STM JOURNALS

International Journal of
Bioinformatics and Computational Biology

ISSN: 2583-9977
Volume 4, Issue 2, 2026
July—December
DOI (Journal): 10.37591/IJBCB

https://journals.stmjournals.com/ijbcb

Review IJBCB

A Lightweight Cost-Sensitive Explainable Ensemble Framework for Early Heart Disease Risk Prediction

Satyam Shinde[1,*], Nikhilesh Mankar[2], Amrut Nikam[1], Swati Shirke[3], Rahul Sonkambale[3]

Abstract

Cardiovascular disease is still one of the leading causes of death, and hence, the early prediction of risk is a very important task in preventive medicine. Although recent studies have shown encouraging results in the application of machine learning algorithms to the prediction of heart disease, it has been noticed that most of the algorithms are more concerned with accuracy-driven optimization than the concerns of safety and false negatives. In medical decision support systems, false negatives are more harmful. This paper presents a light-weight and interpretable machine learning approach for the early risk prediction of heart disease based on structured clinical data. Various models, such as Logistic Regression, Random Forest, XGBoost, and stacking ensemble classifiers, are compared based on clinically meaningful evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The experimental results indicate that ensemble classifiers perform better than individual models, and the unoptimized StackingClassifier performs the best (Recall: 0.8807, F1-score: 0.8930, AUC: 0.9147). Cost-sensitive and threshold-optimized stacking further enhances the recall to 0.9266. To improve transparency and clinical trust, SHAP and LIME are combined to offer global and local explanations. The findings point out ST depression, maximum heart rate reached, type of chest pain, cholesterol, and exercise-induced angina as the important risk factors. The proposed approach shows that simple and interpretable ensemble models can provide accurate heart disease risk predictions.

Keywords: Heart disease prediction, explainable artificial intelligence, ensemble learning, stacking classifier, cost-sensitive learning, threshold optimization, SHAP, LIME

INTRODUCTION

Heart disease is one of the oldest diseases that have continued to pose challenges to the modern healthcare systems and have ever been a major cause of deaths. However, despite all the progress that have been experienced within the last several decades of advances in science and technology, especially in the field of diagnostic technology, there still remain several cardiac ailments that can only be detected at relatively advanced stages of the illness. This is especially in the case where the patient does not have visible and, most crucially, severe manifestations of early phases of illness onset, in which case, unfortunately, success in advanced care can be greatly impeded. Therefore, there remains to this day a major thrust in preventive medical care to detect individuals at risk of cardiac illness [1–3].

Cardiac risk evaluation has been the domain of the doctor's expertise, certain standardized criteria for diagnosing the condition, and certain scoring algorithms for risk calculations [4] involving certain well-verified model parameters, such as the presence

***Author for Correspondence**
Satyam Shinde
E-mail: satyam.shinde24@pcu.edu.in

[1]Research Scholar, Department of CSE, Pimpri Chinchwad University, Pune, Maharashtra, India
[2]Professor, Department of CSE, Pimpri Chinchwad University, Pune, Maharashtra, India
[3]Professor, Department of CSE-AIML, Pimpri Chinchwad University, Pune, Maharashtra, India

Received Date: May 11, 2026
Accepted Date: June 06, 2026
Published Date: June 16, 2026

Citation: Satyam Shinde, Nikhilesh Mankar, Amrut Nikam, Swati Shirke, Rahul Sonkambale. A Lightweight Cost-Sensitive Explainable Ensemble Framework for Early Heart Disease Risk Prediction. International Journal of Bioinformatics and Computational Biology. 2026; 4(2): 26–40p.

of high blood pressure, high cholesterols, symptoms, and cardiac responses, to stress exercises. The advantage herein is the lack of opaqueness and the incorporation of doctor expertise, although perhaps the disadvantage is the obvious linearity model and cut-offs. Such might not be highly appropriate for the establishment of potential relationships between different parameters from the domain of the medical industry, particularly considering the likely diversity of the patients within the domain itself as well as the complexity and nature of health data.

Clinical risk prediction for heart disease, until now, has been the specialty of the doctor's expertise, some strict criteria for the diagnosis, and the statistical risk formula calculations based on some strictly verified parameters like age, blood pressure, cholesterol, symptoms, and the reaction to stress tests. The advantage in these approaches is clarity and doctor expertise, although the disadvantage could be the application limited to a linear model and some specific values. Probably, they lack accuracy in the sense that they fail to deal effectively with the non-linear correlation between different parameters, especially because, in a clinical environment, the patients could be rather varied, while the volume, complexity, and type of health-related data could be challenging in terms of the required flexibility and scalability.

However, out of these approaches, some schemes of ensemble learning have gained popularity due to their robustness and generalization abilities. Random Forests assist in handling the issue of model instability by performing majority classification among different decision trees [5], whereas gradient boosting schemes work towards trying to optimize the classification task using the difficult-to-classify examples. In some of the latest schemes of ensemble learning, stacking is based on meta-learning, which is aimed at effectively using the best possible abilities from different base learners. Many times, there have been instances where an ensemble classifier outperforms other classifiers among applicable clinical features.

However, there has been an increasing complexity of machine learning models over the past few years, which once again poses different challenges for the application of model in the healthcare industry. One of the disadvantages, which is again evident in literature, is that there is too much emphasis on accuracy as a measure of performance. In the context of cardiovascular disease, this is not a very efficient measure of performance as a classification problem. A system can still be able to predict the class of many instances correctly, which may not be able to predict the number of patients who actually have heart disease. Negative predictions of this kind can be a point of concern, as this may lead to very serious problems due to delays in detection [6].

However, there is another important issue, apart from their complexity, which would have to be dealt with in the implementation of both machine learning models and their ensemble models in the medical field, and that would be the fact that these models are not interpretable. Indeed, although models that are relatively complex may happen to be potentially very accurate, they would, on the other hand, be "black boxes," which would provide no explanation whatsoever for their predictions [7]. Therefore, it might be relatively difficult for these medical professionals to wholeheartedly trust and appreciate these models, predictions, and their potential for accuracy, although they may be potentially highly accurate and hence relevant to the medical field [8].

