Decision Analytics Journal 7 (2023) 100242

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients

Alireza Ghasemieh, Alston Lloyed, Parsa Bahrami, Pooyan Vajar, Rasha Kashef *

Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada

ARTICLE INFO

Keywords:
Machine learning
Heart diseases
Feature selection
Labeling
Ensemble learning
Validation metrics

ABSTRACT

Early detection of heart complications is highly effective in treating patients with cardiovascular diseases. Various machine learning methods have previously been used for the early detection of heart diseases. However, existing data-driven machine learning (ML) approaches fall short of providing efficient and accurate heart disease detection. This misdiagnosis causes significant overcrowding in medical care facilities by patients that do not need emergency readmission or fatalities caused by discharging patients requiring emergency. This study proposes a novel model for detecting emergency readmission of heart disease patients by effectively identifying patients who require emergency assistance before the onset of heart attacks or other heart-related complications. A robust Stacking Ensemble Learner (SEL) is developed using ensemble learning to maximize the detection performance. Our SEL method predicts whether a patient with heart problems is required to get admitted as an emergency case after a preliminary admission. To ensure robustness and high accuracy in the prediction results across multiple runs, the XGBoost is used as a meta-learner in the SEL model. The novelty of this paper lies in (1) the use of behavior-based features to create a new class label for emergency readmission, which has not been previously explored in the existing data-driven machine learning approaches, (2) the paper utilizes a comprehensive private dataset from the MIT Laboratory for Computational Physiology, not adopted in clinical studies on heart failure and cardiovascular disease, and (3) The development of a robust Stacking Ensemble Learner (SEL) using ensemble learning, with XGBoost as a meta-learner, also contributes to the novelty of this study, as it achieves higher prediction performance compared to the baseline models, the use of ensemble learning in the SEL model helps to overcome the limitations of unstable training of the individual classification models. Experimental results show that the stacking model provides high accuracy, Recall, and F1 score compared to the baseline models such as logistic regression, k-nearest neighbor, Decision tree, Random Forest, support vector machines, bagging, and boosting. The SEL model has achieved an accuracy of 88% in predicting emergency readmission of heart-disease patients, which is very promising for the production-ready model in clinical practice.

1. Introduction

Heart disease is a major cause of hospitalization and readmission, leading to increased healthcare costs and decreased quality of life for patients. According to the World Health Organization (WHO), cardiovascular diseases are the most frequent cause of death globally. Each year, more than 17.9 million fatalities occur due to heart diseases, accounting for 31 percent of the overall death counts in the world. Thirty-three percent of these deaths are in people under the age of 70. Furthermore, 80 percent of these fatalities are caused by heart attacks [1,2]. Predicting which patients are at high risk of readmission can help healthcare providers prioritize interventions and improve patient outcomes [3,4]. Recent advancements in artificial intelligence have paved the way to develop more accurate models and algorithms for predicting diseases before their onset [5]. Machine learning (ML) models are increasingly used to predict readmission risk in heart-disease patients. These models use statistical algorithms to analyze large datasets of patient information, including clinical and demographic data, to identify patterns and predict outcomes. By analyzing large datasets of patient information, including clinical data such as medication history, hospitalization records, and comorbidities, machine learning models can identify patients at high risk for readmission and prioritize interventions to reduce this risk [5]. Various supervised and unsupervised learning methods are proposed to better detect heart-related diseases. Several supervised machine learning models can be used for this purpose, including logistic regression, decision trees, random forests, and neural networks. These models can be trained on large patient information datasets to identify readmission risk factors and develop predictive models. One key advantage of the data-driven

* Corresponding author.
E-mail address: rkashef@torontomu.ca (R. Kashef).

https://doi.org/10.1016/j.dajour.2023.100242
Received 7 February 2023; Received in revised form 21 April 2023; Accepted 29 April 2023
Available online 6 May 2023

ML models is their ability to handle large and complex datasets, including unstructured data such as clinical notes and images [6]. This allows healthcare providers to identify risk factors that may not be apparent using traditional statistical methods. In addition, ML-based models can continuously learn and adapt to new data. As new patient data becomes available, these models can be retrained to improve their accuracy and performance. However, traditional ML models suffer from data quality, bias, and interpretability [6]. In addition, the inefficiency and inaccuracy of the existing data-driven models would result in overcrowded medical care facilities by patients that do not need emergency readmission or, even worse, fatality caused by discharging patients that require emergency care from the healthcare system. Due to this sensitive nature, finding an accurate model becomes highly necessary. Recently, ensemble models have shown significant improvement compared to traditional single-based classifiers; however, existing ensemble-based approaches used hard voting schemes with no adaptability to the performance of the adopted baseline models, which results in low detection performance. In this paper, we have created a novel model, the stacking ensemble learner (SEL), to detect emergency readmission of heart disease patients early so that overall detection performance is maximized. The proposed model helps identify patients who require emergency assistance before the onset of heart attacks or other heart-related complications. Unlike other publicly available datasets in clinical studies related to heart failure and cardiovascular disease, it was decided to use a much more comprehensive dataset that is not publicly available. To get access to this dataset, a consent form was submitted and approved by the MIT Laboratory for Computational Physiology, after which permission was granted to access the dataset. The dataset was created as part of a study that extracted data from electronic healthcare records of patients admitted due to heart failure at the Zigong Fourth People's Hospital between 2016 and 2019. For this unlabeled data, we used existing behavioral-based features to create a new class label for emergency readmission of patients. Various baseline classification techniques are analyzed in the experimental analysis, and results show that the stacking model achieves the highest prediction performance for improving patient outcomes and reducing healthcare costs.

The main contributions of this paper can be summarized as follows:

- Developing a novel robust Stacking Ensemble Learner (SEL) using ensemble learning to maximize the detection of emergency readmission of heart disease patients by effectively identifying patients who require emergency assistance before the onset of heart attacks or other heart-related complications.
- Using behavior-based features to create a new class label for emergency readmission of patients, which has not been previously explored in existing data-driven machine learning approaches.
- Utilizing a comprehensive private dataset from the MIT Laboratory for Computational Physiology, which has not been previously used in clinical studies related to heart failure and cardiovascular disease.
- Achieving higher prediction in predicting emergency readmission of heart disease patients, very promising for a production-ready model in clinical practice compared to baseline machine learning models.
- Overcoming the limitations of unstable training of individual classification models by using ensemble learning in the SEL model

The rest of the paper is organized as follows: Section 2 presents a literature review and background of previous work on heart diseases and the cardiovascular system. Section 3 discusses the ensemble model for detecting emergency readmission of heart disease-related patients. Section 4 introduces the evaluation metrics used for performance assessment. Section 5 presents the experimental dataset, preprocessing stages, class labeling, and feature selection. Experimental results and analysis are discussed in Section 6. Section 7 concludes the paper and presents future works.

2. Literature review and background

Existing work on heart diseases and the cardiovascular system has been proposed using machine learning, categorized into single-based or ensemble-based methods as discussed below.

2.1. Single-based detection methods

In single-based methods, only one classifier is used; the choice of the classifier depends on the characteristics of the datasets, the distribution of patients features, and the properties of the classification approach in handling non-linear boundaries and data sparsity.

Authors in [6,7] introduced a model for heart disease prediction based on an offline machine-learning model that is integrated with real-time generated by users on Twitter. For feature selection, two different methods are proposed. The first method is univariate and involves creating a subset of features contributing more to determining a data point's label. The other method, called relief, involves using weights to create a degree of importance for features with more effect on the data point's label. Features used in this work include age, sex, blood sugar, chest pain, maximum heart rate, serum cholesterol and so on. Furthermore, different machine learning algorithms such as decision trees, support vector machines and regression are utilized for classification. Finally, the results of the classification algorithms are analyzed to predict the occurrence of heart disease. In [8], a novel IoT-based framework for predicting the probability of heart disease in patients is proposed. The framework consists of three main sections: data generation, storage, and analysis. Human body sensors gather information such as heart rate, respiratory rate, body temperature, blood sugar before and after meals and blood pressure. Data generated in IoT-based systems have high dimensionality and require massive storage capabilities. Thus, they suggested the usage of cloud computing for storing data which also increases the scalability of the whole architecture. The machine learning algorithm used for classification in this work is logistic regression which determines whether a patient could expect heart disease. Finally, the probability of heart disease occurrence is sent to a medical care facility for observation. In [9], they proposed a system that evaluates the possibility of heart disease or heart attack in a patient based on two sets of features. The first set includes features such as body temperature, blood sugar, heart rate, activity, EEG, EMG, ECG, oxygen saturation and respiration rate concerned with human body sensors' data. The second set has features such as age, sex, gender, diet, body mass index, smoking history, cholesterol value and diabetes history, which are gathered by electronic health records and medical tests. Text mining techniques are utilized to find important features in the second set based on electronic medical records. Furthermore, features in both sets are fused and assigned weights to demonstrate their importance for classification for each data point. Features with less importance will have lower weights and be neglected to reduce data's dimensionality in heart-related disease prediction settings. The classifier used for this work is a five-layer feed-forward neural network with backpropagation. The classifier is compared with other machine learning algorithms such as logistic regression, support vector machine, Naïve Bayes and decision tree and demonstrates promising results. An interesting attribute of the proposed system is its ability to recommend a diet based on the patient's heart disease prediction analysis. In [10], they investigated the heart rate and heart disease probability calculation in patients based on a modified artificial plant optimization technique. This algorithm finds the most effective features out of the dataset and resembles the growth process of a plant. Six of the 13 features available were dropped, and the most informative ones were kept. This work evaluates the model's performance based on Naïve Bayes, XGBoost and logistic regression. Performance evaluation is by precision and F1 score. Authors in [11] analyzed feature extraction with a fast Fourier transform. The paper suggests that some features are better captured in the frequency domain. Therefore, a fast Fourier

transform is performed on time series data points before classification for heart disease prediction. The main measurements used for this work are blood glucose, diastolic blood pressure, weight, heart rate and mean arterial pressure. The model has less computational complexity than other proposed methods. Guo et al. [12] proposed a machine learning-based model for predicting cardiovascular disease risk, incorporating various risk factors and clinical data. The model achieved high accuracy in predicting the risk of cardiovascular disease, demonstrating the potential of machine learning for personalized risk assessment. In [13], the authors discussed the potential applications of artificial intelligence and machine learning in precision cardiovascular medicine, including predicting outcomes, diagnosis, and treatment selection. The authors highlighted the importance of interpretability and transparency in developing and implementing machine learning models in clinical settings. Lin et al. [14] developed deep-learning models for predicting 30-day readmission in patients with heart failure. The models incorporated electronic health records and achieved high accuracy in predicting readmission, suggesting the potential of machine learning for improving healthcare outcomes and reducing healthcare costs.

