al.* [8] employed ML techniques to predict HF outcomes using EHR data, incorporating medical notes, laboratory results, and imaging data to achieve expert-level diagnostic accuracy. These studies underscore the role of ML in early HF detection, risk assessment, and personalized intervention strategies. Recently, Visco *et al.* [9] developed a predictive model for cardiovascular diagnostics. Their study identified correlations between worsening HF and creatinine, systolic pulmonary artery pressure (sPAP), and coronary artery disease(CAD). To enhance interpretability, they introduced a GP-based classifier and extracted from it a formula that achieved almost 97% accuracy, outperforming traditional ML models.

Despite these advancements, existing HF risk models face notable challenges. Voors *et al.* [10] showed inconsistencies in multinational cohort data, which affect model generalizability, while Adler *et al.* [11] pointed out limitations in single-center studies, particularly the exclusion of elderly populations. Moreover, many ML models struggle with imbalanced datasets and lack interpretability, which hinders their adoption [12].

We propose an interpretable ML-based predictive model that integrates heterogeneous patient data to improve risk stratification. The model is embedded in an IoT-enabled solution to support integrated hospital–territory–home care, and we show that it outperforms the considered baselines.

C. Used ML-Models

We used two machine learning models: the Logistic Regression Classifier and the Random Forest Classifier. The former is a linear model widely applied in binary classification problems [13]. It applies a weighted sum of the input features and uses the logistic function to estimate the probability of a given class. Due to its simplicity, it is computationally efficient and generally less prone to overfitting, especially in low-dimensional spaces. These characteristics, along with the inherent interpretability of the model's results (i.e., how easily a person can trace and explain the output based on the input features), make the Logistic Regression Classifier particularly valuable in clinical settings [14], where understanding model decisions is of central importance. In contrast, the Random Forest Classifier is a non-linear, ensemble-based method that constructs a collection of decision trees during the training phase and aggregates their outputs to form a final prediction [15]. This approach enables the model to capture intricate patterns and interactions across features, particularly in heterogeneous or high-dimensional datasets. While Random Forests are generally robust and versatile, they tend to lack the transparency of linear models, which may be a drawback in domains where interpretability is a key requirement.

III. Design of the Predictive Model

A. Dataset

The used dataset comprises real clinical records of patients diagnosed with heart failure, collected as part of a structured medical registry. It originally contained 1040 records, each representing a patient and characterized by a unique ID and 65 features that provide a comprehensive view of patient demographics, clinical history, treatment details, and outcomes. The dataset includes: age, height, weight, and body mass index (BMI), along with clinical indicators like ejection fraction (EF), New York Heart Association (NYHA) classification, and specific diagnoses, including ischemic, hypertensive, and dilated cardiomyopathy. Additionally, the dataset documents prescribed medications, including ACE inhibitors, Beta-blockers, and aldosterone antagonists, offering insights into pharmacological management. Longitudinal follow-up data track patient progress, recording mortality outcomes such as death, heart failure exacerbation, and interventions like heart transplants or ventricular assist device (VAD) implantation. The dataset also contains other biochemical information like Na+ (Sodium) or Uric Acid.

1) Features: We extracted 33 out of 65 available features in the dataset (detailed in Table I), including clinical and echocardiographic parameters relevant to cardiac evaluation based on the experience of some of the authors in the HF field and our knowledge of the state of the art (*e.g.,* [9]). The chosen characteristics encompass demographic attributes such as age, BMI, and sex, along with essential clinical indicators such as ejection fraction (EF), NYHA classification, NT-proBNP, creatinine, glucose, and hemoglobin (Hb). Additionally, the dataset contains information on comorbidities such as hypertension, diabetes, dyslipidemia, and respiratory diseases. The selection of features was based on the experience some of the authors gained in the field of HF, taking into account the purposes of the PrediHealth project (*e.g.,* stratifying individuals into patients at risk or not). Specifically, we kept all data related

to a standard clinical and instrumental cardiological examination (without null values). Therefore, key echocardiographic measurements were retained, including left ventricular mass index (LVMI), ventricular dimensions, atrial size, posterior wall thickness, septum thickness, right ventricular diameter, and tricuspid annular systolic excursion (TAPSE). Electrical abnormalities on electrocardiographic evaluation, such as left and right bundle branch blocks (BBSx, BBDx), atrial fibrillation (FA), atrial flutter, and pacemaker presence (PM), were also included. Height and weight were removed since they are inherently represented by the BMI feature. In contrast, other variables such as sodium, potassium, uric acid, urea, free triiodothyronine (FT3), and free thyroxine (FT4) were excluded as they were considered secondary to the main goal of our research. It is worth noting that the *Diagnosis* feature was split into two features to enhance the information granularity and allow for a clearer differentiation between the identified diagnoses. Some other features were deliberately excluded due to the high percentage of missing values. Although the PrediHealth platform collects environmental data (e.g., temperature), the model is used for risk stratification, focusing only on clinical and echocardiographic data.