Explainable Artificial Intelligence may be regarded as one of the most striking fields of research that deal with such issues. In general, the aim of Explainability techniques is to raise the transparency of Machine Learning algorithms by showing which individual variables influenced the given prediction. While Global explanation models analyze which variables affect the risk of diseases most in the data as a whole, local models explain predictions individually for several cases. In the case of heart disease prediction, explainability is an efficient technique that makes validation of predictions against known medical knowledge possible [9].

Along with the issue of interpretability, the issues of computational feasibility as well as the feasibility of application also have a significant role in determining the level of usability of predictive models in real-world applications. Most of the current real-world applications are being developed

based on the concept of deep learning models. Deep learning models require a large number of computational capabilities. This implies that the application of predictive models in real-world applications may not be entirely feasible in a resource-constrained environment of a health setup. Machine learning models are more applicable for usage in real-world applications [10, 11].

One of the areas that could be investigated or not within the context of research related to the prediction of cardiovascular disease could be the inclusion of decision thresholds and the cost of error. The fact is that most classification methods utilize a probability threshold within a default probability framework that does not assign a priority to health concerns [12]. The addition of a classification threshold offers a straightforward method of managing the balance between recall and precision within the model to assign a priority to the early detection of health risks when needed. The cost-sensitive learning method also offers an improvement in terms of assigning a higher priority to the health cost of a missed disease case due to a false negative error [13].

To overcome these limitations, in this research, a light-weight interpretable machine learning system to support early risk prediction of heart disease based on structured health information is proposed. The experimentation is performed in a practical experiment flow indicating real-world usage and not considering developments in the domain as a competition in terms of improvement in predictive capabilities. Contrary to most research papers which suggest a single approach to counter a particular problem, several models based on a holistic approach in a machine learning system are explored to deduce a set of models that co-exist in terms of accuracy and interpretability with light-weight processing.

The proposed framework begins with the preprocessing and feature preparation process for quality data. The design of the basic models will provide a level of performance. Tree models, boost models, and stacking models designed from the concept of multiple models will also be implemented. This will assist in obtaining a fair comparison of the models with similar experimental set-up conditions.

The techniques that fall under this framework are linked to optimal thresholding in classification and cost-sensitive learning for efficient solution of the healthcare problem of “missed diagnoses minimization.” The techniques for classification allow strategic or intentional equilibrium attainment between “recall” and the minimization of false negatives concerning overall performance. There are also multi-criteria assessments of the techniques concerning “Accuracy, Precision, Recall, F1-Score, ROC AUC, and Confusion Matrix Analysis.” This provides a multivariate analysis framework that offers a comprehensive outlook of the analysis of the performance of the techniques [14].

The method of explainability is embedded directly into the methodology, such that it is not merely something that is done after completing a task to analyze data. Global explanations are used for identifying key characteristics regarding persons or patients with risk associated with heart disease, besides acquiring explanations for the predictions done.

Briefly, in this study, a useful, transparent, application-friendly machine learning framework focusing on the prediction of early-stage risks of heart disease will be presented. The proposed approach in this study has addressed directly some critical identified challenges in the related literature by using combining techniques, optimization approaches, and explainability tools, in addition to lightweight models and ensemble methods [15].

Collectively, this work contributes along various important dimensions to machine learning-based early risk prediction of heart disease. First, it presents a lightweight, structured experimental framework that systematically investigates baseline, ensemble, and stacking-based models on clinical tabular data while ensuring a fair comparison under consistent preprocessing and evaluation settings. Second, the framework explicitly focuses on clinical safety since it sets its emphasis upon enhancing recall, minimizing false negatives via techniques, such as classification threshold optimization and cost-

sensitive learning, and not merely upon an accuracy-based assessment. Third, interpretability methodologies to validate predictions based upon sound medical knowledge, setting out to prove that excellent predictivity, interpretability, and efficiency can be attained concurrently, and thus the tool will be applicable in a real-world setting [16].

LITERATURE REVIEW

Cardiovascular Disease Prediction in Preventive Healthcare

The prediction of cardiovascular disease (CVD) has recently been identified as one of the most important application areas in the field of preventive healthcare because of the high mortality rate and associated benefits of early intervention. The early work in this area was largely based on conventional statistical models and risk scoring [17]. These models were linear in assumption with respect to the clinical variables of age, cholesterol, blood pressure, heart rate, and lifestyle variables. Although these models were simple and interpretable, they were limited in their capacity to model the nonlinear interactions of the patient attributes.

With the increase in the usage of electronic health records (EHRs) and the availability of large-scale structured medical data, machine learning (ML) algorithms started replacing traditional statistical analysis. Supervised learning algorithms, like Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Random Forests, became popular for heart disease prediction problems (Figure 1).