A comprehensive review of the current state of machine learning in medicine is presented in [15], discussing the challenges and opportunities of using machine learning in clinical settings. The authors emphasized the importance of collaboration between clinicians, data scientists, and engineers to ensure the development and implementation of effective and ethical machine learning models. Shao et al. [16] proposed machine learning models for predicting cardiovascular events in patients with hypertension. The models incorporated various clinical features and achieved high accuracy in predicting the risk of cardiovascular events, demonstrating the potential of machine learning for personalized risk assessment and prevention. Wang et al. [17] developed machine learning-based prediction models for cardiovascular diseases using health examination data. The models achieved high accuracy in predicting the risk of cardiovascular diseases, suggesting the potential of machine learning for the early detection and prevention of cardiovascular diseases. A machine learning-based model for predicting acute myocardial infarction is presented in [18]. The model incorporated various risk factors and achieved high accuracy in predicting the risk of acute myocardial infarction, demonstrating the potential of machine learning for personalized risk assessment and prevention. Zhou et al. [19] developed a novel machine learning-based model for predicting all-cause mortality in patients with acute myocardial infarction. The model achieved high accuracy in predicting all-cause mortality, suggesting the potential of machine learning for improving risk assessment and management in patients with acute myocardial infarction. Zhu et al. [20] developed machine learning-based prediction models for the development of heart failure in patients with hypertension. The models achieved high accuracy in predicting the risk of heart failure, demonstrating the potential of machine learning for early detection and prevention of heart failure in high-risk patients. Zuo et al. [21] systematically reviewed machine learning models for predicting clinical outcomes in patients with coronary artery disease. The review identified several promising machine learning models for predicting clinical outcomes, highlighting the potential of machine learning for improving risk assessment and management in patients with coronary artery disease. Table 1 provides a summary of the recent work on ML-based models showing the model used, datasets, strengths, and limitations, in addition to the performance of each model.

Table 1 concisely summarizes the main features of several research papers that focus on using single-based machine learning for diagnosing cardiovascular diseases. It includes information on the machine learning models used in each study, such as SVM, Random Forest, Naïve Bayes, Decision Tree, and MLP, as well as the datasets employed. The dataset sources are diverse, ranging from private electronic health records to publicly available datasets such as the MIT-BIH Arrhythmia Database, Cleveland Clinic Foundation Heart Disease Dataset, and Framingham Heart Study. In terms of the benefits of each model, the literature reports several advantages, including improved accuracy, utilization of ECG data and patient-generated data, non-invasive approach, personalized treatment, early detection, low power consumption, and potential for diagnosis. However, several limitations are also outlined, such as limited dataset size, limited validation, no real-world validation, dependence on patient participation, lack of diversity in samples, potential for bias, and unstable training. Nonetheless, there is a great potential for using more robust machine learning models for cardiovascular disease (CVD) diagnosis using ensemble learning, as discussed next.

2.2. Ensemble-based detection methods

Ensemble-based techniques are introduced to overcome the limitations of single-based methods and obtain a robust classification. An ensemble combines two or more classifiers with varying strengths/weaknesses to build a more sustainable model with better performance. Existing ensemble-learning methods use bagging, boosting, and voting schemes [22–25]. Each ensemble works in its domain space with varying performance depending on the choice of the aggregate, the distribution and non-linearity, as we all, the imbalance classes in the dataset.

The work in [22] discusses an ensemble learning model using bagging for detecting Ischemia or coronary artery heart disease. Medical research shows that early diagnosis of this disease increases the survival rate significantly. Thus, finding an accurate model for detecting Ischemia heart disease in the early stages is crucial. This work gathers three types of features: information theory features and features from the frequency and time domains. The classifiers investigated in this model are XGBoost, K-nearest neighbor, decision tree and support vector machines. The best combination for the ensemble model is the XGBoost and support vector machine. The work in [23] is another boosting ensemble model which uses stochastic gradient descent, K-nearest neighbor, random forest, and logistic regression as classifiers. An ensemble model based on the hard voting of these classifiers is built to improve the accuracy of heart disease detection for patients. Basically, the most occurring prediction of the base models is used to determine the label of the ensemble method. The features used for this work are segmented into two groups of discrete and continuous characteristics. Their ensemble method performs best, with the stochastic gradient descent method taking second place in accuracy. Authors in [24] proposed a heart disease prediction model that uses ensemble learning and feature selection techniques. The authors compare the performance of different ensemble methods and feature selection techniques, demonstrating that their proposed model outperforms other models in accuracy and AUC. In [25], an ensemble machine learning model for predicting cardiovascular disease is presented. The authors compare the performance of different machine learning algorithms, including decision trees, random forests, and neural networks. They demonstrate that their ensemble model outperforms each algorithm in terms of accuracy and AUC. Deepika et al. [26] proposed an ensemble learning model for predicting cardiovascular disease. The authors combine multiple decision trees with different feature selection methods, and they demonstrate that their proposed model outperforms other models in terms of accuracy, sensitivity, and specificity. El-Sappagh et al. [27] presented an ensemble of machine learning algorithms for heart disease diagnosis. The authors combine decision trees, random forests, and support vector machines, demonstrating that their proposed model outperforms each algorithm in accuracy, sensitivity, and specificity. In [28], the authors proposed an ensemble deep-learning model based on electrocardiogram (ECG) signals for cardiovascular disease prediction. The authors utilize a combination of convolutional neural networks (CNN) and long short-term memory (LSTM) networks to capture both the spatial and temporal features of ECG signals. The proposed model is trained and evaluated using a large dataset of

Table 1
A summary of the ML-based heart-related prediction methods.

Paper	Model used	Dataset	Benefits	Limitations	Results
Yasin et al. [7]	LSTM, CNN	PTB Diagnostic ECG Database, MIT-BIH Arrhythmia Database	Ultra-low power platform, Secure IoT	Limited dataset size, No real-world validation	Accuracy: 95.62%, Sensitivity: 91.9%, Specificity: 98.2%
Ahmed et al. [8]	ML on Spark: Random Forest, KNN, SVM, Neural Network	Social media data from Twitter, Facebook, and Instagram	Non-invasive approach, Utilizes patient-generated data	Limited dataset size, Dependence on patient participation, lack of diversity in samples	Accuracy: 82.7%
Kumar and Gandhi [9]	3-tier IoT, SVM	Real-time ECG data collected from multiple sensors (MIT-BIH, PTB)	Early detection, Improved accuracy, personalized treatment, low power consumption	Limited dataset size, Limited generalizability	Accuracy: 91.23%, Sensitivity: 98%, Specificity: 97%
Ali et al. [9]	Deep ensemble, feature fusion	MIT-BIH Arrhythmia Database, PTB Diagnostic ECG	Improved accuracy, Non-invasive approach	Limited dataset size, No real-world validation	Accuracy: 99.03%, Sensitivity: 93.6%, Specificity: 99.8%
Sharma et al. [10]	Artificial Plant Optimization Algorithm, KNN, Decision Tree	Cleveland Clinic Foundation Heart Disease Dataset	Improved accuracy, Non-invasive approach	Limited dataset size, No real-world validation	Accuracy: 82.24%, Sensitivity: 95.4%, Specificity: 74.5%
Narayan et al. [11]	FFT, SVM, Naïve Bayes	UCI heart disease, Cleveland	Utilizes ECG data, a Non-invasive approach	Limited dataset size, Limited generalizability	Accuracy of 83.3%
Guo et al. [12]	SVM, Random Forest, Decision Tree, MLP	Framingham Heart Study, National Health and Nutrition Examination Survey	Improved accuracy, Utilizes patient- data	Limited dataset size, No real-world validation	AUC of 0.85
Krittanawong et al. [13]	Deep learning	Various publicly available cardiovascular datasets	Potential for diagnosis	Limited validation, Unstable training	AUC: 0.87–0.98
Lin et al. [14]	Deep learning	MIMIC-III	Improved accuracy, Non-invasive approach	Limited dataset size, No real-world validation	AUC of 0.80
Rajkomar et al. [15]	ML	Various	Wide-ranging review, Potential for diagnosis	Limited validation, Unstable training	High accuracy in predicting various outcomes
Shao et al. [16]	Random Forest, Decision Tree, SVM	Private dataset: Electronic Health Records	Improved accuracy, Utilizes patient- data	Limited dataset size, No real-world validation	AUC of 0.88
Wang et al. [17]	Random Forest, Logistic Regression, Decision Tree, SVM	Health examination data	Utilizes patient-generated data, large sample size	Limited validation, No real-world validation	AUC of 0.84
Zhao et al. [18]	SVM, Random Forest, MLP	MIMIC-III	Improved accuracy, Utilizes ECG data	Limited dataset size, Limited validation	AUC of 0.96 Accuracy: 91.08%
Zhou et al. [19]	ML	MIMIC-III	Improved accuracy, Utilizes ECG data	Limited dataset size, Limited validation	AUC of 0.92
Zhu et al. [20]	ML	Private dataset	Early detection, Utilizes patient-generated data	Limited dataset size, No real-world validation	Accuracy of 90%
Zuo et al. [21]	ML	Various	Wide-ranging review for precision medicine	Limited validation, Potential for bias, Unstable training	Accuracy of 88% and AUC of 0.91

