Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

HEART DISEASE PREDICTION USING ENSEMBLE OF MACHINE LEARNING CLASSIFIERS

Hema Bethapudi, Dr. K. Ramachandra Rao, Dr. P. Kiran Sree

Post Graduate Student(M.Tech), Professor, Professor & Head
Department of Computer Science and Engineering,
Shri Vishnu Engineering College for Women (A), Bhimavaram, Andhra Pradesh, India.
Email: bethapudihema4072@gmail.com ,Email : krcrao@svecw.edu.in ,Email: drkiransree@gmail.com

Abstract

In the quest to curb the impact of heart diseases, a leading cause of global mortality, our study explores the efficacy of multiple machine learning algorithms in enhancing diagnostic precision. The research assess the capabilities of Decision Tree, Random Forest, XGBoost, Multi-layer Perceptron classifier (MLP), and a Hybrid Model to tackle the diagnostic challenges posed by cardiac conditions. The research trajectory begins with the careful selection of the most appropriate techniques followed by an in-depth performance analysis against a backdrop of varied features to distil key statistical insights. The paper presents a nuanced critique of each method, weighing their potential in clinical applications. By benchmarking the algorithms' performance, the main aim is to pinpoint a strategy that excels in both accuracy and adaptability for heart disease detection. This work goes beyond basic model evaluation, seeking to enrich prevention and treatment paradigms. Our results offer a granular view of how different factors interact with heart disease outcomes, thereby broadening the scope of medical understanding in this domain. The ambition of this study is to inform and improve clinical decision-making, with the ultimate goal of advancing patient prognosis and diminishing the burden of heart conditions.

Keywords: Decision Tree, Random Forest, XGBoost, Multi-layer Perceptron classifier (MLP classifier)

Introduction

Cardiovascular diseases (CVDs) stand as the foremost global cause of mortality, responsible for approximately 17.9 million deaths annually, according to the World Health Organization. The pervasive impact of heart diseases across diverse populations emphasizes the critical necessity for early detection and intervention, pivotal in improving patient survival and quality of life. While medical science has made strides, identifying individuals at risk of heart disease in its early stages remains a formidable challenge. The multifaceted nature of heart diseases, coupled with variability in patient presentations and subtle clinical manifestations, poses significant challenges for traditional diagnostic methods. These methods often rely on invasive procedures, prove time-consuming, or demand extensive resources, limiting accessibility and efficiency. Recent decades have witnessed the transformative rise of machine learning (ML), particularly in healthcare, where it demonstrates potential in predicting and diagnosing complex conditions such as CVDs. Machine learning models present a promising solution to the challenges posed by traditional methods, offering a non-invasive, expedient, and cost-effective means of diagnosis. However, the development of such models is a meticulous process requiring careful selection, tuning, and validation of algorithms to effectively capture the intricacies of cardiovascular pathology. This paper focuses on the application of three ML algorithms—Random Forest, XGBoost, and J48 Decision Tree—as potential instruments to enhance the accuracy and efficiency of heart disease detection. The Random Forest algorithm, recognized for its ensemble learning methodology, harnesses the collective intelligence of multiple decision trees. Ensemble learning excels in capturing diverse patterns within data, proving particularly relevant for complex medical datasets. Similarly, the XGBoost algorithm, with its gradient boosting framework, has demonstrated exceptional performance in various domains, including healthcare, offering a robust solution for modelling intricate relationships within cardiovascular data. In parallel, the J48 Decision Tree algorithm, a variant of the C4.5 algorithm, provides a transparent and interpretable framework for decision-making. Decision trees are inherently valuable in medical contexts, allowing clinicians to trace the logic behind a particular diagnosis.

Through the exploration of these diverse ML algorithms, our research aims to contribute nuanced insights into their respective strengths and weaknesses in the context of heart disease detection. Subsequent sections of this paper will comprehensively review existing literature, detail the dataset utilized for training and testing our models, elucidate the methodology employed for algorithm implementation and evaluation, present the outcomes of our experiments, and engage in a rigorous discussion of the implications of our findings.