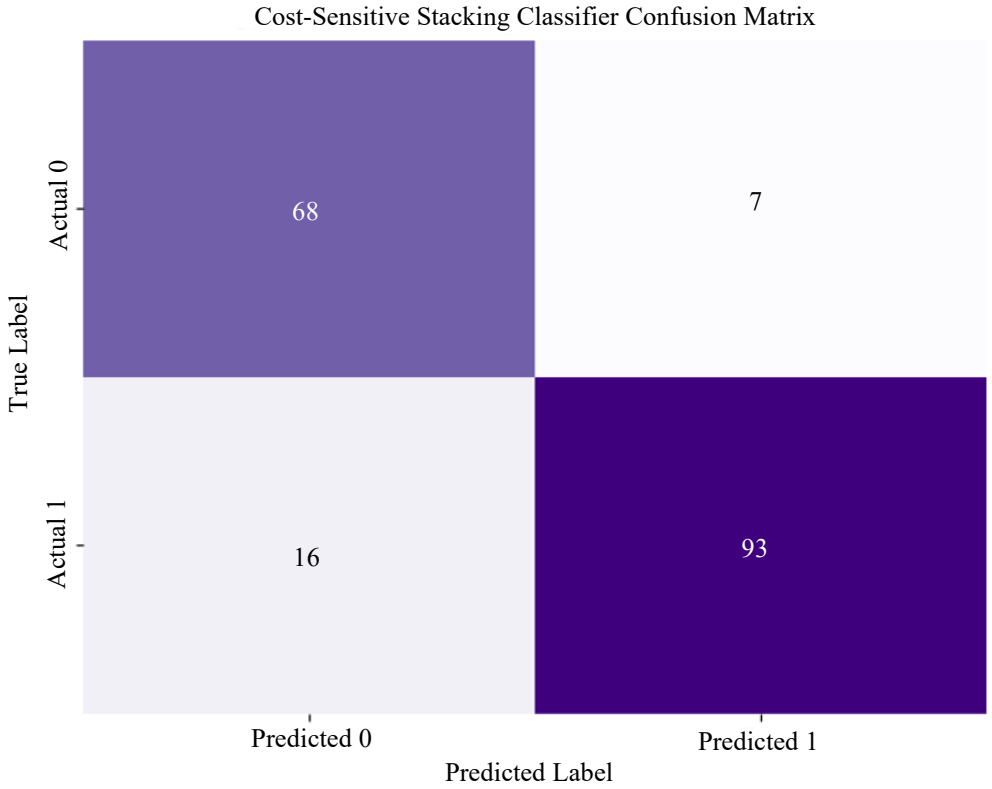

Figure 1. Confusion matrix.

Tree-Based and Ensemble Learning Approaches

To remove the limitation of linear models, tree-based classifiers and ensemble learning methods were explored in depth. Decision Trees assisted in the concept of specifying a nonlinear boundary but suffered from the problem of overfitting. Random Forests was an extension of decision tree classifiers as it was a combination of several decision tree classifiers that would add more strength, recall, and accuracy. Several research papers were prepared to prove the efficacy of Random Forest classifiers, particularly in the presence of interactions and noisy medical data. But it was also understood that over-optimization of

hyperparameters might result in the deterioration of generalization performance, thereby giving rise to a trade-off problem between hyperparameter optimization and hyperparameter robustness [17].

Later on, the development of boosting-based ensemble techniques further enhanced the accuracy of the prediction. Gradient boosting methods, especially XGBoost, were found to be useful in the prediction of cardiovascular disease due to their ability to manage complex patterns, missing values, and feature importance. In previous research, high recall and AUC measures were achieved in the prediction of heart disease using XGBoost. However, boosting techniques have been largely criticized for being black box models, which are not transparent and trustworthy regarding a medical prediction [18].

Stacking Ensemble Learning

To further improve predictions, stacking ensemble learning techniques were introduced. Stacking learns to combine predictions from several individual models by learning a meta-classifier, allowing the system to take advantage of complementary strengths of different learners. Research results show that stacking ensembles often outperform individual classifiers in terms of accuracy and AUC. However, many of the existing stacking-based approaches mainly involve optimizing performance measures, such as overall accuracy, and do not necessarily consider clinically important factors such as false negative reduction. Moreover, interpretability is not usually considered in stacking frameworks, which makes them less practical for use in healthcare applications.

Interpretability and Explainable AI in Medical Prediction

Interpretability has been found to be a crucial requirement to the application of using machine learning models in the medical field. Healthcare professionals need to understand and trust the predictions of the model before applying them to their practice. To overcome this issue, new methods called Explainable Artificial Intelligence (XAI) have been developed. SHAP (SHapley Additive exPlanations) is a method that offers global interpretability by explaining the contribution of each feature to the predictions of the model, whereas LIME (Local Interpretable Model-Agnostic Explanations) concentrates on explaining individual predictions [19].

Although several studies have applied SHAP or LIME to heart disease prediction tasks {ref6}, explainability is often treated as an auxiliary component rather than an integral part of the predictive framework. Consequently, the interpretability provided is limited and not fully aligned with clinical reasoning processes.

Shortcomings in Existing Literature

A major deficiency of current research is the lack of focus on reducing false negatives. In heart disease prediction, false negatives are high risk patients incorrectly classified as healthy, and this has potentially serious clinical implications. Despite this, many studies focus on accuracy as the main metric for evaluation without much thought on how to optimize recall or implement cost sensitive learning strategies. Additionally, a large amount of modern research is based on computationally intensive models, which assume the presence of high-performance computing infrastructure. This assumption limits the application of such models in resource constrained healthcare environments.

We Find a Gap in Research and Are Motivated by It as Follows

Along with the qualitative gaps identified by earlier researchers, the quantitative comparison between the models that became available in the literature of the recent past also reveals inconsistencies in the performance of models of the various learning paradigms. Table “model performance in existing works” – gives a summarized comparison of resultative performance of selective models used in most of the occasions in forecasting heart disease, and these are the trend lines grounded on the observations of several studies carried out.

According to the literature under review, the predictive framework that is balanced in terms of both the need to perform and the need to be interpreted and deployed practically is indeed required.

Fragmented explainability integration, lack of enough recall and reduction of false negative, less application of cost sensitive and threshold-based optimization techniques and large computational complexity of current methods were the major gaps raised.

The current paper has filled these gaps, with a lightweight ensemble-based framework that focuses on recall maximization based on cost-sensitive learning, incorporates a SHAP-driven explainability, and is computationally affordable. The proposed framework will be used to provide clinically meaningful predictions in a manner that transparency is ensured, as well as deploy ability of the system, by systematically comparing the traditional machine learning models, ensemble-based methods and stacking, longing to delivery and assist the clinics in their decision-making processes. In this context, Table 1 supports several findings in literature. The old models, like the Logistic Regression, are more interpretable and have relatively poor predictive power. Ensemble and tree-based models (Random Forest and XGBoost) are shown to be more accurate, have a higher recall, and AUC, which proves their validity in their ability to represent nonlinear relationships in medical data. Un-tuned ensemble strategies tend to are more general-purpose than more aggressive variants and tend to be robust, rather than over-optimized.