ECG recordings from multiple sources. The authors also compare the performance of their proposed model with other deep learning models and traditional machine learning models. The results show that their proposed model outperforms others in accuracy, sensitivity, specificity, and AUC. They conclude that their proposed ensemble deep learning model can effectively predict cardiovascular disease based on ECG signals. The model can potentially be applied in clinical settings to assist healthcare professionals in the early diagnosis and treatment of cardiovascular disease. However, further validation and optimization of the model are necessary before it can be deployed in practical applications. Table 2 compares recent work on ensemble learning to diagnose heart disease or predict the risk of developing.

As shown in Table 2, we can observe that existing ensemble methods can be effective tools for diagnosing heart disease. Existing literature using ensemble learning for CVD falls short in addressing the following challenges:

- Limited dataset size: Some studies are limited to small-size datasets and lack external validation, limiting their results' scalability.
- No feature selection: Some studies do not perform feature selection, which may result in overfitting and reduced performance on new data.
- Unstable training: Some studies report unstable training of individual classification models with a high chance of overfitting, which may impact the model's overall performance.
- Limited to numerical-based patient features: Some studies are limited to dealing with numerical features only, which may not represent the larger population and may limit the generalizability of their results.

Overall, these limitations suggest a need for more robust and generalizable machine learning models for the early detection of heart diseases. The proposed ESL model in the abstract addresses some of

Table 2
A summary of the ensemble-based ML models for heart-related diagnosis.

Paper	Datasets	Model used	Limitations	Benefits	Results
[22] Tao et al. (2018)	Magnetocardiography recordings of 189 subjects with and without ischemic heart disease	Magnetocardiography-based machine learning methods	Small dataset size, limited to patients who underwent MCG testing, limited external validation	High accuracy in detecting and localizing ischemic heart disease using MCG-based features	Sensitivity of 92.5%, specificity of 98.3%
[23] Atallah and Al-Mousa (2019)	Cleveland Clinic Foundation Heart Disease dataset	Majority voting ensemble method with six machine learning algorithms	No mention of feature selection limited to one dataset	High accuracy and low false positive rate compared to individual models	Accuracy of 91.14%, F1-score of 0.91
[24] Asif and Wajid (2020)	Cleveland Clinic Foundation Heart Disease dataset	Ensemble learning with four machine learning algorithms and feature selection	Limited to one dataset, no external validation	Higher accuracy and reduced overfitting compared to individual models	Accuracy of 87.13%, F1-score of 0.88
[25] Chang et al. (2018)	Cleveland Clinic Foundation Heart Disease dataset	Ensemble learning with six machine learning algorithms	No feature selection limited to one dataset	Improved accuracy and sensitivity compared to individual models	Accuracy of 81.6%, sensitivity of 86.3%
[26] Deepika and Prabhu (2019)	Cleveland Clinic Foundation Heart Disease dataset	Ensemble learning with four machine learning algorithms	No feature selection limited to one dataset, unstable training	Improved accuracy and reduced overfitting compared to individual models	Accuracy of 86.6%, F1-score of 0.85
[27] El-Sappagh et al. (2019)	Cleveland Clinic Foundation Heart Disease dataset	Ensemble learning with eight ML algorithms	No feature selection limited to one dataset	Improved accuracy and reduced overfitting compared to individual models	Accuracy of 84.3%, F1-score of 0.84
[28] Jia et al. (2020)	Shanghai Chest Hospital Cardiovascular dataset	Random forest, support vector machine, artificial neural network, logistic regression	Limited to one dataset, no feature selection	High accuracy in predicting cardiovascular disease in Chinese patients	Accuracy of 89.5%, AUC of 0.90

these limitations by using behavior-based features, a comprehensive private dataset, and a robust stacking ensemble learner with a robust meta-learner to achieve higher prediction performance compared to baseline models. Compared to the existing ensemble-based approaches that used hard voting schemes, in this paper, we have created a novel and robust model, the stacking ensemble learner (SEL), to early detect emergency readmission of heart disease patients such that the overall detection performance is maximized.

3. The Stacking Ensemble Learner (SEL) model

In this paper, a stacking ensemble learner (SEL) is developed to maximize the detection performance while achieving robust and stable performance. A stacking ensemble classifier is a machine learning model that combines multiple base models to improve prediction accuracy. In this approach, the outputs of the base models are used as inputs to a higher-level model, known as a meta-model, which learns to make the final predictions.

3.1. The ESL architecture

The stacking scheme consists of two levels: level 0 and level 1. In level 0, the base models are trained on the training set, and their outputs are stored as new features in a new dataset. The validation set is then fed to the base models, and their predictions are used as inputs to the meta-model. This forms level 1, where the meta-model is trained on the validation set using the base models' outputs as features. Compared to the existing ensemble-based approaches that used hard voting schemes, our ESL model is a stacking ensemble classifier that uses a stacking Scheme through two levels, level 0 and level 1. Compared to any aggregate model in the ensemble process, stacking produces accurate predictions by leveraging the strengths of high-performing models on the classification task. The ESL, in contrast to bagging, implies that the models are often distinct (i.e., heterogeneous ensemble) and fitted to the same dataset (e.g., instead of samples of the training dataset). Unlike boosting, the ESL uses a single model to effectively integrate the predictions from the adopting models (e.g., instead of using a sequence of models that adjust the forecasting results of previous models). To determine how to combine the predictions most effectively from two or more base machine-learning algorithms, the proposed ESL model employs a meta-learning algorithm. The preprocessed data, including independent attributes vector, denoted by x and class attribute, denoted by y, are fed into all level-0 models (baseline models). After the predictions have been made, the outputs of level-0 models, denoted by $\hat{y}$, are fed into the metamodel at level 1 as inputs. This means that our meta-model is trained on the predictions obtained by baseline models on out-of-sample data. In addition to the above-specified inputs, the meta-model is also provided by the class attribute to determine which classification produced the accurate or wrong prediction. The level-1 classifier in this design oversees weighing the level-0 classifiers' output and producing the final classification result. Utilizing such an architecture has the benefit of typically producing greater performance than separate models. The level-1 classifier is critical in the stacking ensemble classifier's decision-making process. It is responsible for taking the outputs from the level-0 classifiers, essentially the predictions from the individual models, and combining them to generate the final classification result. The level-1 classifier can be any machine learning algorithm, such as XGBoost, logistic regression, decision tree, or support vector machine, which takes the level-0 classifier outputs as input features and produces a final classification result. The key idea behind the level-1 classifier is to leverage the strengths of different level-0 classifiers and create a more accurate and robust prediction by weighing their outputs appropriately. This way, the stacking ensemble classifier can overcome the limitations of individual classifiers and produce more reliable and accurate predictions. The weighing mechanism used by the level-1 classifier can be simple or complex, depending on the application and the nature of the level-0 classifiers. For example, one simple weighting mechanism uses equal weights for all level-0 classifiers, while a more complex mechanism could involve learning the weights using a separate optimization algorithm. The final classification result is then determined based on the weighted outputs from the level-0 classifiers with their predictions, as shown in Eq. (1).

$$h\_\{1\}(x), h\_\{2\}(x), \ldots, h\_\{k\}(x),$$

where h is the class label for a data point x using a given classifier (1)

These predictions are then used as input to the level-1 classifier, which produces the final prediction:

$$H(x) = g(f(h\_\{1\}(x), h\_\{2\}(x), \ldots, h\_\{k\}(x))) \quad (2)$$

where f is the function that aggregates the predictions of the level-0 classifiers, and g is the function that produces the final prediction.

The weights of the level-0 classifiers are learned during the training of the level-1 classifier. This is done by minimizing a loss function that compares the output of the level-1 classifier to the true labels:

$$L(y, H(x)) \tag{3}$$

y is the true label of the input x; the weights can be learned using any optimization algorithm, such as gradient descent or Newton's method. During prediction, the ESL classifier first applies the level-0 classifiers to the input and then uses the learned weights to produce the final prediction using the level-1 classifier. The Level-0 model output is:

$$Z = [z_1, z_2, \dots, z_n], z_i \in Rk \tag{4}$$

where Z: Level-0 model output matrix of size n × k, where n is the number of instances and k is the number of level-0 classifiers, z_i: The output vector of the ith instance of the dataset, where each element corresponds to the probability of the corresponding class according to the k level-0 classifiers. The Level-1 model input is:

$$X = [x1, x2, \dots, xn], xi \in Rk \tag{5}$$

where X: Level-1 model input matrix of size n × k, where n is the number of instances and k is the number of level-0 classifiers, x_i: The input vector of the ith instance of the dataset, which consists of the level-0 classifiers' output probabilities for each class. Finally, The Level-1 model output:

$$y = [y1, y2, \dots, yn], yi \in Rc \tag{6}$$

y: Level-1 model output matrix of size n × c, where n is the number of instances and c is the number of classes, y_i: The output vector of the ith instance of the dataset, which consists of the predicted probabilities of each class according to the level-1 classifier. The objective function of the Stacking ensemble classifier:

$$\hat{yi} = \text{argmax} j(y_{i,j}), i = 1, 2, \dots, n \tag{7}$$

$\hat{yi}$ is the predicted class label for the ith instance of the dataset, obtained by selecting the class with the highest predicted probability in yi. $y_{i,j}$ is the predicted probability of the jth class for the ith instance of the dataset. The pseudocode of the SEL is shown in Algorithm 1. The stacking ensemble classifier (ESL) proposed in the paper uses a two-level architecture to improve the accuracy of cardiovascular disease prediction.