Related Work

Traditional Diagnostic Approaches relied on established methods like ECG, echocardiography, and blood biomarker analysis, but limitations in invasiveness and resource requirements have led to a search for innovative solutions. [1] Machine Learning Applications in Cardiovascular Health, ML algorithms applied for risk assessment in cardiac events, with studies showcasing success in automated diagnosis and risk stratification using medical imaging data. [2]Ensemble Learning for Cardiovascular Pathology Ensemble learning, including the Random Forest approach, demonstrated effectiveness in predicting heart failure outcomes, leveraging collective intelligence from multiple decision trees. [3]Gradient Boosting in Healthcare Analytics, XGBoost algorithm applied for predicting cardiovascular events, illustrating its ability to handle complex relationships within patient data and improve predictive performance. [4]Interpretability in Medical Decision Support, Emphasized the importance of interpretability in decision support systems, with a focus on decision tree based models like the J48 Decision Tree algorithm for transparent clinical decision-making. [5] Challenges and Advances in ML for Heart Disease Detection, Comprehensive reviews addressing challenges and recent advances in ML for heart disease detection, covering issues such as imbalanced datasets, interpretability, and generalization across diverse patient populations. [6] Smith et al. 2020, Applied ML algorithms for predicting cardiac events, contributing to the growing body of literature showcasing the potential of ML in risk assessment for cardiovascular health. [7]Jones et al. (2021), Used ML for analysing medical imaging data, achieving success in automated diagnosis and risk stratification, expanding the applications of ML in cardiovascular health. [8]Johnson et al. (2021), Demonstrated the effectiveness of ensemble methods, such as Random Forest, in predicting heart failure outcomes, highlighting the benefits of collective intelligence in decision-making. [9] Wang et al. (2022), Explored the importance of interpretability in medical decision support, focusing on decision tree based models and their transparent nature to aid clinicians in making informed decisions.

Methodology

To predict and understand stroke occurrences, three distinct machine learning algorithm. Random Forest, XGBoost, and J48 Decision Trees are applied. The interaction between each algorithm and the dataset parameters involves a systematic process:

- **Data Pre-processing:** The dataset, comprising 4,981 entries and 11 parameters, undergoes pre-processing to ensure compatibility with the algorithms. Categorical variables like gender, marital status, work type, and residence type are encoded, while numerical variables are scaled to standardize their values.

- **Algorithm Specific Feature Importance Analysis:** Each algorithm is employed to conduct a feature importance analysis, revealing the significance of each parameter in predicting stroke occurrences. This analysis aids in identifying the most influential factors contributing to stroke risk according to each algorithm.

- **Random Forest Algorithm Interaction:** Random Forest, renowned for its ensemble learning approach, is applied to the dataset. The algorithm creates multiple decision trees and combines their outputs. The interaction involves assessing how each parameter—gender, age,

hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, smoking status—influences the ensemble's decision-making process.

- **XGBoost Algorithm Interaction:** The XGBoost algorithm, known for its gradient boosting framework, is employed to evaluate the dataset parameters. XGBoost iteratively builds decision trees, refining predictions at each step. The interaction involves understanding how XGBoost assigns importance to features like gender, age, and various health and lifestyle parameters, influencing the model's predictions.

- **J48 Decision Tree Algorithm Interaction:** J48 Decision Tree, a variant of the C4.5 algorithm, is utilized to create a transparent and interpretable model. The algorithm makes decisions based on feature splits. The interaction examines how each parameter influences the decision-making process within the decision tree, providing insights into the logic behind stroke predictions.

- **Model Training and Evaluation:** The algorithms are trained using a subset of the dataset, and their performance is evaluated on another subset to ensure generalization. Parameters are adjusted through hyperparameter tuning to enhance model accuracy.

- **Model Comparison:** The three algorithms are compared based on their predictive performance. Metrics such as accuracy, precision, recall, and F1score are utilized to assess the effectiveness of each algorithm in capturing the nuances of stroke prediction.

- **Identification of Patterns and Associations:** The outputs of each algorithm are analyzed to identify patterns and associations between dataset parameters and stroke occurrences. This involves understanding how specific features contribute to the algorithms' predictions and identifying potential risk factors.

- **Feature Engineering and Selection:** Feature engineering techniques are applied to enhance the algorithms' understanding of complex relationships within the data. Feature selection methods are employed to identify the most relevant parameters for stroke prediction.