Table 1. Comparison of model performance in existing works.

Model	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression	0.7989	0.8529	0.7982	0.8246	0.8882
Untuned Random Forest	0.8587	0.9029	0.8532	0.8774	0.9108
Tuned Random Forest (Previous)	0.8315	0.8750	0.8349	0.8545	0.9023
Tuned Random Forest (Revisited)	0.8533	0.8727	0.8807	0.8767	0.9114
XGBoost	0.8750	0.9135	0.8716	0.8920	0.9105
StackingClassifier (Untuned)	0.9057	0.8750	0.8807	0.8930	0.9147
StackingClassifier (Tuned, Previous)	0.5924	0.5924	1.0000	0.7440	0.9154
StackingClassifier (Tuned, Extensive)	0.5924	0.5924	1.0000	0.7440	0.9154
StackingClassifier (Cost-Sensitive)	0.8641	0.8559	0.9266	0.8899	0.9145

It is important to note that stacking-based models have good overall performance, but tuned stacking variants have enormous recall with high accuracy which implies that they have bias in drawing decision boundaries with respect to the positive class. This conduct points to one of the typical constraints of available ensemble research where recall is enhanced without proper thresholding or cost-controlling design. This problem is addressed partially by the cost-sensitive stacking method, which enhances recall and provides balanced performance, thus the necessity to focus specifically on false negative reduction in medical prediction tasks.

Alignment with Proposed Methodology

Consideration of other research methods will be conducted within the scope of the proposed research methodology.

According to the literature review, the methodological design of the current study is informed by three key objectives: (i) enhancing recall so that to reduce falses in detection of heart diseases at an early stage, (ii) interpretability of the model by incorporating multi-explanatory techniques, and (iii) computational efficiency that the model should have so that it can be deployed in the real-world. Therefore, the methodology includes a set of baseline machine learning models combined with ensemble learning plans, stacking classifiers, cost-sensitive threshold optimization, and then SHAP based interpretability analysis is done to fit the predictive performance to clinical reasoning.

Conceptual Literature Gap

The development of prediction methods of heart diseases, traditional statistical models, and advanced ensemble learning methods. The diagram also presents three fundamental gaps that have been found in

the existing literature, namely the limitation of integration of explainability, the inadequacy of attention to false negative reduction and the absence of lightweight deployment-concentrated structures. The framework suggested resides between ensemble learning and cost-sensitive optimization and explainable AI as a solution to address these gaps and make clinical use cases attractive (Figures 2–5).

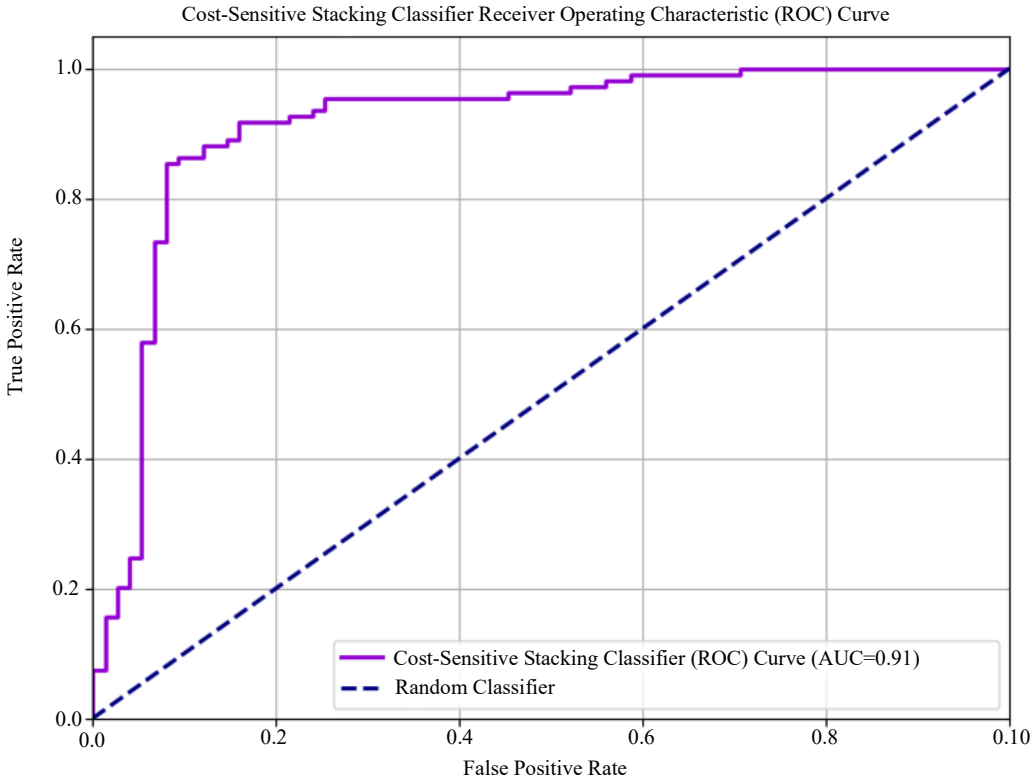

Figure 2. Architecture and workflow of the cost-sensitive stacking classifier model.