Algorithm 1: SEL (D, X, Y)
Input:
- Training data D, with features X and labels Y
- Base classifiers h1, h2, ..., hm and a Meta-classifier g
Output:
- Ensemble model f
Begin // Level 0 Training Phase
For i = 1 to m do:
    Train hi on D
    For each (x, y) in D do:
        Append hi(x) to Xi
        Append y to Yi
End For
// Level 1 Training Phase
Train g on (Xi, Yi)
// Ensemble Model Prediction
For each new input x do:
    For i = 1 to m do:
        Append hi(x) to Xi
    y_ensemble = g(Xi)
End For
Return y_ensemble as the predicted output.

3.2. The level-0 and level-1 adopted models

In Algorithm 1, the function f can be any aggregation function, such as a simple average or a weighted average. The function g is typically a logistic regression, XGboost, or a neural network. At level 0, any set of classifiers can be adopted such that the choice of the input pool of classifiers depends on the following factors for building an aggregate, including (1) handling imbalance classes, (2) dealing with sparsity in data, (3) addressing nonlinear boundaries, (4) handling categorical features, and (5) reducing the overfitting. In this study, to build a robust ensemble which satisfies the above aggregation factors, six baseline classifiers are adopted, including Logistic Regression (LR), k-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost). These models were chosen because they are commonly used in the literature and have been shown to perform well on a range of classification tasks in CVD diagnosis, as shown in Section 2. In particular, Logistic regression can handle large datasets with probabilistic outputs, KNN is non-parametric and can handle nonlinear decision boundaries, Decision Tree (DT) can handle both categorical and numerical features, and RF combines multiple decision trees to improve performance and reduce overfitting. It is useful for handling high-dimensional datasets and can handle both categorical and numerical features; SVM is effective for handling datasets with a high number of features and can handle non-linear decision boundaries. It is also effective for handling imbalanced datasets. The XGBoost uses gradient boosting to improve performance and reduce overfitting. Combining the strengths of the adopted baseline models at level 1 ensures that categorical features as handled, the chance of overfitting is minimized, non-linear decision boundaries are addressed, and imbalanced classes are tackled.

This paper explores several meta-learners, including Support Vector Machines (SVM), Linear regression (LR), Multi-Layer Perceptron (MLP), and XGBoost. However, the XGBoost approach outperformed the other meta-learners and was selected for use in the final ESL model. The proposed architecture of the two-level ESL model is shown in Fig. 1. The use of multiple classifiers at level 0, followed by a meta-learner at level 1, helps to improve the accuracy and robustness of the overall prediction model while minimizing overfitting.

Each classifier is optimized by tuning its hyperparameters using grid search to find the best combination of hyperparameters for a machine learning model by systematically trying different values for each hyperparameter and evaluating the model's performance for each combination of hyperparameters. The hyperparameters tuned for the random forest classifier were the number of trees (n_estimators) and the maximum depth of each tree (max_depth). The best combination of hyperparameters was found to be n_estimators = 100 and max_depth = 3. The probability parameter was activated for the support vector machine (SVM) classifier, allowing the classifier to output probabilities of the predicted classes. For the decision tree classifier, the maximum depth of the tree was set to 3, which means the tree will not split further once it reaches a depth of 3. For the k-nearest neighbors (KNN) classifier, the number of nearest neighbors used to predict the class of a new data point was set to 5. For the XGBoost classifier, the hyperparameters tuned were the number of threads (n_jobs), learning rate (learning_rate), number of trees (n_estimators), maximum depth of each tree (max_depth), objective function used for optimization (objective), and the type of boosting algorithm used (booster). The best combination of hyperparameters was n_jobs = −1, learning_rate = 0.1, n_estimators = 100, max_depth = 3, objective = 'binary:logistic', booster = 'gbtree'. For the meta learner, we have increased the learning rate, the number of estimators and max-depth to enhance the results and add more stability in the final classifications. A summary of the hyperparameter used for each baseline model is summarized below.

- Random forest (RF): n_estimators = 100, max_depth = 3
- SVM: Probability activated

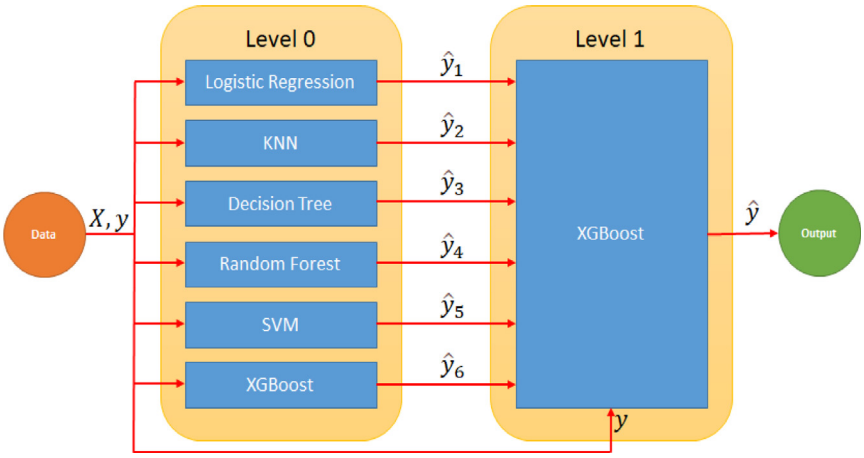


Fig. 1. The architecture of the Stacking Ensemble Learner (SEL) classifier.

- Decision tree (DT_): max_depth = 3
- KNN: Number of Neighboor: 5
- XGBoost (Baseline): n_jobs = −1, learning_rate = 0.1, n_estimators = 100, max_depth = 3, objective = 'binary:logistic', booster = 'gbtree'
- XGBoost (Meta Learner) : n_jobs = −1, learning_rate = 0.2, n_estimators = 200, max_depth = 5, objective = 'binary:logistic',

The stacking ensemble classifier offers several benefits over single models, such as improved accuracy and reduced overfitting. Additionally, it can handle different types of data and models, making it a versatile technique. However, it also has some limitations, such as increased computational complexity due to training multiple models and the risk of creating an overly complex model that may not generalize well. To comprehensively evaluate the prediction provided by the proposed model, multiple external quality metrics (i.e., using existing and generated class labels) are used to examine the model's performance compared to the baseline models, as discussed next.

4. Performance metrics

It is crucial to align the evaluation metrics with the problem definition to have a realistic understanding of model performance. In this way, various criteria are discussed below.

4.1. Recall

Suppose the patient has serious heart failure, which is an emergency case. A patient comes to the hospital, and the model is used to decide whether the patient is in an emergency case or not. Now the model classifies it as a non-emergency case. It means the misclassification jeopardizes the patient's life. Therefore, the false negative (FN) is the most crucial parameter to evaluate the model performance. Therefore *Recall* (Eq. (8)) is considered the primary evaluation metric.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{8}$$

4.2. Precision

On the other hand, if the model classifies non-emergency cases as emergency cases, it creates a huge problem for the hospital emergency department. The actual emergency cases will not receive sufficient medical service. But it is not a problem for a few misclassification numbers; further diagnosis shows they are not emergencies. Therefore, false positives should also be considered not to be so high. In this way, the *Precision (Eq. (9))* metric is used.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{9}$$

4.3. F1 Score

F1 Score (Eq. (10)) is used to have the effects of both aforementioned metrics in one formula, and it facilitates comparing the performance of different models. It should be noted that in the formula, Recall and Precision have the same effect on the F1 Score; however, it is mentioned that the Recall score is more important than Precision, and for final evaluation, Recall should be considered the most important metric.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

4.4. Model stability evaluation

Besides the models' different scores, another important evaluation parameter is model stability. If a model in one training turn generates a satisfactory result but, in another turn, generates an unacceptable one, it is not reliable. Therefore, to measure the model's stability and avoid the overfitting issue, The *cross-validation* method is used, and the *average performance* (Eq. (11)) and the *standard deviation of the performance* (Eq. (12)) of results are calculated. This way, the model's stability can be understood. The N-Fold cross-validation splits the dataset into N segments, and each training iteration takes one segment as the test set and the rest as the training set. Then the model is trained and evaluated. This process continues for N iterations. In each iteration, the performance of the model is stored. At the end of the process, the performance average and standard deviation are calculated as the final model performance score.

$$\mu_{Performance} = \frac{1}{n}\sum_{i=1}^{n} Performance_i \tag{11}$$

$$\sigma_{Performance} = \sqrt{\frac{\sum_{i=1}^{n}(Performance_i - \mu_{Performance})^2}{N}} \tag{12}$$

4.5. ROC curve

The receiver operating characteristic (ROC) curve determines the ability of a classifier to diagnose binary classes. It uses the true positive rate (TPR) (Eq. (13)) against the false positive rate (FPR) (Eq. (14)).

$$TPR = \frac{True\ Positive}{Total\ Positive} \tag{13}$$

$$FPR = \frac{False\ Positive}{Total\ Negative} \tag{14}$$

The area under the curve (Eq. (15)) shows the performance of the binary classifier regardless of the actual instance distribution. In other

words, it calculates the probability of two classes in which $X_1 > X_2$.

$$\begin{aligned} A &= \int_{x=0}^{\infty} TPR(FPR^{-1}(x))dx \\ &= \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T)f_0(T)dT'dT = P(X_1 > X_0) \end{aligned} \tag{15}$$

5. Experimental setup and data preprocessing

A Python implementation of the stacking ensemble is offered by the scikit-learn library. Next, a discussion on the dataset, preprocessing stages, and performance evaluation are provided.