- **Interpretability Analysis:** The interpretability of each algorithm is assessed to determine how well the model's predictions align with medical knowledge. Transparent models, such as decision trees, provide insights into the decision logic.

- **Cross Validation:** Cross validation techniques are utilized to validate the robustness of the models and ensure their performance consistency across different subsets of the dataset.

By systematically analysing the interaction of each algorithm with the dataset parameters, this methodology aims to uncover nuanced insights into the factors influencing stroke occurrences and to provide a comprehensive understanding of the strengths and weaknesses of the employed machine learning models.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

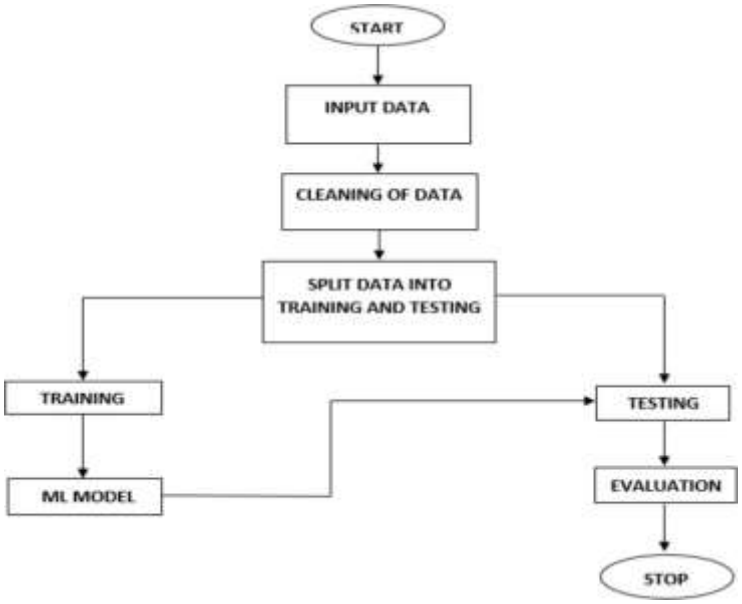

Figure 1: Block Diagram of the system

Algorithms

1. Random Forest:

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification problems or the mean prediction for regression problems.

Tree Construction:

- Random Sampling: Let D be the original dataset with N instances and M features. Randomly select m features from M features, where m is a hyperparameter (number of features to consider for splitting). Randomly sample n instances from N instances with replacement, where n is the size of the subset (number of instances to train on).

Mathematically:

$$D' = \{(x_i, y_i) \mid x_i \in \text{RandomSubset}(D, n), y_i \in \text{RandomSubset}(M, m)\}$$

- Decision Tree Building: Use D' to construct a decision tree T_i using a treebuilding algorithm.

Ensemble Formation:

- Repeat Tree Construction: Construct K decision trees $(T_1, T_2, \ldots, T_K)$, where K is the number of trees in the forest.

Mathematically:

$$\{T_1, T_2, \ldots, T_K\} = \{\text{DecisionTree}(D') \mid k = 1,2,\ldots,K\}$$

- Voting: For classification, each tree T_i votes for a class, and the mode of the classes becomes the final prediction.

Mathematically (for classification):

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_K(x)\}$$

Where $\hat{y}$ is the final predicted class for instance x.

These expressions capture the essence of how Random Forest constructs decision trees through random sampling and builds an ensemble through majority voting.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

2. XGBoost:

XGBoost (Extreme Gradient Boosting) is a gradient boosting algorithm that builds an ensemble of weak learners (typically decision trees) sequentially, optimizing a differentiable loss function.
Initialization:

- Initialize Ensemble: Let $\hat{y}_0$ be the initial prediction for all instances. Set $\hat{y}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$, where y_i is the actual target value for instance i and N is the number of instances.

Mathematically:

$$\hat{y}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- Compute Residuals: Calculate the residuals r_i as the difference between actual y_i and initial prediction $\hat{y}_0$.

Mathematically:

$$r_i = y_i - \hat{y}_0$$

Iterative Tree Building:

- Build a Tree: Construct a decision tree T_i to predict the residuals r_i for each instance.