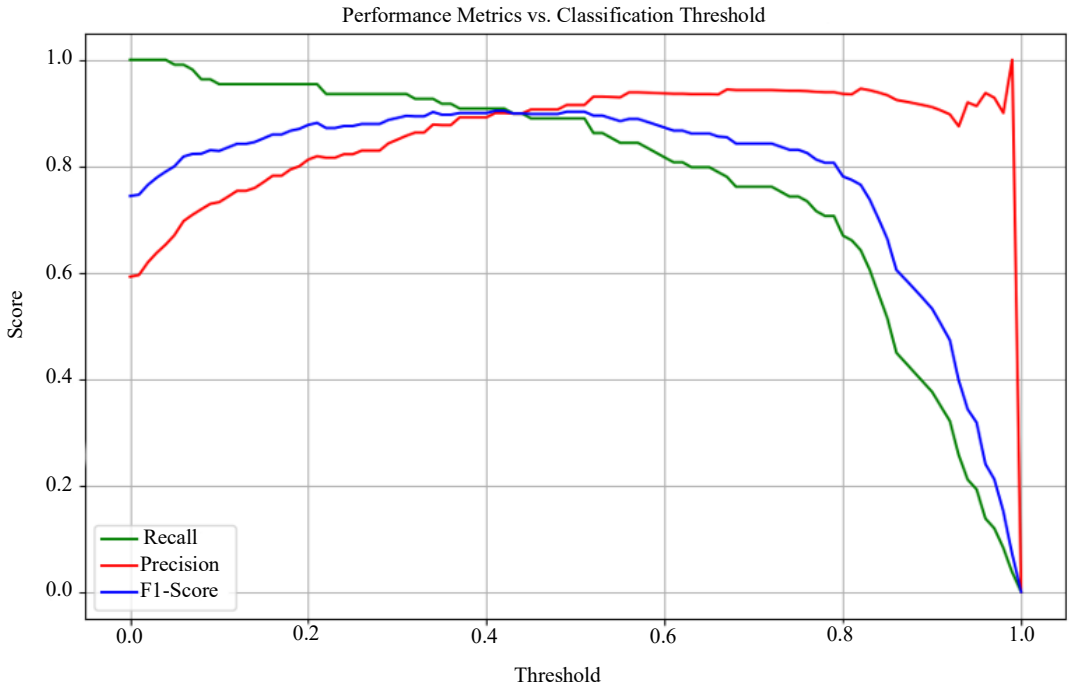

Figure 3. Variation of performance metrics with respect to classification threshold values.

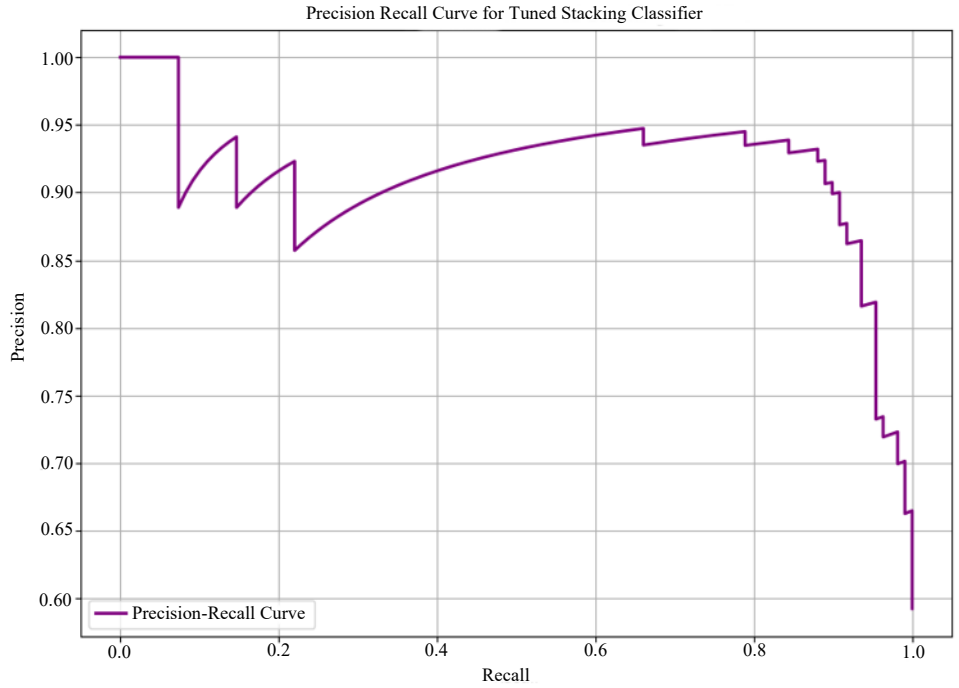

Figure 4. Precision–Recall curve showing the trade-off between sensitivity.

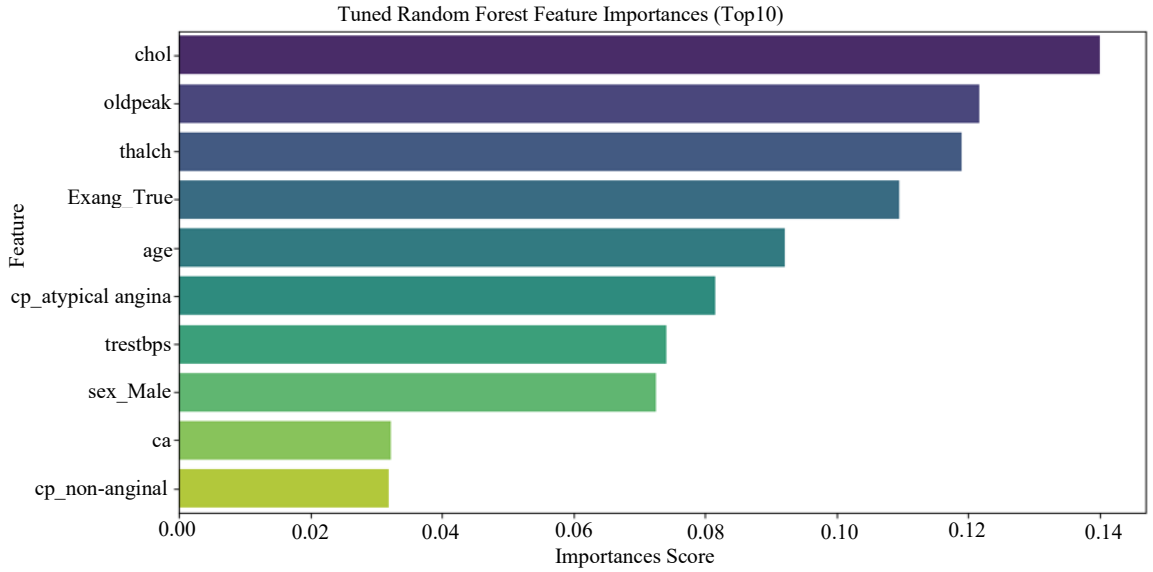

Figure 5. Top ten feature importance scores obtained from the tuned random forest model.

