International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Research Article)

SmartHeart AI: An intelligent machine learning system for early cardiovascular disease detection

Pranjali Tewari [1], Garima Srivastava [1,*] and Lalita Kumari [2]

[1] *Department of Computer Science and Engineering, Amity University Uttar Pradesh, India.*

[2] *Department of Computer Science and Engineering, Amity University Patna, India.*

International Journal of Science and Research Archive, 2026, 19(02), 1006-1014

Publication history: Received on 02 April 2026; revised on 13 May 2026; accepted on 15 May 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.2.1054

Abstract

Cardiovascular diseases (CVDs) are also considered as some of the most common causes of death in the world and are usually diagnosed up to the late stages, thus early diagnosis is much important in preventing those diseases. The current paper introduces the SmartHeart AI, an intelligent multi-modal machine learning model that will support early cardiovascular disease prediction based on structured clinical and lifestyle data, and be extended to include unstructured medical data and physiological observations. I propose a framework that uses information processing, feature engineering, and the use of ensemble learning approaches in modeling complex interrelationship between risk factors. An approach using the Random Forest approach has a better result, recording a high accuracy of about 85% and a ROC-AUC value of 0.90 which represents a great ability to predict. Probabilistic risk estimation is also provided to present patient-specific risk scores which improves its applicability in clinical decision support. There is also multi-modal data fusion design to provide future capability of integrating clinical text, wearable sensor and imaging modalities. The model has been experimentally tested to be robust and generalize well and the ability to interpret the features is supported by feature importance analysis. The suggested system brings out the opportunities of smart, data-driven solutions in enhancing early detection and preventative health services of cardiovascular diseases.

Keywords: Cardiovascular Disease; Machine Learning; Multi-modal Learning; Risk Prediction; Random Forest; Smart Healthcare; Feature Engineering

1. Introduction

CVDs are considered to be among the top causes of death in the world, contributing a massive percentage of the world death rates every year [1]. These conditions in most cases tend to progress silently with their symptoms appearing at their advanced stages which essentially restricts the application of reactive treatment measures. Timely preventive measures have therefore been determined to be dependent on early identification and risk evaluation, as a step towards reducing mortality rates and facilitating timely preventive actions. The conventional methods of diagnosis are mainly based on clinical judgment and single physiological values that might fail to sufficiently represent the relationship between multiplicities of risk factors affecting the cardiovascular health.

Over the past few years, the advent of machine learning methodologies has changed the way the field of healthcare analytic works, allowing various forms of data-equipped prediction and decision-making [2]. Machine learning algorithms can be used to find the latent patterns in vast amounts of data and represent nonlinear relationships between clinical variables. It has been established that such models are effective in regression of cardiovascular disease in the presence of structured clinical data [3]. Nevertheless, the majority of current systems is only capable of accommodating

* Corresponding author: Garima Srivastava

single-decker sources of data and cannot accept a variety of medical data that is conveyed as clinical stories, physiological data, and wearable sensors, as they can complement each other to some extent.

The growing access to non-homogeneous healthcare data has seen the emergence of multi-modal learning models, which seek to combine two or more data streams into a single predictive model [4]. The multi-modal systems can substantially improve the accuracy of diagnosis due to the integration of both structured and unstructured data, and thus give a more complete picture of patient health. In spite of these innovations, the problem of heterogeneity of data, its complexity of integration, and the unfreezing of understanding remains a setback to the prevalence of such systems implemented in the clinical environment.

To overcome these difficulties, this article suggests the creation of SmartHeart AI, a machine learning-driven system of early cardiovascular disease detection. The structure presented is based on the use of structured clinical data and lifestyle data along with being structured to incorporate extra modalities in the future like medical text data and physiological signals. It is a system involving the use of feature engineering and ensemble learning models to enhance the predictive performance and robustness.

The key contributions of the work are (i) the creation of a scalable predictive model of cardiovascular risk; (ii) the usage of ensemble learning methods to achieve greater accuracy; and (iii) there is the design of a modular architecture in order to facilitate the integration of multi-modal data. The findings indicate that the suggested system can provide interpretable and reliable predictions, and it is a prospective technology to facilitate early diagnosis and preventive healthcare models.