Medication data, including beta-blockers, ACE inhibitors, and aldosterone antagonists, were incorporated to assess treatment effects. Also, two derived characteristics were included: "lifespan," representing the number of days between diagnosis and outcome, and "label." The features, grouped in "Clinical Features" and "Echocardiographic Features," that serve as the foundation of our model are detailed in Table I. The feature categorization into Clinical and Echocardiographic Features reflects their complementary roles in HF risk stratification.

2) Pre-processing: We preprocessed data to ensure data consistency and reliability. The goal was to include the highest number of data and have, at the same time, a balance between patients at HF risk and those not:

- Date values were converted into a standardized date-time format, and a new feature, "lifespan," was introduced to represent the number of days between the initial characterization date and the recorded outcome date ("death").
- We hold patients for whom we had information on their death, or the follow-up period was larger than or equal to three years. At the end, 464 patients remained in the dataset.
- Patients with at least one missing value in the features used by our model were removed. At the end, we removed 99 patients from the dataset, and then 365 patients remained.
- Patients with inconsistent values or placeholder values in terms of classification have been removed. At the end, we removed 8 patients from the dataset, and we reached the final dimension of the dataset of 357 patients.
- As for the outcome variable, we used a binary classification for the values in the dataset. We employed a threshold-based labeling approach, using a three-year threshold to classify patients into: "at risk" and not "at risk." It is justified by technical reasons and clinical bases. This threshold divides the dataset almost perfectly; in fact, we have 45% of samples with the label "at risk" and 55% of samples with the label "not at risk." From a clinical perspective, this threshold aligns with a time frame during which close monitoring is essential to mitigate adverse outcomes in patients at HF risk.

B. Predictive Model Workflow

Our predictive model (also meta-model or simply model from here onwards) is based on an ensemble learning approach, a modified stacking technique, to classify patients at HF risk or not. Standard stacking combines multiple heterogeneous models trained on the same set of features; our approach leverages domain-specific feature segregation to enhance predictive accuracy. The input data are divided into two distinct feature groups: clinical features, including demographic data, patient history, laboratory results, and risk factors, and echocardiographic features. Each feature group is processed separately using two identical ML models (*i.e.,* also clinical and echocardiographic specialized models from here onwards), each trained exclusively on one of the feature sets. This separation ensures that each model specializes in learning patterns from its respective data type without interference from other features. Both the specialized models are implemented using Logistic Regression Classifiers. In detail, we employ a five-fold cross-validation strategy using GridSearchCV, which systematically tests various combinations of hyperparameters for a split to determine the best configuration that maximizes the accuracy. In addition to these two models, we employ a third model trained on the complete set of features (*i.e.,* the merge between clinical and echocardiographic features). Indeed, we used a Random Forest Classifier to capture complex and non-linear relationships across the feature space that could be difficult to detect for Logistic regression classifiers. Using all the features together also allows for capturing interactions and patterns that span across different data types, potentially uncovering valuable insights that might be lost when the two sets of features are treated separately. A single train-test split (with 20% of the data reserved for testing) is exploited to evaluate the performance of these three models. The choice of the three models was informed by preliminary evidence suggesting their effectiveness alone and in combined use.

The outputs of the three models above, along with their confidence in the prediction, are used as inputs for our meta-model, which is implemented as a Logistic Regression Classifier responsible for aggregating the predictions of these three models into a single prediction. The idea underlying our model is inspired by the research presented by Pinem *et al.* [7], where they trained multiple models on the same set of features.