5.1. Data sources

Unlike other publicly available datasets in clinical studies related to heart failure and cardiovascular disease, it was decided to use a much more comprehensive dataset that is not publicly available. To get access to this dataset, a consent form was submitted and approved by the MIT Laboratory for Computational Physiology, after which permission was granted to access the dataset. The dataset was created as part of a study that extracted data from electronic healthcare records of patients admitted due to heart failure at the Zigong Fourth People's Hospital between 2016 and 2019 [29,30]. In total, personal and clinical and biomedical information on 2008 patients with various types of heart failure was collected on the day of hospital admission, as well as three mandatory follow-up visits after 28 days, 3 months and six months to record patients' status or mortality. The information recorded for each patient included personal information such as age, gender, weight, height, and occupation, as well as clinical characteristics such as respiratory rate, diabetes, liver or kidney disease, dementia and other heart failure-related characteristics such as pulse, systolic and diastolic blood pressure, type of heart failure, white and red blood cell counts, etc. [29,30]. To ensure completeness of data, other categorizations and findings such as "NYHA cardiac function, Killip Grade, Glasgow Coma Scale (GCS), echocardiographic findings, left ventricular ejection fraction (LVEF), left ventricular end-diastolic diameter, mitral valve peak E wave velocity (m/s), mitral valve peak A wave velocity (m/s), E/A, tricuspid valve regurgitation velocity, and tricuspid valve regurgitation pressure" were captured as well during follow up visits [29,30]. In addition to medical data, information such as the number of visits, ward of initial admission and total readmissions, and the date and reason for the patient's release were also recorded. Overall, the dataset includes 166 primarily numerical features, although some categorical data were recognized after an initial inspection.

5.2. Class labels creation

This paper aims to develop a model that could classify patients suffering from heart failure into two categories, Emergency and Non-Emergency, based on their clinical characteristics recorded upon readmission to the hospital. After evaluating the dataset, it was discovered that this information is not classified for each patient, so it was necessary to create a new feature called **Emergency** by combining seven features, which were available for all patients. As a result, all patients were marked as **Non-Emergency** unless they met any of the following conditions: being admitted to the Emergency ward upon readmittance, having returned to the emergency department within six months of being discharged, having died during hospitalization, having died within 28 days, three months, or six months, or having been discharged to a morgue for burial. After completing this classification, 1392 patients were labeled as emergency cases, and the remaining 616 were labeled as non-emergency.

5.3. Data cleaning

As the first step in preparing the dataset for creating a machine learning model, the seven features used to create the new Emergency feature and two additional features, patient identifiers, were dropped. Next, each feature's number and percentage of missing values were calculated. In total, out of 157 features, 114 features had missing values of various degrees. Forty-three features had more than 50% missing values, which were dropped as it was decided there was no way to reliably impute these missing values without jeopardizing the validity of the dataset. In addition, 55 features had less than 5% missing values, so it was decided to remove these missing values. The remaining 16 features, which had between 5% and 50% missing values, had their missing elements replaced with the median of the feature. Next, features were renamed to improve the readability and usability of the dataset, as some features had random and additional characters. Finally, features were analyzed and placed into two categories of continuous and categorical. As a result, 77 features were marked as continuous, and the remaining 37 as categorical.

5.4. Correlation analysis

A feature-to-feature correlation matrix is generated, highlighting features with a more than 30% correlation coefficient. We have observed several highly correlated (negative and positive) features, which is understandable and expected since most features related to a patient's clinical characteristics are biologically related. Finally, a feature-to-class correlation is calculated, which shows the degree of correlation of all features to the Emergency class; some features are highly correlated, positive, and negative.

5.5. Data preprocessing

In the data preprocessing phase, every continuous feature was evaluated for the presence of outliers. As was observed previously, a few features seemed to have outliers. It was decided to use 1% and 99% as the lower and upper quantiles for determining and removing outliers from each feature. Using 5% and 95% eliminated 90% of the dataset, making it unusable. It is important to note that the main reason behind this massive elimination of elements is that elements with values as outliers in one feature do not necessarily represent outliers across all features. Although features for most of the non-emergency patients fall within the normal distribution plot of most features, patients who are considered to have emergency status usually have an unusually high or low value in one or more features, which is why it was decided to keep the outlier boundaries to 1% and 99%. We have observed that some features are more normally distributed after removing outliers. In addition, the cleaned dataset was analyzed, and the feature "AIDS" had to be dropped since it only had one category for all the remaining patient records after outlier removal. Next, all continuous features were normalized to a range of between 0 and 1 since the range of each feature was widely different, which would have required unnecessarily high computational capacity for building the model. As the final step in the preprocessing phase, all categorical variables were converted into dummy variables (with the first element dropped to avoid falling into the multicollinearity trap). The final correlation matrix shows some features are still highly correlated with each other, which will be handled in the next section.

5.6. Feature selection

After preprocessing, there are 719 elements left in the dataset, with 113 features represented. However, to select the best set of statistically significant features in explaining the Emergency class, the features must pass through correlation elimination and nested cross-validation. In the correlation elimination step, highly correlated features are removed

Table 3
Accuracy of the SEL at different l = 2:6 with various meta-learners.

	SVM	LR	MLP	XGBoost
l = 2	LR, KNN 80.91%	KNN, RF 81.70%	LR, KNN 78.30%	LR, SVM 82.47%
l = 3	LR, KNN, RF 81.75%	KNN, RF, LR 82.46%	LR, KNN, SVM 80.93%	LR, SVM, RF 83.79%
l = 4	LR, KNN, RF, SVM 82.20%	KNN, RF, LR, SVM 83.90%	LR, KNN, SVM, DT 82.46%	LR, SVM, RF, KNN 85.54%
l = 5	LR, KNN, RF, SVM, DT 84.11%	KNN, RF, LR, SVM, DT 85.34%	LR, KNN, SVM, DT, RF 83.86%	LR, SVM, RF, KNN, DT 86.98%
l = 6	**LR, KNN, RF, SVM, DT, XGBoost 86.37%**	**KNN, RF, LR, SVM, DT, XGBoost 86.54%**	**LR, KNN, SVM, DT, RF, XGBoost 84.12%**	**LR, SVM, RF, KNN, DT, XGBoost 88.23%**

until a final set of features with a correlation coefficient of less than 40% is left. This step reduces the number of features to 74. First, the dataset is split into 80% for training and 20% for testing using 5-fold cross-validation. Next, nested cross-validation (NCV) is employed within hyperparameter tunning to evaluate the importance and significance of the remaining features of the dataset. The NCV process has been carried out in the following stages (1) a feature ranking is performed using the Relieff score, (2) using the training set and looping over the features, starting with the least important features to the most important feature with 10-fold cross-validation and hyperparameter tuning, and (3) the best number of features and hyperparameters are chosen based on the minimum misclassification rate. This process produces a list of 9 features that are statistically significant and can be used in building the machine learning model. These features are: Sys.BP (systolic.blood.pressure), NeutrophilRatio (neutrophil.ratio), ThrombinTime (thrombin.time), LactateDehdro (lactate.dehydrogenase), AdmissionWard_GeneralWard, Typ2RespFail_TypeII (type.II. respiratory. failure), ReAdmin6Mnts_1 (re.admission.within.6.months), and AgeCategory_(89,110). As the final step, the data balance of the Emergency class was checked, and it was determined that the class was not balanced, as the number of records marked as an emergency was twice as much as non-emergency records. Therefore, it was decided to resample the dataset by up-sampling the minority class, which resulted in a balanced dataset without changing the shape of any of the features within the dataset. Fig. 2a and b show the pair plots of the chosen nine features with the Emergency class represented by the blue and red colors (0 and 1, respectively).

Fig. 2a shows the unbalanced dataset, whereas Fig. 2b shows the balanced dataset. As seen from these plots on their diagonal line, the number of class 0 is almost half of class 1 across all features in Fig. 2a. In contrast, they are of similar sizes in Fig. 2b. It is also important to note that the process of up-sampling the dataset has not changed the shape of features, as can be observed by comparing the upper and lower halves of both plots.

5.7. The conceptual model

Based on the features that were determined in the previous section, the following conceptual model can be formulated:

$$f^* = \beta_0+\beta_1 F_1+\beta_2 F_2+\beta_3 F_3+\beta_4 F_4+\beta_5 F_5+\beta_6 F_6+\beta_7 F_7+\beta_8 F_8+\beta_9 F_9. \quad (16)$$

F_1 to F_9 are: Sys.BP (systolic.blood.pressure), NeutrophilRatio (neutrophil.ratio), ThrombinTime (thrombin.time), LactateDehdro (lactate.dehydrogenase), AdmissionWard_GeneralWard, Typ2RespFail_TypeII (type.II.respiratory.failure), ReAdmin6Mnts_1 (re.admission.within.6.months), and AgeCategory_(89,110).