METHODOLOGY

In this section, the author expounds on the nature of the methodologies employed in designing and developing an explainable and light weight machine learning system, to predict early heart disease. The framework incorporates data preprocessing, model development, ensemble learning, optimization methods, and interpretability models to guarantee reproducibility, clinical safety and feasibility of deployment.

The Overall Design Is Designed on the Concept of Hierarchical Design

The proposed framework comprises a clinical decision support pipeline that is an end to end and sequential. Unlike the classical models, which are largely founded on the issue of predictive accuracy, the proposed alternative is focused on minimizing false negatives, increased interpretability, and the

requirement to be computationally friendly. This is one of the main factors that would ensure the framework would succeed when implemented in a real-life healthcare environment.

These are a series of methodological steps which include the following:

- Feature engineering and item Data Preprocessing.
- Training tree and yield models.
- Human > Stack-based ensemble learning.
- Object Classification Threshold Optimization.
- Type Multicriteria Performance Evaluation.
- LIME and Explainability Analysis based on item SHAP.

All the phases cover important limitations identified in previous research, like model transparency, too large of a computational cost and insufficient attention to false negative reduction.

Dataset Description

Each experiment is carried out on a structured clinical dataset with both socio-medical and diagnostic attributes that are typically utilized during cardiovascular risk assessment. They are age, sex, the type of chest pain, the rest heart pressure, serum cholesterol, fasting blood sugar, the electrocardiography outcomes, the highest heart rate, the angina in exercise, the ST depression (oldpeak), the count of large vessels, and the thalassemia-related characteristics, which are routine variables in clinical diagnosis studies.

The target variable is binary, as it takes the form of the presence or absence of heart disease, so it is appropriate to the early screening systems of the disease.

The Data Undergoes Preprocessing and Cleanup Stages as Described Below

Data preprocessing is done to enhance resilience and remove biasness in training a model. Missing numbers are replaced by the median value whereas the missing categorical values are replaced by the most common category, as a best practice in medical data preprocessing [1].

The noise and dimensionality are sieved-off by rejecting the irrelevant or redundant features. Categorical variables are coded by use of one-hot encoding, and numerical features are normalized to enable equal feature scaling to their respective values. The data obtained after the refining process is further separated into training and testing data sets, which can be evaluated impartially.

The modeling of the basis model entails the modeling of an elementary valuation utilizing assumptions with regards to the interplay amid the components.

This is the background cause or the reason why it is being used as a baseline classifier in this application by using Logistic Regression as a simple and interpretation based and generally applicable form of classifier to be used in clinical risk prediction endeavors [2]. It provides a starting point by which it is possible to measure gains that the ensemble-based models have achieved in terms of recall and precision and minimizing false negative.

Tree-Based Models

The tree-based models are nonlinear interaction of clinical variables. It uses an Rf classifier that uses a combination of multiple decision trees to increase stability and generalization.

Both tuned and untuned cases are considered. Hyperparameter optimization is used with great caution as over-optimization has been shown to decrease generalization performance in medical data [3].

Boosting

XGBoost is applied as a gradient-boosting algorithm that is optimized to work with tabular and high-dimensional clinical data and is optimized on structured tabular data [4, 5]. It enhances the performance through repeated learning on misclassified instances and regularization to prevent the occurrence of

overfitting instances of misclassification of performance criteria and information depth through regularization and refraining against overfitting and overtraining sample samples and bias tendencies in information representation and implementation by conformity to pattern consistency and sample execution characteristics of the density and field of information expression by wavelength fluctuations [6].

This is done to ensure consistency of experiments using the same preprocessing and evaluation procedures.

Stacking-Based Ensemble

Stacking ensemble learning is used to enhance predictive reliability further. Stacking combines the outputs of several base learners with a meta-classifier that allows one to capitalize on the ability of complementary model strengths to be leveraged.

The stacking configurations that are analyzed are:

- Untuned StackingClassifier with Logistic Regression as meta-learner.
- Tuned StackingClassifier with Logistic Regression as meta-learner.
- StackingClassifier with Random Forest as meta-learner.

These setups facilitate systematic evaluation of recall stability, precision, and false negative behavior [7].

Hyperparameter Tuning

Structured search strategies which include grid and randomized search are used to perform hyperparameter optimization because these methods help prevent overfitting while increasing system performance and system reliability, according to Pedregosa et al. (2011) [8]. Model performance must be balanced with generalization capability because only specific baseline models show high performance.

The medical field does not find default probability thresholds suitable because its diagnostic process requires higher recall rates than it needs precise results, according to Rajkomar (2019) [9]. Threshold optimization is, therefore, applied to selected models, particularly stacking classifiers, to explicitly control the recall–precision trade-off.

Lower thresholds increase sensitivity and reduce false negatives, while higher thresholds increase precision and reduce false positives.

Cost-sensitive learning is incorporated through the implementation of greater penalties for false negative errors than for false positive mistakes which matches actual clinical risk assessment needs, according to Obermeyer and Emanuel (2016) [10]. This approach aligns model optimization with patient safety objectives and is suitable for early screening deployments.

Model performance assessment uses confusion matrix metrics which include accuracy, precision, recall, F1 score, and ROC AUC according to Ahmad et al. (2018) [11].

Explainability Analysis

The framework includes explainability as its fundamental component. The SHAP method calculates global feature attribution by determining how each input variable affects the predictive accuracy of the model [12]. Researchers consistently identify ST depression and maximum heart rate and chest pain type and cholesterol levels as the primary risk factors.