Integrating a set of models within a meta-model offers several potential advantages over a single ML model:

1) **Interpretability and Modularity**. By separating clinical and echocardiographic features, the contribution of each data type can be analyzed independently. Additionally, a third model trained on all features captures joint relationships across the data.
2) **Risk of Overfitting**. A single model trained on all features may overfit due to redundant or spurious correlations. By using specialized models for each group of features

TABLE I
DESCRIPTION OF THE FEATURES USED IN THE DATASET.

Feature	Description	Type
Clinical Features		
Primary_Diagnosis	Primary diagnosis of the patient	Categorical
Secondary_Diagnosis	Divided into Primary and Secondary if two diagnoses are present	Categorical
HFpEF	Heart failure with preserved ejection fraction classification	Categorical
EF	Ejection Fraction (%) – Left ventricular pumping function	Numeric
NYHA	New York Heart Association (NYHA) class (I-IV), measures heart failure severity	Categorical
Age	Age of patients in years	Numeric
BMI	Body Mass Index	Numeric
Sex	Sex of patients	Categorical (1 = Male, 0 = Female)
Hypertension	High blood pressure	Categorical (1 = Yes, 0 = No)
Dyslipidemia	Abnormal lipid levels	Categorical (1 = Yes, 0 = No)
Diabetic	Blood sugar disorder	Categorical (1 = Yes, 0 = No)
Bronchopneumonia	Lung disease	Categorical (1 = Yes, 0 = No)
Beta-Blocker	Use of beta-blockers	Categorical (1 = Yes, 0 = No)
ACE_SART	Use of ACE inhibitors or sartans	Categorical (1 = Yes, 0 = No)
Anti-Aldosterone	Use of aldosterone antagonists	Categorical (1 = Yes, 0 = No)
Echocardiographic Features		
PARETE POST	Posterior heart wall thickness	Numeric
SETTO	Septum thickness	Numeric
LVES_DIAM	Left ventricular end-systolic diameter	Numeric
LVED_DIAM	Left ventricular end-diastolic diameter	Numeric
VDx (PARAST)	Right ventricular diameter	Numeric
LVMI	Left ventricular mass index	Numeric
ASx	Left atrium size	Numeric
TAPSE	Tricuspid annular systolic excursion	Numeric
RS	Respiratory syndrome	Categorical (1 = Yes, 0 = No)
BBSx	Left bundle branch block	Categorical (1 = Yes, 0 = No)
BBDx	Right bundle branch block	Categorical (1 = Yes, 0 = No)
NT-proBNP	Heart failure biomarker	Numeric
Blood Creatinine Level	Kidney function marker	Numeric
Glucose	Blood sugar level	Numeric
FA	Atrial fibrillation	Categorical (1 = Yes, 0 = No)
Flutter	Atrial flutter	Categorical (1 = Yes, 0 = No)
PM	Pacemaker presence	Categorical (1 = Yes, 0 = No)
Hb	Hemoglobin level	Numeric

(clinical and echocardiographic), we reduce the risk of one type of feature dominating the learning process.

3) **Robustness**. This structure makes the system less sensitive to missing data. If a patient lacks echocardiographic measurements, the clinical model can still contribute to predictions. The metamodel also balances predictions, reducing the impact of errors from individual models.
4) **Data availability**. We built a model based on variables readily available in clinical reality.

IV. STUDY DESIGN

A. Research Question

We studied the following main Research Question (RQ): ***Does our predictive model outperform well-known ML techniques in classifying at-risk patients?***

In addition to the three ML-based models used in our meta-model, we considered two baselines–*i.e.,* Decision Tree[2] and Support Vector Classifier[3] (SVC). We choose these baselines because they represent the standard for comparison in AI application contexts and medical domains (*e.g.,* [18]).

B. Experimental Methodology

We conducted a comparative analysis between our model and the baselines that were trained on the entire set of 33 features in Table I. All models considered in the experimental assessment were trained and evaluated using the final dataset with 20% of the data reserved for testing. Baselines were implemented using standard algorithms from the `scikit-learn` library in Python. Just as for training the models underlying our meta-model, the baseline models have been optimized using a five-fold cross-validation strategy and GridSearchCV, which systematically tests various combinations of hyperparameters. This methodology ensures a fair comparison of the models when studying RQ.

[2]It is a tree-structured model that recursively splits the data based on feature values to predict an outcome. It is intuitive and interpretable and can handle both classification and regression tasks [16].

[3]It is a variant of Support Vector Machines [17] designed to identify the optimal hyperplane that separates data points of different classes with the maximum margin. It captures non-linear relationships using kernel functions.