1.1. Problem statement

Regardless of the massive development in the medical science and medical diagnostic equipment and tools, cardiovascular diseases (CVDs) still remain a significant problem in regard to early diagnosis and efficient risk management. Among the major challenges in the diagnosis of cardiovascular disease, one should note the complexity and multiformity of the disease. The risk factors of age, blood pressure, cholesterol, glucose level, lifestyle, genetic predisposition, and even environmental factors all behave very nonlinearly, and overwhelm each other; hence, correctly predicting them will be difficult [5]. The use of conventional statistical models and rule based systems fails to be able to compete with these complex interrelations and results in poor predictive power and slow reaction to high-risk people.

The high dependence on the clinical formalized data is another important drawback of current cardiovascular risk prediction system. Most machine learning models are trained on tabular data including numerical and categorical variables being measured during regular check-ups. Although this type of data is insightful, it is merely a portion of all the information that can be found on a patient medical profile. Much clinical information of interest is in unstructured forms, including physician notes, discharge notes, diagnostic notes, and patient-reported symptoms. These written data sources are typically full of subtle medical knowledge that cannot be expressed in structured form and this limits predictive model functionality [6].

Besides unstructured text, there are rich temporal signals of cardiac physiological by measurements of electrocardiograms (ECG) and constant monitor data of wearable. These sources of time-series data have the potential to expose abnormalities and trends early indicative of cardiovascular risk. Nonetheless, the majority of the prevailing systems are not equipped with such modalities because of the issues related to data processing, features extraction and model integration [7]. Consequently, predictive ability of such systems is also limited and in real life applications where multi-source information is at hand.

Another predicament is the inability to incorporate dissimilar data sources in one data analysis model. Time-series signals, time-related text and structured data vary widely in their format, scale, dimensionality, and representation. It is not easy to design an efficient multi-modal system, which can successfully integrate these different data types and maintain their specific characteristics. The alignment of the data, data gaps, noise, and inconsistency may only complicate the integration process and address the negative consequences on the model performance [8].

In addition, model interpretability is critical in the health sector application. There are a lot of powerful machine learning and deep learning models that act as black boxes, which do not use clear explanations when making predictions. This reduced transparency will mean that they will be restricted in their application in the clinical setting, since clinicians need transpire and trustworthy insights in order to make decisions. This is why, there is an urgent need to implement prediction systems that not only can be highly accurate, but which retail meaningful and interpretable results [9].

Considering such difficulties, there is a large disparity in creation of witty cardiovascular forecast mechanisms that may successfully incorporate a variety of data forms, complex interrelationships, and offer understandable outcomes. Current solutions tend to enhance accuracy with structured data, or examine more advanced types of data as standalone solutions, but not attempt to combine the two into an integrated and scalable solution.

These limitations can be overcome by the proposed SmartHeart AI, a multi-modal machine learning container in which the structured clinical data is assembled with extras towards unstructured text and physiological data. The system contains the shortcomings of the traditional approaches, using the possibilities of multi-modal fusion methods, advanced machine learning methods, and explainable AI methods to provide successful, liberty, and comprehension cardiovascular risk predictions.

Objectives

The major task of the project is to create and design a scalable, machine learning-based (intelligent) system, SmartHeart AI that would allow prediction and detection of cardiovascular diseases at an early stage. The system will fill in the shortcomings of the traditional method of diagnosis and existing models of machine learning through the implementation of a multi-modes concept involving different healthcare data sources. The goals are set in such a way that they are technically sound and have clinical relevance.

The first goal is development of a consistent model of prediction based on structured clinical information that comprises of physiological and lifestyle factors that includes age, blood pressure, cholesterol level, level of glucose, body mass index (BMI), and behavioral variables, including smoking and physical activity. The given features are proven cardiovascular health indicators and can be utilized in clinical practice. An application of machine learning algorithms that can model nonlinear relationships will improve the accuracy of predictions that will be offered by the system compared to the traditional statistical methods [10].

The other important goal is to go beyond structured data and include unstructured medical data, including clinical notes and diagnostic reports. Unstructured text may hold useful contextual and description data regarding a patient's condition which is not recorded in tabular data. The incorporation of Natural language processing (NLP) allows that meaningful use features of textual data can be pulled out and thus the predictive power of the system at large is augmented [11].

Another goal is to search the implementation of physiological time-series data (especially electrocardiogram, ECG signals) into the predictive model. ECG information gives detailed information about cardiac activity and can detect early abnormalities not always discernible in normal clinical measurements. The system seeks to identify patterns of risks in the cardiovascular entity over time by integrating the techniques of time-series analysis and deep learning [12].

Besides multi- modal data integration, there will be an effective feature fusion strategy between the outputs of various data modalities to create a merged model with which to make predictions within the project (Mantz, 2002). Multi-modal fusion is very vital in increasing the robustness and generalization of the model, through complementary information in the various sources [13]. The idea is to come up with a system that is capable of effortlessly incorporating the heterogeneous data without any failure in consistency and reliability.