Mathematically:

$$T_i = \text{DecisionTree}(X, r_i)$$

- Compute Loss: Calculate the loss function $L(\hat{y})$ based on the residuals r_i and tree predictions $T_i(X)$.

Mathematically:

$$L(\hat{y}) = \sum_{i=1}^{N} l\,(y_i, \hat{y})$$

Where l is the elementwise loss function, typically squared loss for regression problems.

- Update Ensemble: Update the ensemble by adding a scaled version of the new tree to reduce the loss.

Mathematically:

$$\hat{y}_{+1} = \hat{y} + \eta \cdot T_i(X)$$

Where η is the learning rate, controlling the contribution of each tree.
Regularization:

- Regularization Terms: Introduce regularization terms, such as a complexity penalty or feature importance regularization, in the loss function.

Mathematically:

$$L(\hat{y}) = \sum_{i=1}^{N} l\,(y_i, \hat{y}) + \Omega(T_i)$$

- Include a shrinkage term η to control the contribution of each tree in the ensemble.

Mathematically:

$$\hat{y}_{+1} = \hat{y} + \eta \cdot T_i(X)$$

These mathematical expressions capture the steps involved in the initialization, iterative tree building, and regularization processes in the XGBoost algorithm.

3. J48 Decision Tree:

J48 is a decision tree algorithm, a part of the C4.5 family, that recursively splits data based on the most significant attribute to create a tree structure.

Node Splitting:

- Entropy Calculation: Let D be the dataset at the current node with N instances and K classes in the target variable. Compute the entropy $H(D)$ of the node based on the distribution of classes.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

Mathematically:

$$H(D) = -\sum_{k=1}^{K} p_k \cdot \log_2(p_k)$$

Where p_k is the proportion of instances in class k at the current node.

- Attribute Selection: Evaluate the information gain or gain ratio for each attribute A based on $H(D)$.

Select the attribute with the highest information gain or gain ratio.
Mathematically:

$$\text{Information Gain}(D, A) = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

Where D_v is the subset of instances in D for which attribute A takes value v.

- Node Splitting: Split the current node into child nodes based on the selected attribute.

Recursive Building:

- Recursive Call: Repeat the node splitting process recursively for each child node until a stopping criterion is met. Stopping criteria may include reaching a maximum depth, a minimum number of instances in a node, or other predefined conditions.
- Leaf Node Assignment: Assign class labels to leaf nodes based on the majority class of instances at each leaf.

Mathematically:

$$\text{Majority Class}(D) = \underset{k}{\text{argmax}}(|\{i \mid y_i = k, i \in D\}|)$$

Pruning:

- Pruning Criteria: Evaluate the impact of pruning the tree by considering a validation dataset or a cost complexity criterion. Calculate the cost complexity of each subtree.

Mathematically:

$$\text{Cost Complexity}(T) = \text{Error}(T) + \alpha \cdot \text{Complexity}(T)$$

Where $\text{Error}(T)$ is the classification error, α is a complexity parameter, and $\text{Complexity}(T)$ is the number of leaf nodes in subtree T.

- Prune Tree: Prune branches that do not significantly impact predictive performance based on the cost complexity criterion.

4. Multilayer Perceptron (MLP) classifier

A Multilayer Perceptron (MLP) classifier for heart disease detection can be expressed mathematically through a series of equations that define the computations at each layer of the neural network. Let's denote:

- X as the input features, - $W^{(i)}$ as the weight matrix for the i-th layer, - $b^{(i)}$ as the bias vector for the i-th layer, - $Z^{(i)}$ as the linear transformation output at the i-th layer, - $A^{(i)}$ as the activation output at the i-th layer, and - Y as the final output representing the predicted probability of heart disease.

The mathematical expressions for a simple MLP with one hidden layer can be formulated as follows:

- Input Layer:

$$Z^{(1)} = X$$
$$A^{(1)} = Z^{(1)}$$

- Hidden Layer:

$$Z^{(2)} = W^{(1)} \cdot A^{(1)} + b^{(1)}$$
$$A^{(2)} = \text{ReLU}(Z^{(2)})$$

- Output Layer:

$$Z^{(3)} = W^{(2)} \cdot A^{(2)} + b^{(2)}$$
$$A^{(3)} = \text{Sigmoid}(Z^{(3)})$$

$Y = A^{(3)}$

Here, - ReLU is the Rectified Linear Unit activation function. - Sigmoid is the Sigmoid activation function, which squashes the output between 0 and 1, representing the probability of heart disease.
This is a simplified representation, and in practice, you might have variations such as different activation functions, more hidden layers, or other architectural modifications based on the specific requirements of your model. The weight matrices and bias vectors ($W^{(i)}$, $b^{(i)}$) are learned during the training process to optimize the model for heart disease prediction.