C. Metrics to Assess the Predictions of the Models

The used evaluation metrics represent the standard for comparing ML techniques in the medical domain [19]:

- **Accuracy:** It is the proportion of patients that are correctly identified, and it is computed as: $\frac{TP+TN}{TP+TN+FP+FN}\%$. TPs (True Positives) refer to patients correctly identified as people at HF risk, while TNs (True Negatives) refer to patients correctly identified as patients not at HF risk. Instead, FPs (False Positives) represent patients mistakenly identified as at risk, while FNs (False Negatives) represent patients mistakenly identified as not at risk. Accuracy assumes values between 0% and 100%. The higher the value, the better it is.
- **Precision:** It indicates the proportion of patients predicted as at risk who truly are at risk and is calculated as: $\frac{TP}{TP+FP}\%$. Also, precision assumes values between 0% and 100%. The higher the value, the better it is.
- **Sensitivity (or recall):** It measures the proportion of actual positive patients that are correctly identified as having HF risk. It is computed as follows $\frac{TP}{TP+FN}\%$. Also, this metric

TABLE II
MEDICAL EVALUATION METRICS OF THE PROPOSED MODEL AND SPECIALIZED ONES

Model	Metric	Value
Meta-model	Accuracy	78%
	Precision	70%
	Sensitivity	91%
	F1-Score	79%
	DOR	20
Clinical model	Accuracy	72%
	Precision	67%
	Sensitivity	79%
	F1-Score	78%
	DOR	16
Echocardiographic model	Accuracy	78%
	Precision	71%
	Sensitivity	88%
	F1-Score	78%
	DOR	16.31
Random Forest Classifier	Accuracy	76%
	Precision	69%
	Sensitivity	88%
	F1-Score	76%
	DOR	14.5

assumes values between 0% and 100%. The higher the value, the better it is.

- **F1-Score:** It is the balanced harmonic mean of precision and sensitivity and is computed as: $2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}\%$. This metric assumes values between 0% and 100%. The higher the value, the better it is.
- **Diagnostic Odds Ratio (DOR):** It is a comprehensive metric that integrates sensitivity and specificity. It is computed as follows: $\frac{TP \times TN}{FP \times FN}$. The higher the value, the better it is.

Given the goal of PrediHealth, high precision is desirable, as it ensures that most of the patients at risk will take part in the telemonitoring program.

V. RESULTS, IMPLICATIONS, AND VALIDITY THREATS

A. Main Findings and Discussion

The results of our meta-model are summarized in Table II (top). Although the overall precision of 78% appears moderate at first glance, the high sensitivity (91%) ensures that nearly all patients at HF risk are correctly identified. The precision of 70% reflects a relatively high number of false positives, which is acceptable given the overall objective of the PrediHealth project. Furthermore, the high DOR (20) underscores the model's strong discriminative power between patients at HF risk and those not, reinforcing its suitability in the PrediHealth project. We can note in Table II that the meta-model allows obtaining better results as compared with its underlying ML-based models. The only difference concerns the precision value of the Echocardiographic model, which is 71%, while the precision value is 70% for the meta-model. These results support the conclusion that combining the predictions of the three underlying models captures data complementary aspects.

Analyzing the meta-model's Logistic Regression coefficients reveals that the Clinical base learner carries the highest weight (0.956), followed by the Echocardiographic model (0.873) and the Combined model (0.630). This strongly quantifies the benefit of our feature-group separation. Furthermore, the use of Logistic Regression for the base learners provides inherent interpretability via odds-ratios, supporting clinicians. For instance, in the clinical model, higher NYHA class (coef: 0.139) and age (0.048) increase HF risk, while BMI (-0.08) and ACE/SART therapy (-0.08) act as protective factors. In the echocardiographic model, NT-proBNP (0.367) is a strong risk predictor, whereas higher Hemoglobin (-0.128) is protective.