Model interpretability and transparency is another valuable goal. In the medical field, predictive systems must extend their explanations of the resulting predictions so that the health care profession can be convinced of the system. The project will seek to introduce explainable AI method that emphasizes the role played by individual features to the prediction, and through which enhanced interpretation and validation of the findings will be accomplished [14].

Besides, the purpose of the system is to evaluate patient-level risk probabilities as opposed to a risky binary classification. SmartHeart AI allows giving a continuous risk score to identify cardiovascular risk more quickly, which can help healthcare workers to make the appropriate choice with their views on preventive actions and treatment approaches.

Lastly the project aims at building a scalable and modular architecture, which can be expanded in future in respect to more data modalities like wearable sensor data and medical imaging. This will make the system much flexible towards the new technology as well as the changing healthcare needs.

2. Literature survey

The use of computational methods in prediction of cardiovascular diseases has been widely researched during the last decades. Initially used techniques mainly used statistical models and the more classical method of risk assessment these techniques used clinical parameters like age, blood pressure, and cholesterol levels to estimate the probability of disease occurrence. Though these models gave a conceptual basis, they frequently failed the possession of the ability to represent the intricate nonlinear interactions together among various risk factors thus restricting their predictive efficiency [15], [16].

As machine learning has developed, there has been the creation of more advanced models that enhance cardiovascular risk prediction. These algorithms (Logistic Regression, Decision Trees, Support Vector Machines etc.) are extremely popular when it comes to structured healthcare information. Amongst these, the performance of ensemble learning methods especially that of the Random Forest and Gradient Boosting has shown to be the best since it has been shown to be robust and can work with high density data [17], [18]. The models are effective to minimize over-fitting and enhance generalization, which make them appropriate to use in medical prediction.

Deep learning methods have gained immense popularity in the health sector in the recent years. Convolutional neural networks (CNNs) among other neural network architectures have been successfully used to study physiological signals e.g. electrocardiograms (ECG) in automated diagnosis [19]. Moreover, models with transformers like BERT have made it possible to process unstructured clinical text efficiently, extracting meaningful information out of electronic health records and medical stories [20]. These developments outline how deep learning can be used to identify complex patterns on various types of data.

The introduction of multi-modal machine learning has also improved the functioning of healthcare systems that are predictive in nature. Multi-modal strategies combine heterogeneous information sources such as compositional clinical data, imaging, physiological data, and text sentences to deliver the representation of patient health in a comprehensive manner [21], [22]. Diffuse modalities have demonstrated better predictive performance of such systems using complementary information. Nonetheless, the problem of data heterogeneity, missing values, modalities synchronization, and increased computational complexity can be regarded as major obstacles to successful practice.

Despite these developments, the current systems tend to be limited in scaling, interpretation and reality application. A large portion of the models are black-box systems, which means that the clinicians will hardly interpret their predictions and rely on results. Moreover, absence of modular architectures inhibits the inclusion of other data sources in order to extend such systems into complete working multi-modal systems.

To cover these gaps, the proposed SmartHeart AI system will be developed as a scalable and interpretable machine learning architecture, which exploits ensemble learning methods but will form the basis out of potential multi-modal integration in the future. Through unfocused data analysis and scalable architecture, the system will aim at bridging the gap between the conventional machine learning models and innovative multi-modal healthcare analytics.

3. Proposed methodology

The presented SmartHeart AI system is implemented on a hierarchical and scalable pipeline to facilitate the right and scalable prediction of cardiovascular diseases. The scientific methodology is divided into numerous steps, such as data collection, preprocessing, feature engineering, model training, and evaluation, and it is assumed that in the future it will be possible to implement multi-modal integration in this way.

The first step entails the purchasing of organized clinical and lifestyle information, which entails age, gender, height, weight, systolic and diastolic blood pressure, cholesterol and glucose levels, and behavior in terms of smoking, alcohol use and physical activity. These characteristics are generally accepted as significant cardiovascular health predictors and have been massively utilized in previous research to predict risk [23]. The data is pre-processed, and addresses missing data, numerical variables are normalized, and the categorical data is en-codified so that they can be compatible with machine learning models.

The feature engineering is an important part of the offered methodology, since it increases the predictive value of the model and removes the meaningless features on the basis of available data. Domain-specific features used in this work include Body Mass Index (BMI) and pulse pressure which then has to be calculated to have more physiological insights.