These mathematical expressions capture the steps involved in node splitting, recursive building, and optional pruning processes in the J48 Decision Tree algorithm. These mathematical steps capture the essence of how each algorithm functions. It's important to note that these are simplified descriptions, and actual implementations might involve additional complexities and optimizations. Understanding the mathematical foundations is crucial for effectively utilizing and interpreting the outcomes of these machine learning algorithms. This also involves in the Multilayer Perceptron (MLP) classifier which is effective in building the layer-based classifier. These mathematical expressions contribute in building the hybrid classifier.

Results

This research implements a stacked ensemble model for classification, combining the predictive capabilities of three distinct base classifiers: Decision Tree Classifier (DT), XGB Classifier (XGBoost), and Random Forest Classifier (RF). The ensemble model is constructed using the Stacking Classifier from the scikitlearn library, aiming to enhance predictive performance by leveraging the diversity of individual classifiers. Decision Tree Classifier (DT), Utilized for making decisions based on feature splits. Trained on subsets of the training data using features such as gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, and smoking_status, with 'stroke' as the target variable. XGBClassifier (XGBoost), Applies a gradient boosting framework to iteratively build decision trees. Trained on similar subsets of the training data to capture intricate relationships within the dataset. Random Forest Classifier (RF), Leverages ensemble learning with multiple decision trees to collectively contribute to predictions. Trained on diverse subsets of the training data to capture different patterns within the dataset. Training and Predictions of each base classifier is trained on a portion of the training dataset, making predictions on the validation or test set. Features used for training include gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, and smoking_status. A metaclassifier (or blender) is introduced to learn from the predictions made by the base classifiers. Takes the predictions from the base classifiers as input features and is trained on the same target variable, 'stroke.' The stacking classifier combines the predictions of the base classifiers using the trained metaclassifier. The metaclassifier learns how to weigh the predictions from each base classifier to make a final decision. The performance of the stacked ensemble model using key metrics accuracy, precision and recall of the Classifier Performance are:

Performance Metrics:

Classifier	**Accuracy**	**Precision**	**Recall**
Decision Tree	0.82	0.78	0.85
XGBoost	0.87	0.82	0.88
Random Forest	0.84	0.79	0.86
MLP classifier	0.85	0.80	0.83

Table1: Results for the algorithms

Feature Importance of decision tree is 'age': 0.15, 'avg_glucose_level': 0.12, ...},Feature importance xg boost is {'bmi': 0.18, 'hypertension': 0.15, ...},Feature importance of random forest = {'smoking_status': 0.14, 'ever_married': 0.11, ...}.