6. Experimental analysis and results

Three key performance indicators are obtained from the models to compare the results. The key indicators are the Accuracy, Recall, F1 score, and AUC of the ROC curve used in [31–35] of the models. A recall is a much more important indicator in this paper because Recall depends on the False Negative (FN) results. A Low FN rate translates to a high Recall value. Therefore, the best machine learning model for emergency case prediction should maximize Recall value and provide high accuracy. F1 score helps to indicate both Precision and Recall matrices. Minimal FN and FP classification results give a higher F1 score.

6.1. Experimental setup and meta learner selection

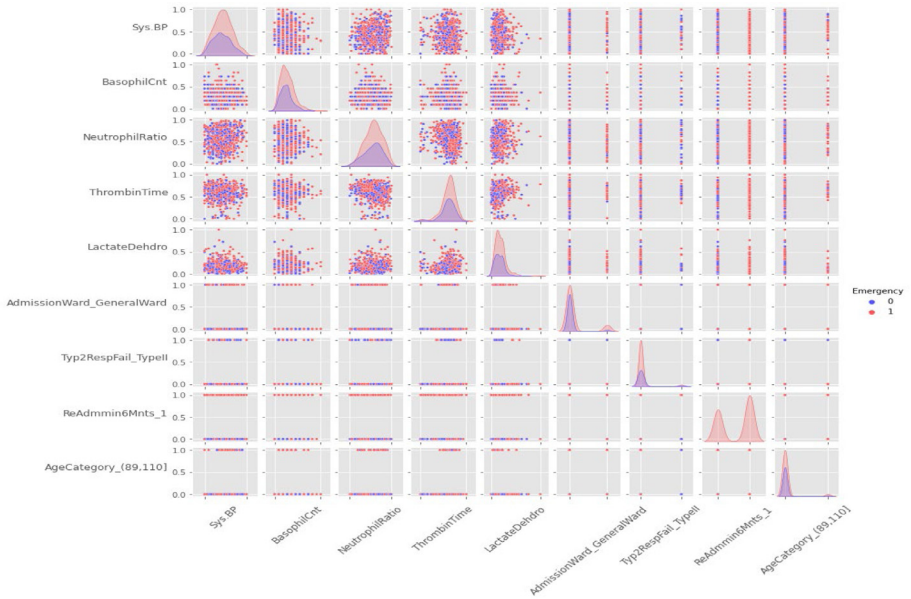
Assume a set of candidate baseline models S1 = {LR, KNN, DT, RF, SVM, and XGboost} at level-0. At level-1, assume a set of meta-learners, S2 = {SVM, LR, MLP, XGBoost}. Assume l is the number of selected baseline models at level-0 such that l = 2:6. For example, when l = 2, two baseline models are chosen from the set S1, l = 3 means that three baseline models are selected, etc. The following experiments have been conducted to validate the section of the adopted models at level-0 and level-1. In Table 3, we evaluated the accuracy of the ESL model using a variable set of baseline models at level-0 with different meta-learners at level-1. In each cell, the best-performing baselines are reported with their performance accuracy. For example, at l = 2, the best-performing pair-wise aggregate is {LR, SVM} with an accuracy of 82.47%, and at l = 4, the best-performing 4-tuple aggregate is {LR, SVM, RF, KNN} with a performance accuracy of 85.54%.

For an aggregate of size l = 6, we can observe that adding a boosting classifier (i.e., XGBoost) at level-0 increases the accuracy of the ESL using different meta-learners at level-1, as XGBoost uses gradient boosting to improve performance and reduce overfitting. In addition, at l = 6, we can observe an improvement in the accuracy of up to 88.23% with the XGBoost as a meta-learner in level-1. We can thus perceive that adding boosting at level-0 and level-1 enhances the performance of the SEL while handing overfitting and imbalanced classes.

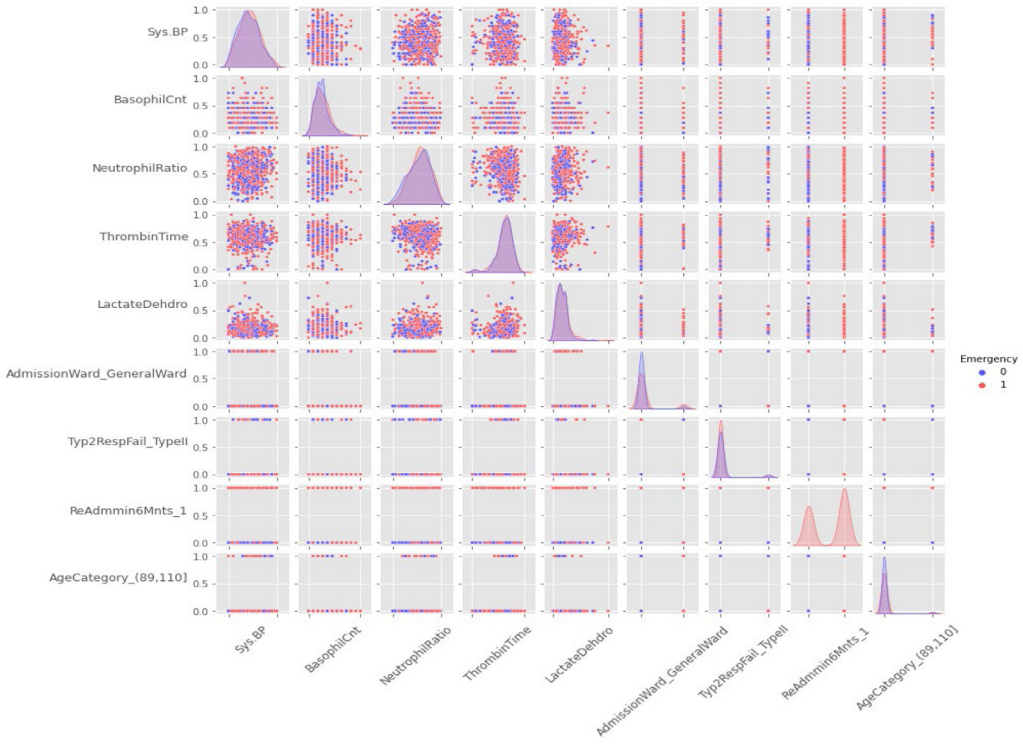
We can see from Table 3 that the XGBoost is the best-performing meta-learner with a set of six baseline models using LR, SVM, RF, KNN, DT, and XGBoost. Based on the above results, in the next experiments, the architecture of the ESL comprises the XGBoost as the meta-learner with LR, SVM, RF, KNN, DT, and XGBoost as baseline models in level-0.

6.2. N-fold cross validation

To test the chance of overfitting, we have evaluated the performance of the ESL with various segment ratios of training, validation, and testing sets using 5-fold and 10-fold, with 60%, 20%, and 20% training/validation/testing split, as shown in Table 4. The N-fold *cross-validation* method splits the dataset into N segments, and each training iteration takes one segment as the test set and the rest as the training set. Then the model is trained and evaluated. This process continues for N iterations. In each iteration, the performance of the model is stored. At the end of the process, the average accuracy and standard deviation are calculated as the model performance score. This way, the model's stability and overfitting can be explained.



(a) Unbalanced Pair Plot of Final Features



(b) Balanced (bottom) Pair Plot of Final Features

Fig. 2. The pair plot of final features. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

From Table 4, at 5-fold and 10-fold, comparing the training, validation, and testing accuracy, the SEL model is robust with no overfitting. Using six classifiers at level 0, followed by XGBoost as a meta-learner at level 1, helps to improve the accuracy and robustness of the overall prediction model while minimizing overfitting.

Next, we compared the performance of the ESL model to various machine learning algorithms such as Linear Regression (LR), K-nearest Neighbor (KNN), Decision Trees (DT), Support Vector Machines (SVM), Naïve Bayesian (NB), stochastic gradient descent (SGD), and traditional ensemble models such as Random Forest (RF), and XGBoost. To better show the performance of the proposed ESL, we have assessed its performance against the state-of-the-art ensemble-based classifiers using Bagging, Boosting, AdaBoost and Majority [36,37].

Table 4
Accuracy of the SEL using variable N-fold cross-validation.

	5-fold	10-fold
Training accuracy	87.7% ± 0.15	89.4% ± 0.27
Validation accuracy	86.6% ± 0.66	87.5% ± 0.63
Testing accuracy	87.1% ± 0.35	88.3% ± 0.80

6.3. Single-based models vs. the ESL model

Fig. 3 shows the comparisons of the models across all metrics. The proposed ESL stacking model offers high accuracy with less fluctuation than other models. The average accuracy of the stacking model is 88%.

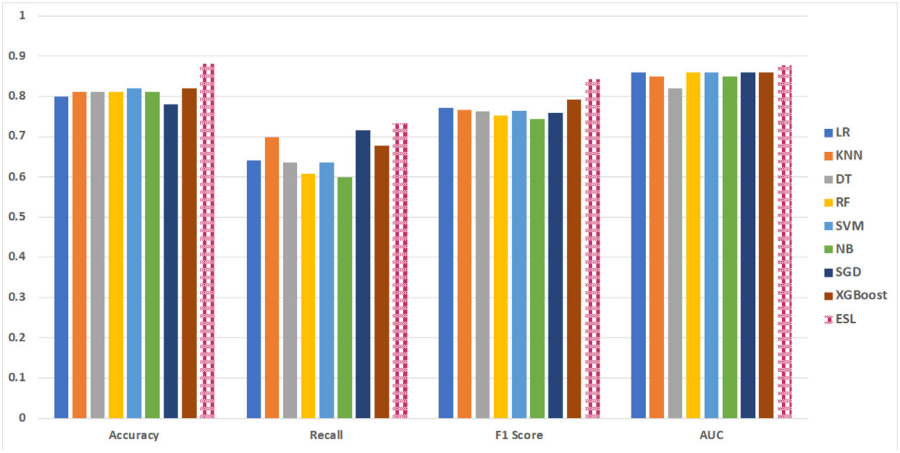


Fig. 3. Performance evaluation: Baseline techniques Vs. the proposed Stacking ESL model.