LIME is used to create local explanations that show how each patient prediction works, which helps doctors to comprehend the decision-making process for each specific case [13, 14].

Reproducibility and Practical Deployment

The research experiments use the same preprocessing methods and evaluation standards and validation procedures to achieve reproducibility of results according to reference 19. The framework

enables hospitals to use standard computer systems because it does not require advanced computing equipment for their operations.

This section will elaborate on the methodologies that will be used in designing and developing an explainable and light-weight machine learning system for the early prediction of heart disease. There are various methodologies that have been developed, and these methodologies have been designed in some particular way in order to be more reproducible and effective in their applications. This includes diversity in tasks, such as data processing, model development, ensemble learning techniques on models, optimization techniques applied to models, and efficiency of models based on various real-life parameters.

Overall Framework Design

The proposed framework adopts a systematic end-to-end pipeline approach for clinical decision support. Unlike other frameworks that mainly emphasize predictive capability, the proposed framework emphasizes the importance of avoiding false negatives, interpretability, and efficiency. Avoiding false negatives is the most critical factor in the healthcare domain, as a high-risk patient missed by the system can be dangerous [15].

They consist of a sequence of methodological steps, which implies the following:

- Preprocessing and feature engineering of item Data.
- Boosting models and item Tree.
- item Ensemble Learning with.
- purpose This is the optimization technique that utilizes standard deviation standard to cluster data points.
- Multivariate model assessment on different criteria.
- Explanatory Analysis SHAP and LIME.
- It concerns the gap that exists in the previous studies on the issues of transparency, complexity, and enhancement of the false negative rate that is the case of every stage.

Dataset Description

These tests have been performed on a disciplined data set which is a combination of both socio-medical and medical traits, which have a genesis on the traditional cardiovascular analysis tests. The medical characteristics are age, sex, chest pains, the blood pressure at rest, the cholesterol amount, the quantity of glucose, the electrocardiogram, the maximum heart rate achieved, the induced angina, the ST alterations, old peaks, the typical colors of the vessels under the fluoroscopic appearance, and the typical thalassemia characteristic.

Target Variable in case target variable, one is working with heart disease and the same is modeled as a binary classification problem. This is especially useful in the case of early screening when efforts are trying to distinguish between the patients that need to be addressed in a high-level facility other than just being diagnosed.

Data Preprocessing and Cleaning

Preprocessing of the data will be handled, so that any model being used acts very well without any bias. The data is initially screened for missing values, inconsistencies, and an irrelevant feature. Numerical attributes with missing values are replaced by the median since median are less affected by outliers. Nominal features missing values replaced the most popular category [16].

This removes the noise and unnecessary complexity in the data by removing irrelevant columns that have non-informative data. The categorical variables are encoded using the one-hot method, where a categorical variable is converted into a numeric variable that can be handled by the machine learning algorithm. The numerical attributes are then scaled using standard techniques of normalization such that the attributes with larger ranges of values do not dominate the training process.

In the binary target variable, it is a yes or no response that a patient is diseased or not. The outcome of the processed information to be evaluated is separated into two sections, a training set and a test set.

This Is Because the Baseline Model Has Not Been Developed Yet

The Logistic Regression will be employed as a baseline classifier due to its simpleness, interpretability, and common application on medical prediction tasks. This model offers an apt point of reference that more sophisticated and ensemble-based models can be compared to.

The baseline model serves to put into perspective the improvements in Recall, Precision, and reduction of False Negatives by the ensemble methods as meaningful and by clinical relevance common sense.

Tree-Based Models

The tree-based models can be used to model the non-linear relations that exist between clinical features. Random Forest classifier is a model that is trained on a set of decision trees, and it can account for interactions between features with a lesser tendency to overfit because of the combination of features.

Untuned and tuned Random Forest were tested. The tuning tries various configurations of parameters to determine which among them can be used to affect recall, precision and overall performance. The findings established that tuning enhanced the performance, however, excessive or improperly validated tuning indeed worsens generalization, so optimization of the trade-off is recommended [17].

Boosting

XGBoost algorithm implementation occurs based on a gradient boost approach due to its effectiveness and accomplishment in operating structured tabular data of the medical domain. The method improves performance learning through the errors that are made by the other models and regularized methods are used in the technique.

Like other models, to train, or to evaluate XGBoost it employs the same preprocessing steps and the same evaluation metrics. The model performance statistics of XGBoost indicate that the model has good recall and precision. This helps in making predictions on risks in the medical profession.

Stacking-Based Ensemble

To improve the accuracy of prediction, one engages an ensemble learning methodology that is stacked.

Multi-classifier Stacking is used whereby a meta-classifier is used to take the final prediction using base classifiers.

Through it, the suggested architecture takes advantage of a wide range of learning algorithms. strengthen homogenization act in broad skills of artificial intelligence.

Stacking Arrangements: These Are the stacking Arrangements of the Following

Stacking-Based Ensemble

To improve the accuracy of prediction, one engages an ensemble learning methodology that is stacked.

Multi-classifier Stacking is used whereby a meta-classifier is used to take the final prediction using base classifiers.

Through it, the suggested architecture takes advantage of a wide range of learning algorithms. strengthen homogenization act in broad skills of artificial intelligence.

Hyperparameter Tuning

Search hyperparameter Hyperparameter search is applied selectively through structured search paths. Hyperparameter search is not done rigorously to all models, since a hyperparameter search is done to models that are good in terms of a baseline measure. It can avoid omissions or overfitting.

Based on the findings of the parameter tuning process, it has been seen that the optimal performance may not be achieved with extreme values of the parameter as well, and therefore, appropriate validation should be made when employing the parameter tuning in healthcare.

Where the classification threshold is set to achieve optimal accuracy, the results are depicted in Table 3 of the Appendix: Set 1.