TABLE III
MEDICAL EVALUATION METRICS OF THE BASELINE ML MODELS

ML Model	Metric	Value
Decision Tree	Accuracy	67%
	Precision	60%
	Sensitivity	79%
	F1-Score	68%
	DOR	4.81
SVC	Accuracy	68%
	Precision	63%
	Sensitivity	72%
	F1-Score	68%
	DOR	4.76

The results in Table III suggest that the meta-model outperforms the baseline models. It achieves an accuracy of 78%, compared to 67% for the Decision Tree and 68% for SVC. The sensitivity of our model reaches 91%, which is well above the 79% and 72% for Decision Tree and SVC, respectively. This suggests a better capability of the meta-model to correctly identify patients at HF risk, and not when compared with the baselines. The precision of the meta-model (70%) is higher than the precision of the baselines (60% and 63% for Decision Tree and SVC, respectively). This indicates that the meta-model has a better balance between minimizing FPs and ensuring the inclusion of at-risk patients in the telemonitoring program. Additionally, the proposed model achieves an F1-score of 79%, outperforming Random Forest and SVC (68% for both models). Its DOR (20) further confirms its superior ability to distinguish between patients at HF risk and those not, compared to Decision Tree (4.81) and SVC (4.76). A McNemar's test on the test split yielded no statistically significant differences ($p > 0.05$) between the meta-model and the individual baselines, largely due to the limited sample size of the test set ($n \approx 72$). Nonetheless, the consistent improvement in sensitivity and DOR trends justifies the use of the meta-model. In a bootstrapping analysis,[4] performance metrics were computed for each resampled test set, and their empirical distributions were used to estimate mean values and 95% confidence intervals (CI). The meta-model showed a mean sensitivity of 87.7% with a 95% CI of [76.7%, 96.9%], a mean accuracy of 74.4% with a 95% CI of [65.3%, 84.7%], and a mean diagnostic odds ratio DOR of 20.1 with a 95% CI of [4.7, 64.4]. These results indicate moderate variability in the performance estimates, particularly for the DOR, reflecting the limited size of the test set.

B. Implications

Our predictive model uses a modified stacking ensemble. It achieved high sensitivity while maintaining adequate accuracy. This outcome suggests that a modified stacking ensemble approach can effectively identify a significant portion of

[4]To assess robustness, we performed this analysis (1000 iterations) on the test set predictions. Bootstrapping consists of repeatedly sampling, with replacement, from the test set to generate more datasets of equal size, assessing the estimation variability without retraining the model.

patients at HF risk. This is relevant for practitioners, as the model works as a valuable screening tool. Researchers could be interested in improving the model to enhance its accuracy while maintaining high sensitivity.

Our model integrates clinical and echocardiographic features through a flattened representation, enabling joint inference over heterogeneous data sources. This may improve interpretability by separating the contribution of global clinical variables from echocardiography-derived predictors in the resulting risk scores. Empirically, our results show that aggregating multiple base learners via a meta-learning layer improves predictive performance over individual models, supporting the use of ensemble learning and motivating further research on robust and interpretable meta-learning strategies. This point is clearly relevant for researchers and practitioners.

C. Threats to Validity

1) External validity: Our model was developed/evaluated using 357 samples, which represents a real clinical environment capturing authentic medical complexities, but this poses a threat to external validity. The model's performance might differ in other populations with varying demographic profiles, healthcare practices, or different diagnoses. Validation on an external or temporally separated cohort represents a critical next step, currently constrained by strict data privacy regulations (e.g., GDPR) governing clinical datasets. Furthermore, the set of 33 selected features might not be universally collected or hold the same predictive power in different contexts.

2) Internal Validity: The number of baselines used in the empirical assessment could affect internal validity. There could be other ML models outperforming our meta-model. However, the baseline models are among the most used in the medical field, justifying their choice.

3) Construct Validity: One key concern about construct validity is how we define patients at HF risk, as it may not capture when a patient faces a risk of worsening health. As for the metrics used to assess our predictive model and baselines, we used standard performance measures. They could not provide a complete picture of the phenomenon understudy. Lastly, our model assumes that medical records are accurate. Any errors in these records could affect the prediction.

4) Conclusion validity: Our definition of the binary outcome (at risk vs. not at risk) could have affected outcomes. The choice of 33 features can also affect internal validity.

5) Reliability validity: Due to the sensitive nature of patient health information and regulations like GDPR, the raw patient-level dataset used in this study cannot be made publicly available. This limits replications.

VI. Conclusion

We proposed a predictive model based on machine learning (ML) techniques to identify patients at Heart Failure (HF) risk. The model adopts an ensemble learning approach using a modified stacking technique, which combines the outputs of three underlying ML-based models. We evaluate our model on a real-world dataset, and the results show good performance in stratifying patients at HF risk. This is especially relevant for the telemonitoring program within the PrediHealth project in which our model has been developed. We also compared our model against selected baseline approaches, showing that it achieves superior performance.