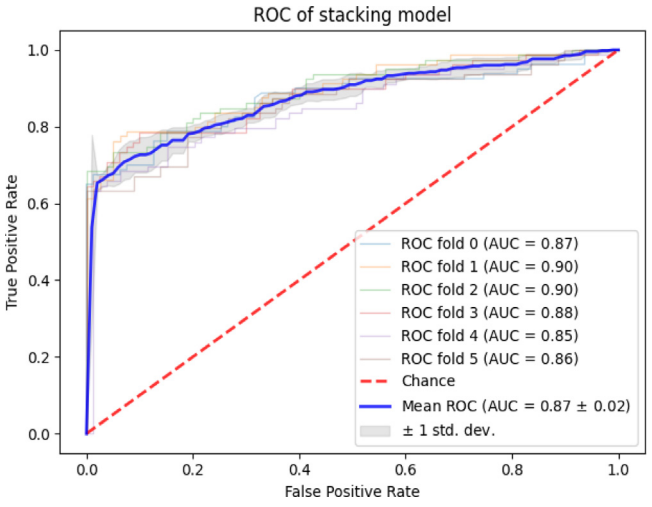


Fig. 4. The ROC curve (5-fold): The ESL model.

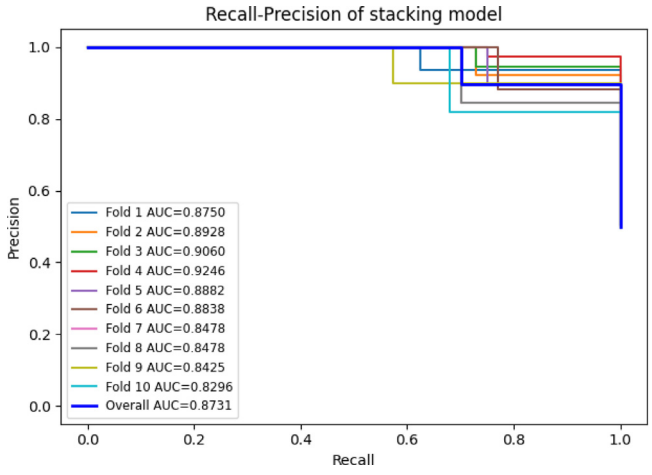


Fig. 5. Recall–Precision curve (10-fold): The ESL model.

The Recall of the proposed stacking model outperforms all the other models. The Recall value of the stacking model is around 74%. The Stacking model provides high F1 scores, around 84%. The proposed Stacking model offers higher accuracy, Recall, and F1 score than other baseline models. In addition to the key indicators, we also obtained the ROC and Recall–Precision curves from the model evaluation program to see the performance of the stacking model, as shown in the Figs. 3 and 4. Fig. 4 shows the AUC of the ESL model with stable results across multiple folds. Fig. 5 shows the Recall–Precision curve of the Stacking model. The overall AUC of the average Recall–Precision curve of the stacking model is around 88%.

6.4. Ensemble-based models vs. the ESL model

Bagging involves creating multiple subsets of the original dataset and training a base model on each subset. The predictions of the base models are then combined to form the final prediction. Boosting is responsible for training multiple weak models sequentially, with each subsequent model attempting to correct the errors of the previous model. The predictions of the weak models are then combined to form the final prediction. AdaBoost combines boosting with weighted data. The algorithm assigns weights to each data point, with more weight given to data points that are difficult to classify. The base model is then trained on the weighted data, and subsequent models are trained to correct the errors of the previous models. The predictions of the models are then combined to form the final prediction. The Voting ensemble trains multiple base models on the same dataset, and the predictions of the models are then combined using either a majority voting approach or a weighted voting approach. In this subsection, we have compared the performance of the ESL against Bagging, Boosting, AdaBoost, and Voting, as illustrated in Fig. 6. As observed in Fig. 6, the proposed ESL model outperforms state-of-the-art ensembles across all evaluation metrics. In addition, the model's accuracy is stable across multiple folds and trails, which shows the robustness of the model compared to existing ensembles.

7. Discussion

From a medical perspective, accurate and reliable prediction of heart-related diseases and medical conditions is crucial for early diagnosis and timely treatment. Machine learning models have become essential in medical research and clinical practice, providing valuable insights into diagnosing, prognosis, and treating various heart-related diseases. In recent years, there has been a growing interest in developing ensemble models that combine multiple machine learning algorithms to improve the accuracy and reliability of disease prediction. This paper analyzed various classification techniques and found that the proposed SEL stacking model achieved the highest prediction performance compared to the baseline models such as logistic regression, k-nearest neighbor, decision tree, random forest, support vector machines, bagging, Boosting Adaboost, and Voting. This result

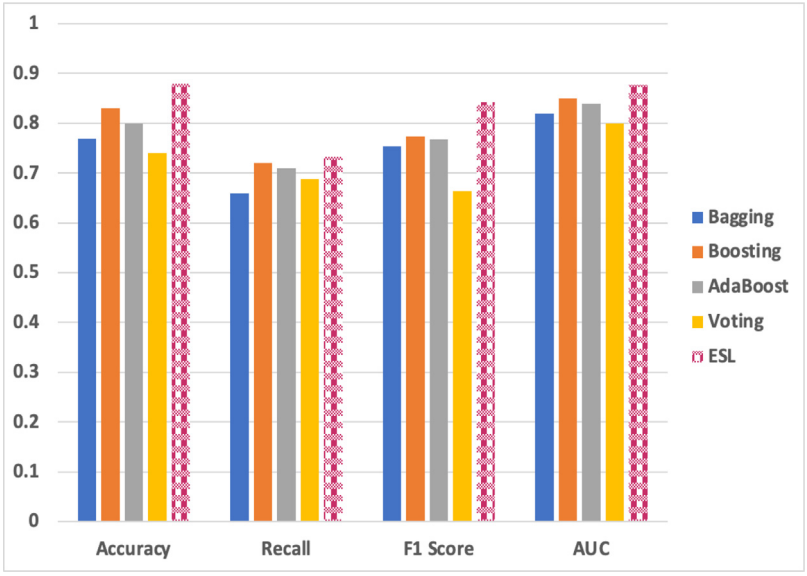


Fig. 6. The ESL model Vs. State-of-the-art ensembles.

indicates that the stacking model can effectively identify patterns and relationships among different data features to make accurate predictions. In medical applications, accurate prediction of disease risk can help healthcare professionals make informed decisions about patient care, leading to better health outcomes. To ensure robustness and high accuracy in the prediction results across multiple runs, we used XGBoost as a meta-learner in the stacking ensemble model with efficient hyperparameter tuning. XGBoost is a powerful and widely used ensemble machine learning algorithm that is known for its high accuracy and computational efficiency. Using XGBoost as a meta-learner, the study achieved comparable performance to baseline XGBoost itself, indicating that a stacking ensemble model is an effective approach for disease prediction. We found that the stacked ensemble classifier provides various advantages over individual and existing ensemble models, including enhanced accuracy and decreased overfitting, resulting in improving the generalization performance of the model. Moreover, the stacking ensemble model can accommodate various types of data and models, making it a versatile approach that can be applied to different medical applications. However, we noted that the heightened computational complexity resulting from training multiple models could be a limitation of the stacking model. In medical applications, where time and computational resources are often limited, this may pose a challenge. Therefore, further research is needed to optimize the computational efficiency of the stacking model in clinical settings.

8. Conclusion and future directions

Stacked ensembles are robust and flexible machine-learning methods that can provide more steady predictions. This paper proposes a two-level stacking ensemble. Based on the experimental results, we conclude that the proposed stacking model is a potentially more robust approach than the individual model approach to classify emergency cases from non-emergency cases. In this work, we worked with a new unlabeled data set which did not have the class attribute(label); thus, a novel feature integration technique was employed to classify the data set records with "emergency" cases from "non-emergency" case labels. The developed model is in a testing trial mode. Once we generalize it with more classifiers using deep learning over multiple datasets, we plan to move forward to a production-ready model in clinical practice. Future directions to implement the developed model with different geographical locations that would include more people with diverse demographics. While the stacking ensemble model has shown promise in various medical applications, the heightened computational complexity resulting from training multiple models can be a significant limitation, particularly in clinical settings where time and computational resources are often limited. To address this limitation, future research could explore using transfer learning techniques to leverage pre-trained models and reduce the required training data. In addition to optimizing the computational efficiency of the stacking model, future research could also explore the use of more diverse and heterogeneous models in the ensemble. The stacking model's performance depends on the choice of base models. Incorporating a more comprehensive range of models can improve the model's ability to capture complex patterns in the data. Another potential future direction is the integration of other data sources, such as genomics, proteomics, and electronic health records, to further improve the accuracy and reliability of disease prediction. The stacking model can capture a more comprehensive picture of the patient's health status and improve disease prediction and diagnosis by incorporating multiple data sources. In addition to heart-related diagnoses, ESL can be applied for COVID-related diagnoses, similar to the work in [38–40]. Future direction would involve adopting deep learning [41,42] to build a more reliable multi-dataset diagnosis model using additional metrics such as hamming loss, Jaccard score and cross-entropy loss.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

This research is Funded by Toronto Metopolian University, Faculty of Engineering and Architectural Science.