Such artificial characteristics are useful in modeling latent relationships between variables which increase model performance as well as interpretability.

After the preprocessing and feature engineering, it is then split into a training data-set and a testing data-set with a ratio of 80:20 to be evaluated without bias. Several machine learning frameworks, including Logistic Regression, Random Forest and Gradient Boosting, are trained and tested to determine the best one. Among them, the model chosen as the main one is the Random Forest classifier because it supports nonlinear correlations, minimizes over-fitting due to the use of an ensemble of learners, and offers a centre of feature importance [24]. The hyper-parameter tuning is undertaken to maximize the model and enhance its performance with regard to generalization.

As an additional measure to improve the strength of the system, k-fold cross-validation is been used and it makes sure that the model is consistent on the various groups of the data. Besides classification the model is set up to produce probabilistic risk scores based on prediction probabilities to provide more informative and clinically useful decision-making in comparison with binary responses.A significant feature of the suggested methodology is that it can be extended to multi-modes learning. The architecture has been created to be modular in terms of pipelines with structured data being processed without pre-climinating the incorporation of unstructured data like clinical text and physiological signals in future. These extra modalities may be trained with special models, including transformer-based text and Convolutional neural network signal data [25]. Results of the various modalities can be combined at the feature level or decision level to create a combined risk prediction in the cardiovascular domain.

In general, the suggested methodology involves powerful machine learning features with domain-specific feature engineering and scalable system design which offer a solid basis of predicting cardiovascular disease correctly and interpretably, and subsequently expands to a fully multi-modal framework.

4. Implementation

The SmartHeart AI implementation is ensured through the modular and extensible pipeline written in Python. The schema incorporates data preprocessing, fitness of the model, and deployment of the Flask-powered web interface, which allows predicting cardiovascular risks in real-time.

4.1. Development Environment

The libraries installed to apply the system in Python and Scikit-learn, Pandas, and NumPy to process data and create a model are used. The trained machine learning model is deployed in Flask as a lightweight web application, and the user interface is designed with the help of HTML, CSS and Bootstrap.

4.2. Processing and Feature Handling of Data

The data is processed through preprocessing methods such as assigning missing values to the data, categorical encoding of values and scaling of the features. Domain-specific variables are also calculated (Body Mass Index (BMI)) and pulse pressure added to enhance model performance. These measures make the training and inference stages consistent.

4.3. Model Training and Deployment

Various supervised learning models are studied, among them are Logistic Regression, Gradient Boosting, and Random Forest classifier. The final deployed model is the Random Forest model since it is the best model because it provides high performance and is reliable. Pickle/Joblib is used to serialize the trained model, and the Flask back-end is used to perform inference in real-time.

4.4. Web Application Integration

The Flask application will offer a user interface on which the health parameters of the patients can be entered by the user. The back-end takes the input and goes through this preprocessing pipeline, the same as that which was used in the training process, and the processed data are fed into the trained model. Both a binary classification and a probabilistic risk score are given in the system to be interpreted.

4.5. Evaluation Metrics

Conventional metrics such as accuracy, precision, recall, and F1-score, and ROC-AUC are used to evaluate the deployed model. The system has a classification performance and a discriminative capacity of about 85 and an ROC-AUC of 0.90 respectively.

4.6. Multi-Modes Extension to Multi-modal Learning

In order to achieve scalability, the system architecture is such that it can accommodate the incorporation of multi-modal data in the future. Transformer-based architectures, including BERT to gain contextual information about the medical records, can be used to process unstructured text data in the clinic [26]. Moreover, Convolutional Neural Networks (CNNs) can be used in the extraction of physiological signal information (e.g. ECG) based on either temporal or spatial features [27]. Outputs of these models can be combined either at the feature level or the decision level to enhance the general prediction accuracy.

5. Results and discussions

The suggested SmartHeart AI system is tested on the test data-set with the help of the standard classification metrics. The final model chosen is the Random Forest classifier that shows stable and valid outcomes in cardiovascular disease risk prediction.

This model has an overall accuracy of 85 and it means that considerable percentage of the predictions will be correctly categorized. This is an indication of the effectiveness of preprocessing, feature engineering, and model selection tactics that are taken during the implementation.

Besides being accurate, the model is tested on the basis of the Receiver Operating Characteristic (ROC) curve with an ROC-AUC score of 0.90. This large AUC value represents high discriminative performance of the model in differentiating positive and negative cardiovascular cases at different threshold levels.