The default classes in disease related diagnosis tasks may fail to provide the significance among target classes. The above-discussed problem may be relevant to implementation of a threshold optimization strategy of some models that favor stacking classifiers.

The value of probability threshold will enable the balancing point between recall and precision to be controlled since a lower probability threshold will result in the increase in the value of recall since there will be more patients under the category of high risk, and the value of precision increase since the number of false positives will be reduced.

Cost-S

The chosen models have embraced approaches to cost-sensitive learning and prioritize highlighting their significance as opposed to patient safety. Otherwise, the penalty or loss would be higher with false negative than with a false positive. This happens to be the case due to the radical implication of an individual with heart disease.

The approach is also meant to harmonize the optimization techniques of models to healthcare objectives. It is also a safe way of being deployed in early screening models.

subsection Performance Evaluation Metrics In this study the metrics are set to appear as performance evaluation indicators, embodying personal preferences, in the event of process inefficiency, and over time to establish the variability of results in this study, will be presented as performance evaluation indicators, reflecting personal choices, in case of process inefficiency, and in time to determine the variability of outcomes.

tests the performance of a classification model on a specific problem with numerous parameters relying on the confusion matrix:

Accuracy: This refers to the rightness of all the predictions given. The computations of the recall are accorded equal importance since the significance of the recall is very high in eradicating the issue of false negatives. Precision: The predictability of all the predictions made, and F1-score: the score of precision and recall. Both classes are separable as the ROC AUC score quantifies the separability of both classes [18].

Of interest is the number of false positive and false negative.

Explainability Analysis

Explainability is not perceived as an afterthought but is a part of the framework. SHAP analysis allows obtaining global explanations, which compute the significance of features to predictions made by the model. It is interesting to note that the given analysis shows that such significant risk factors as the ST depression (oldpeak), the highest heart rate achieved, the type of chest pain, and cholesterol have been noticed.

LIME is utilized in the provision of local explanations in relation to patient-level prediction. These local explanations are aimed at making the doctors know why a particular patient has been categorized into low-risk or high-risk patients [11]. subsection Reproducibility and Practical This indicates that it can be reproduced and is practical.

Experiments are performed in a similar way in terms of preprocessing, evaluation measures, and validation schemes to be able to make sure that they are reproducible. The suggested solution has been designed in a manner that it can operate efficiently without depending on the high-performance computing powers [19].

CONCLUSION

The research has enriched the machine learning literature by providing a lightweight interpretable paradigm of early predicting the risk of heart disease based on the structurally represented clinical information. The construction of this paradigm has placed heavy concentration on interpretations, and applicability to the task whereby the greater the focus has been on the recall, the false negative reduction and interpretability rather than the optimal development in terms of accuracy. Learning strategies have been completely compared using a shared experimentation framework to the evaluation of the cardiovascular disease risk in the study.

The experiments clearly showed that ensemble technology can even perform better than single classifiers. Although the Logistic Regression classifier was chosen as a preferred benchmarking method because of its simplicity and inability to acquire more complex clinical relationships, it resulted in more false negatives. RF and XGBoost had a tremendous enhancement on the recall and robustness as tree-based and boosting algorithms. And among all the methods that were experimented to find the characteristics of balanced performance measures, the ensemble stacking methods were the most convenient to use considering the sensitivity and the reliability balance. With the simple StackingClassifier, balanced techniques yielded the highest F1-score and AUC. Besides, threshold selected and cost-sensitive ensemble techniques displayed the highest values of recall.

It was found that an interrelationship between strategies in the optimization of the classification threshold and cost sensitive learning provided value in terms of modifying the attitude of the classifiers to the real-life priorities. In reality, which can be observed the cost of false negative example is more than that of false positive example. This technique is based on the following principle.

The model of explainability was implemented in the model by utilizing the SHAP and LIME. It was also noted that the world had explanations of ST depression, heart rate achieved, type of chest pains, angina during exercise and color of cholesterol and major vessels fluoroscopically. Local explanations also furthered this information and applied the same information to predict patients. This is to ensure that even in this proposed approach, there is also validity.

In conclusion, it can be seen that the results confirm that the machine learning ensembles of a lightweight nature, together with a method of explanation and an optimization heuristic inspired by clinician expertise, can be part of an efficient and reliable model with the capability to predict early risk of cardiovascular disease. There appears to be promise with regards to application within a DDS.

FUTURE WORK

Although the proposed framework has excellent performance and interpretability capabilities, there are some research directions in the future that still need consideration. First, since this study is done on one structured heart disease dataset, future studies should be aimed at testing the framework on different data.

Third, the methods involving threshold optimization and cost-sensitive learning rely on fixed settings. For future studies, there could be adaptive or dynamic methods to choose the threshold levels depending on clinical aspects or availability of healthcare resources. Furthermore, more advanced methods in cost-sensitive training could allow greater alignment in model training and clinical priorities in diagnosis.

Fourth, although the explanations offered by SHAP and LIME in the study were interpretable, the need for validation by clinician-in-the-loops in future studies to establish the reception and application of explanations is warranted. Exploratory work on the effect of explanations on decision-making would be beneficial.

Lastly, future studies can be focused on incorporating the proposed framework into a clinical decision support system so that there can be smooth communication between the clinical staff and the predictive models. Additionally, mechanisms for data monitoring and model update can also be investigated to address data drift and clinical patterns that continue to change.

REFERENCES

1. World Health Organization. World Health Organization cardiovascular diseases (CVDs) fact sheet. World Health Organ. 2020;42(1):207–16.
2. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
3. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29 (5):1189–1232.
4. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 785–94.
5. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59.
6. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
7. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 1135–44.
8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
10. Obermeyer Z, Emanuel EJ. Predicting the future—Big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9.
11. Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; 2018 Aug 15–18; Washington, DC, USA. p. 559–60.
12. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):93.
13. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
14. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol. 2021;6(6):633–41.
15. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2015;24(1):44–65.
16. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
17. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2018;22(5):1589–604.
18. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73.
19. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