References

[1] J. Smith and A. Doe, "Challenges in stratification and treatment personalization for alzheimer's disease," *Alzheimer's Research & Therapy*, vol. 14, no. 1, pp. 55–67, 2022.

[2] I. Odrobina, "Clinical predictive modeling of heart failure: Domain description, models' characteristics and literature review," *Diagnostics*, vol. 14, no. 4, p. 443, 2024. [Online]. Available: https://doi.org/10.3390/diagnostics14040443

[3] M. Toumpourleka, D. Patoulias, A. Katsimardou, M. Doumas, and C. Papadopoulos, "Risk scores and prediction models in chronic heart failure: A comprehensive review," *Current Pharmaceutical Design*, vol. 27, no. 10, pp. 1289–1297, 2021.

[4] A. Bramanti, A. Corallo, G. Clemente, L. Greco, M. Garofano, M. Giordano, C. Pascarelli, G. Mitrano, M. P. Di Palo, F. Di Spirito, M. Amato, M. Bartolomeo, R. Sorbo, M. Ciccarelli, P. Bramanti, and P. Ritrovato, "Exploring the role of voice assistants in managing noncommunicable diseases: A systematic review on clinical, behavioral outcomes, quality of life, and user experiences," *Healthcare*, vol. 13, p. 517, 02 2025.

[5] G. D. Filippo, S. Singh, M. Mazzotta, G. Sisto, M. Lazoi, G. Mitrano, C. Pascarelli, G. Fimiani, S. Romano, M. Garofano, and A. Bramanti, "Predihealth: Telemedicine and predictive algorithms for the care and prevention of patients with chronic heart failure," 2025. [Online]. Available: https://arxiv.org/abs/2504.03737

[6] D. Medhi and et al., "Artificial intelligence and its role in diagnosing heart failure: A narrative review," *Cureus*, vol. 16, no. 5, 2024.

[7] T. H. Pinem and M. Yan Rianto, "Integrating multiple machine learning models to predict heart failure risk," *Telematika*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274483186

[8] A. Guo, M. Pasque, F. Loh, D. L. Mann, and P. R. O. Payne, "Heart failure diagnosis, readmission, and mortality prediction using machine learning and artificial intelligence models," *Current Heart Failure Reports*, 2020.

[9] V. Visco and et al., "An explainable model for predicting worsening heart failure based on genetic programming," *Computers in Biology and Medicine*, vol. 182, p. 109110, 2024.

[10] A. A. Voors, W. Ouwerkerk, F. Zannad, M. Gheorghiade, A. Mebazaa, B. Merkely, M. Schou, K. Dickstein, C. Ariti, N. J. Samani, M. Metra, L. L. Ng, D. J. van Veldhuisen, C. C. Lang, P. van der Harst, and P. van der Meer, "Development and validation of multivariable models to predict mortality and hospitalization in patients with heart failure," *European Journal of Heart Failure*, vol. 19, no. 5, pp. 627–634, 2017.

[11] E. D. Adler, A. A. Voors, L. Klein, F. Macheret, O. O. Braun, M. A. Urey, M. Zhu, I. Sama, M. Tadel, C. Campagnari, and B. Greenberg, "Improving risk prediction in heart failure using machine learning," *European Journal of Heart Failure*, vol. 22, no. 1, pp. 139–147, 2020.

[12] R. M. Califf and M. J. Pencina, "Predictive models in heart failure: who cares?" *Circulation. Heart failure*, vol. 6 5, pp. 877–8, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:10796288

[13] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1958.

[14] P. Schober and T. R. Vetter, "Logistic regression in medical research," *Anesth Analg*, 2021.

[15] L. Breiman, "Random forests," *Machine Learning*, 2001.

[16] W. A. Belson, "Matching and prediction on the principle of biological classification," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1959.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, 1995.

[18] M. Shehab, L. Abualigah, Q. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. I. Alsalibi, and A. H. Gandomi, "Machine learning in medical applications: A review of state-of-the-art methods," *Computers in Biology and Medicine*, vol. 145, p. 105458, 2022.

[19] A. S. Glas and et al., "The diagnostic odds ratio: a single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 56, no. 11, 2003.