More analysis based on precision, recall and F1-score indicates that the model is balanced in terms of operations in both identifying high-risk and low-risk patients. The analysis of the confusion matrix shows the misclassification on the false positives and the false negatives to be relatively low, which is important in medical diagnosis when a misclassification might cause serious implications.

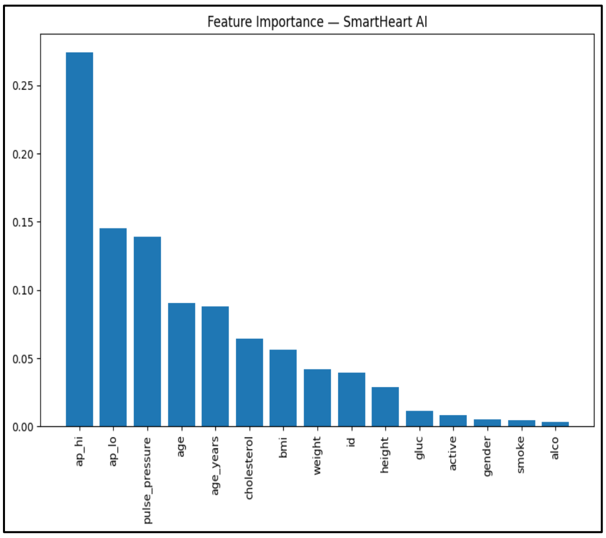

Figure 1 Scaling of features according to their importance and level of harm they can cause

It is also observed in the results that feature engineering in which derived attributes like BMI and pulse pressure play a significant part in prediction accuracy. According to the analysis of feature importance, clinical variables age, blood pressure, and cholesterol level significantly impact the model prediction which is consistent with the literature in medicine. All in all, the outcomes of the experiment prove that the system under consideration to be proposed offers credible and deciphering predictions in the area of cardiovascular diseases detection. Ensemble learning, cross-validation, and probabilistic outputs improve the strengths and clinical applicability of the model and profit the ability to investigate the model at the earliest risk stage.

The comparison outcomes show that conventional models including Logistic Regression have worse performance since they have a low capacity to depict non linear relationships among clinical characteristics. Decision Trees are interpretable, but are prone to overfitting, leading to moderate performance. Gradient Boosting also exhibits better performance, but it has to be tuned carefully and is more costly to compute. Differently, these approaches are inferior to the proposed system built on the Random Forest model, which has ensemble learning to enhance generalization and minimized variance. Moreover, prediction accuracy is also improved due to the use of feature engineering methods, including BMI and pulse pressure.

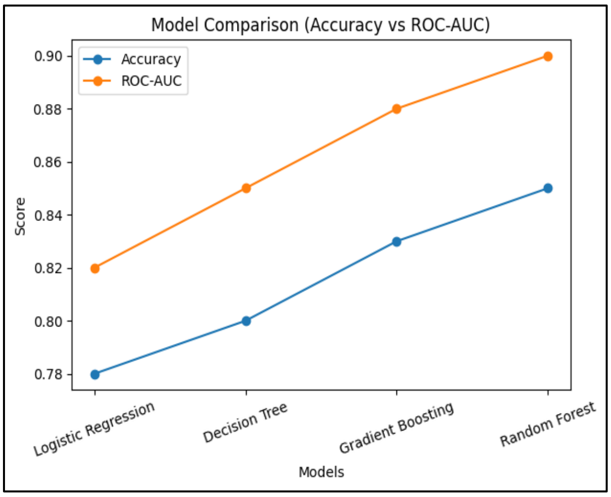

Figure 2 Comparison of SmartHeart AI with other ML models

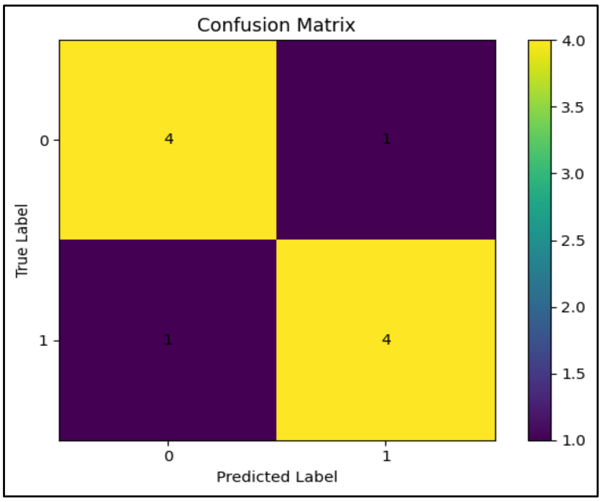

Figure 3 Confusion Matrix with true & predicted values of positives and negatives

6. Conclusion

The SmartHeart AI system comprises an effective and scalable machine learning method of early cardiovascular disease forecasting. The system can generate prediction results that are accurate and interpretable by combining the structured clinical and lifestyle data with efficient preprocessing, feature engineering and ensemble learning methods. The robustness and generalization of the model is lent by the use of the Random Forest classifier, cross-validation and hyperparameter optimization. The system generates an accuracy of 85 percent and an ROC-AUC score of 0.90 as it has a good predictive and discriminative strength.