References

[1] M. Motwani, D. Dey, D.S. Berman, G. Germano, S. Achenbach, M.H. Al-Mallah …, P. Slomka, Machine learning for predicting all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis, Eur. Heart J. 40 (19) (2019) 1451–1458.

[2] Z.I. Attia, S. Kapa, X. Yao, F. Lopez-Jimenez, T.L. Mohan, P.A. Pellikka ., P.A. Noseworthy, Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction, JAMA Cardiol. 4 (7) (2019) 577–584.

[3] H. Cho, H. Lee, S. Kim, S. Lee, K.H. Kim, Machine learning-based predictive model for acute myocardial infarction using electronic health records, Int. J. Med. Inform. 128 (2019) 47–53.

[4] A.M. Alaa, T. Bolton, E. Di Angelantonio, J.H. Rudd, M. van der Schaar, Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423, 604 UK biobank participants, PLoS One 14 (5) (2019) e0213653.

[5] L. Saba, M. Biswas, V. Kuppili, E.C. Godia, H.S. Suri, D.R. Edla ., J.R. Laird, The role of machine learning in cardiac imaging, including structure, function, and electrophysiology, IEEE Trans. Biomed. Eng. 67 (10) (2020) 2664–2692.

[6] Muhammad Yasin, Temesghen Tekeste, Hani Saleh, Baker Mohammad, Ozgur Sinanoglu, Mohammed Ismail, Ultra-low power, secure IoT platform for predicting cardiovascular diseases, IEEE Trans. Circuits Syst. I. Regul. Pap. 64 (9) (2017) 2624–2637.

[7] Hager Ahmed, Eman MG. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali, Heart disease identification from patients' social posts, machine learning solution on Spark, Future Gener. Comput. Syst. 111 (2020) 714–722.

[8] Kumar, Priyan Malarvizhi, Usha Devi Gandhi, A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases, Comput. Electr. Eng. 65 (2018) 222–235.

[9] Farman Ali, Shaker El-Sappagh, SM. Riazul Islam, Daehan Kwak, Amjad Ali, Muhammad Imran, Kyung-Sup Kwak, A smart healthcare monitoring system for heart disease prediction based on deep ensemble learning and feature fusion, Inf. Fusion 63 (2020) 208–222.

[10] Prerna Sharma, Krishna Choudhary, Kshitij Gupta, Rahul Chawla, Deepak Gupta, Arun Sharma, Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine learning, Artif. Intell. Med. 102 (2020) 101752.

[11] Subhashini Narayan, E. Sathiyamoorthy, A novel recommender system based on FFT with machine learning for predicting and identifying heart diseases, Neural Comput. Appl. 31 (1) (2019) 93–102.

[12] L. Guo, Y. Zhang, G. Sun, A machine learning-based model for predicting cardiovascular disease risk, IEEE Trans. NanoBiosci. 19 (2) (2020) 250–256.

[13] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, T. Kitai, Artificial intelligence in precision cardiovascular medicine, J. Am. Coll. Cardiol. 69 (21) (2019) 2657–2664.

[14] C.H. Lin, H.I. Lin, J.S. Lai, Deep learning models for predicting 30-day readmission in patients with heart failure, J. Med. Syst. 44 (10) (2020) 176.

[15] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358.

[16] H. Shao, Y. Li, Q. Li, Y. Yang, H. Li, Y. Tang, Machine learning models for predicting cardiovascular events in patients with hypertension, J. Med. Syst. 44 (5) (2020) 92.

[17] Y. Wang, Y. Wang, G. Xu, Y. Liu, X. Liu, Machine learning-based prediction models for cardiovascular diseases using health examination data, J. Med. Syst. 45 (2) (2021) 1–10.

[18] Y. Zhao, Y. Zhang, Q. Shi, Y. Fan, Y. Chen, A machine learning-based model for predicting acute myocardial infarction, Int. J. Med. Inform. 141 (2020) 104146.

[19] Y. Zhou, Y. Liao, Z. Wu, A novel machine learning-based model for predicting all-cause mortality in patients with acute myocardial infarction, Int. J. Med. Inform. 143 (2020) 104260.

[20] H. Zhu, Y. Wang, X. Chen, W. Chen, Machine learning-based prediction models for the development of heart failure in patients with hypertension, J. Med. Syst. 45 (3) (2021) 1–10.

[21] Y. Zuo, J. Lin, Y. Lei, Machine learning models for predicting clinical outcomes in patients with coronary artery disease: A systematic review, Int. J. Med. Inform. 146 (2021) 104358.

[22] Rong Tao, Shulin Zhang, Xiao Huang, Minfang Tao, Jian Ma, Shixin Ma, Chaoxiang Zhang, et al., Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods, IEEE Trans. Biomed. Eng. 66 (6) (2018) 1658–1667.

[23] Rahma Atallah, Amjed Al-Mousa, Heart disease detection using machine learning majority voting ensemble method, in: 2019 2nd International Conference on New Trends in Computing Sciences, ICTCS, IEEE, 2019, pp. 1–6.

[24] M. Asif, U. Wajid, Heart disease prediction using ensemble learning and feature selection techniques, Healthc. Inform. Res. 26 (4) (2020) 279–289.

[25] P.Y. Chang, J.Y. Hsu, M.H. Wang, Ensemble machine learning for cardiovascular disease prediction, Comput. Biol. Med. 96 (2018) 120–126.

[26] N.C. Deepika, S.S. Prabhu, Ensemble learning techniques for cardiovascular disease prediction, Procedia Comput. Sci. 165 (2019) 14–21.

[27] S. El-Sappagh, A.A. Hendawi, M. Elmogy, Ensemble of machine learning algorithms for heart disease diagnosis, J. Med. Syst. 43 (9) (2019) 1–11.

[28] X. Jia, Y. Zhang, Y. Chen, Y. Wang, Machine learning models for predicting cardiovascular disease in Chinese patients, J. Med. Syst. 44 (6) (2020) 1–8.

[29] Zhongheng Zhang, Yan Zhao, Linghong Cao, Ziyin Xu, Rangui Chen, Lukai Lv, Ping Xu, Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data (version 1.1), PhysioNet (2020) http://dx.doi.org/10.13026/q712-yv80.

[30] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation [Online] 101 (23) (2000) e215–e220.

[31] R. Kashef, A. Niranjan, Handling large-scale data using two-tier hierarchical super-peer P2P network, in: Proceedings of the International Conference on Big Data and Internet of Things, 2017, pp. 52–56.

[32] G. Hass, P. Simon, R. Kashef, Business applications for current developments in big data clustering: an overview, in: 2020 IEEE International Conference on Industrial Engineering and Engineering Management, IEEM, IEEE, 2020, pp. 195–199.

[33] R. Kashef, ECNN: Enhanced convolutional neural network for efficient diagnosis of autism spectrum disorder, Cogn. Syst. Res. 71 (2022) 41–49.

[34] S.A.A. Shah, A.H. Saleh, M. Ebrahimian, R. Kashef, Early detection of heart disease using advances of machine learning for large-scale patient datasets, in: 2022 IEEE Canadian Conference on Electrical and Computer Engineering, CCECE, IEEE, 2022, pp. 274–280.

[35] P. Vajar, A.L. Emmanuel, A. Ghasemieh, P. Bahrami, R. Kashef, The internet of medical things (IoMT): A vision on learning, privacy, and computing, in: 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME, IEEE, 2021, pp. 1–7.

[36] N. Razfar, R. Kashef, F. Mohammadi, Assessing stroke patients movements using inertial measurements through the advances of ensemble learning technology, in: 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), IEEE, 2021, pp. 1482–1489.

[37] K. Chadaga, S. Prabhu, N. Sampathila, R. Chadaga, S. KS, S. Sengupta, Predicting cervical cancer biopsy results using demographic and epidemiological parameters: a custom stacked ensemble machine learning approach, Cogent Eng. 9 (1) (2022) 2143040.

[38] A. Pradhan, S. Prabhu, K. Chadaga, S. Sengupta, G. Nath, Supervised learning models for the preliminary detection of COVID-19 in patients using demographic and epidemiological parameters, Information 13 (7) (2022) 330.

[39] K. Chadaga, C. Chakraborty, S. Prabhu, S. Umakanth, V. Bhat, N. Sampathila, Clinical and laboratory approach to diagnose COVID-19 using machine learning, Interdiscip. Sci.: Comput. Life Sci. 14 (2) (2022) 452–470.

[40] K. Chadaga, S. Prabhu, V. Bhat, N. Sampathila, S. Umakanth, R. Chadaga, COVID-19 mortality prediction using machine learning: A deep forest approach, in: 2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics, DISCOVER, IEEE, 2022, pp. 245–250.

[41] M. Woźniak, M. Wieczorek, J. Siłka, Bilstm deep neural network model for imbalanced medical data of IoT systems, Future Gener. Comput. Syst. 141 (2023) 489–499.

[42] Z.A. Shirazi, C.P. de Souza, R. Kashef, F.F. Rodrigues, Deep learning in the healthcare industry: theory and applications, in: Computational Intelligence and Soft Computing Applications in Healthcare Management Science, IGI Global, 2020, pp. 220–245.