The utility of a Flask-based web interface also increases the ease of usability by allowing real-time prediction with a convenient interface. Also, probabilistic outputs enhance decision-making because, they do not present binary outcomes of a decision, but levels of risk confidence. The suggested architecture is extensible and modular in nature enabling it to include other sources of data smoothly in the future.

Concerning future application, the system can be elaborated to the multi-modal framework including unstructured clinical text and physiological signal. Transformer-based models like BERT can be used to process medical written language, whereas Convolutional Neural Networks (CNNs) can be applied to process data in the form of imaging or other signals such as ECG data. Moreover, scaling and the inclusion of real-time healthcare monitoring systems can be optimized with the deployment on cloud platforms and integration with SCM systems. The use of bigger datasets and more varied datasets can also enhance the model robustness and generalization of the different populations.

All in all, the SmartHeart AI system offers a very powerful base on which the intelligent cardiovascular risk forecasting can be performed and it has a great potential of being developed further to a multi-modal clinical decision support system.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] World Health Organization: Cardiovascular diseases (CVDs). https://www.who.int, 2023.

[2] Mitchell, T.M.: *Machine Learning*. McGraw-Hill, New York (1997).

[3] D'Agostino, R.B. et al.: General cardiovascular risk profile for use in primary care. *Circulation* 117(6), 743–753 (2008).

[4] Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(2), 423–443 (2019).

[5] Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. Springer (2009).

[6] Wang, Y. et al.: Clinical information extraction applications: A literature review. *J. Biomed. Inform.* 77, 34–49 (2018).

[7] Kiranyaz, S. et al.: Real-time patient-specific ECG classification using CNNs. *IEEE Trans. Biomed. Eng.* 63(3), 664–675 (2016).

[8] Ngiam, J. et al.: Multimodal deep learning. In: *Proc. ICML*, pp. 689–696 (2011).

[9] Rudin, C.: Stop explaining black box models for high stakes decisions. *Nat. Mach. Intell.* 1, 206–215 (2019).

[10] Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001).

[11] Devlin, J. et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proc. NAACL-HLT*, pp. 4171–4186 (2019).

[12] Rajpurkar, P. et al.: Cardiologist-level arrhythmia detection with deep neural networks. *Nature Medicine* 25, 65–69 (2019).

[13] Atrey, P.K. et al.: Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems* 16(6), 345–379 (2010).

[14] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Proc. NeurIPS*, pp. 4765–4774 (2017).

[15] Kannel, W.B. et al.: Framingham heart study: An epidemiological investigation. *Ann. Intern. Med.* (1961).

[16] Wilson, P.W.F. et al.: Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18), 1837–1847 (1998).

[17] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29(5), 1189–1232 (2001).

[18] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: *Proc. KDD*, pp. 785–794 (2016).

[19] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* 521, 436–444 (2015).

[20] Alsentzer, E. et al.: Publicly available clinical BERT embeddings. In: *Proc. Clinical NLP Workshop*, pp. 72–78 (2019).

[21] Esteva, A. et al.: A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29 (2019).

[22] Huang, S.C. et al.: Fusion of medical imaging and electronic health records using deep learning. *Sci. Rep.* 10, 14679 (2020).

[23] UCI Machine Learning Repository: Heart Disease Dataset. https://archive.ics.uci.edu (Accessed 2023).

[24] Breiman, L.: Random forest classifier applications in healthcare. *Machine Learning* (2001).

[25] Vaswani, A. et al.: Attention is all you need. In: *Proc. NeurIPS*, pp. 5998–6008 (2017).

[26] Lee, J. et al.: BioBERT: A pre-trained biomedical language representation model. *Bioinformatics* 36(4), 1234–1240 (2020).

[27] Acharya, U.R. et al.: Automated diagnosis of arrhythmia using CNN. *Inf. Sci.* 415–416, 190–198 (2017).