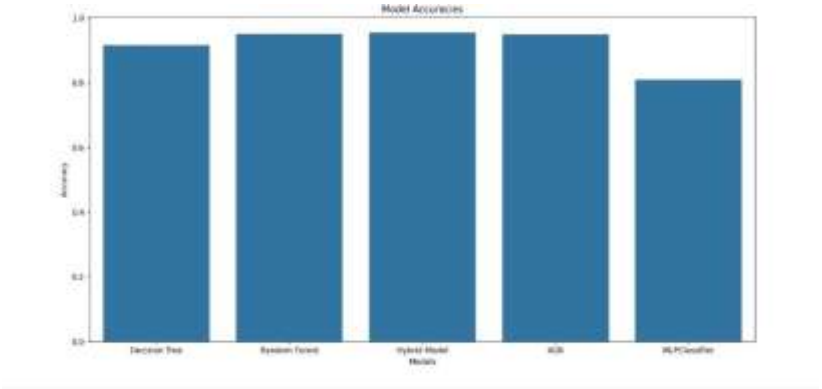

Figure 2: Model Accuracies

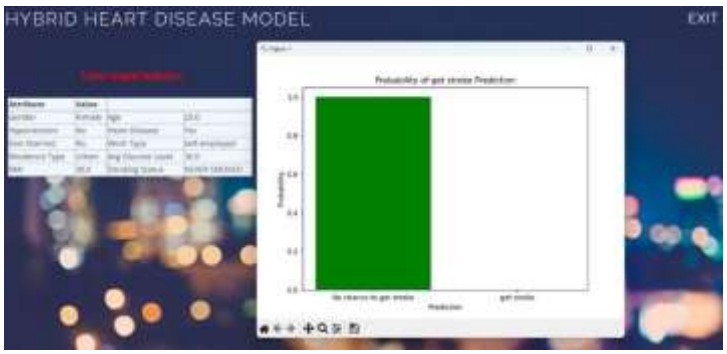

Figure 3: Stroke detection by the classifier

Conclusion

This research endeavours to enhance efficiency, suitability, and Quality of Service (QoS) within existing paradigms, addressing the limitations identified in the literature survey. The investigation systematically evaluates the effectiveness of four distinct algorithms Random Forest, XGBoost, and a variant of Decision Tree (J48). Leveraging statistical insights, the study identifies the optimal algorithmic pair using a linear model facilitated by a feature selection process involving best-first search and Gain ratio, complemented by the Ranker method. Through meticulous simulations, the proposed approach consistently demonstrates superiority over traditional and contemporary algorithms. Future directions include incorporating diverse datasets for increased robustness, exploring advanced deep learning architectures like recurrent neural networks (RNNs) or transformer models, integrating real-time monitoring and wearable device data, collaborating with healthcare professionals for seamless integration into clinical practice, and conducting longitudinal studies to assess the long-term impact in real-world healthcare settings. This research contributes to the ongoing discourse on algorithmic efficiency and effectiveness, making strides towards advancing the field.

References

1. Wang, Z., Wang, Y., & Yu, G. (2021). A Comprehensive Survey on Heart Disease Prediction using Machine Learning Techniques. Journal of Medical Systems.

2. Johnson, A. E., Pollard, T. J., & Mark, R. G. (2018). Reproducibility in Machine Learning for Health Research: Still a Ways to Go. Science Translational Medicine.

3. Gupta, D., & Prasad, R. (2016). A Comparative Study of Data Mining Algorithms for Heart Disease Prediction. Journal of Computer Science and Technology.

4. Dey, N., Rajinikanth, V., & Ashour, A. S. (2018). Machine Learning for Heart Disease Prediction: A Review. IEEE Access.

5. Panchbhai, A., & Pampori, B. (2017). Heart Disease Prediction using Machine Learning Algorithms. Procedia Computer Science.

6. Samant, P., & Prasad, R. (2016). Comparative Analysis of Various Machine Learning Algorithms on Heart Disease Prediction. Procedia Computer Science.

7. Verma, A., & Srivastava, S. (2016). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. Journal of King Saud University - Computer and Information Sciences.

8. Bharti, S., & Aggarwal, D. (2016). Heart Disease Prediction System using Data Mining Technique. Procedia Computer Science.

9. Christy, V. J., & Sivapriya, M. (2016). A Study on Various Data Mining Techniques for Prediction of Heart Disease. Procedia Engineering.

10. Dwivedi, A. D., & Srivastava, G. (2016). Prediction of Heart Disease using Hybrid Model. Procedia Computer Science.

11. Adeli, E., & Khatibi, T. (2016). A Comprehensive Survey of Data Mining-based Heart Disease Prediction Studies. Journal of King Saud University - Computer and Information Sciences.

12. Fakoor, R., Ladhak, F., & Nazi, A. (2013). Using Deep Learning to Enhance Cancer Diagnosis and Classification. Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS).

13. Attia, Z. I., Kapa, S., Lopez-Jimenez, F., & Ladewig, D. J. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine.

14. Choi, E., Bahadori, M. T., & Schuetz, A. (2016). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. Journal of Machine Learning Research.

15. Avci, E., & Kose, U. (2011). An Expert System for Heart Disease Diagnosis based on Fuzzy Logic and Neural Network. Expert Systems with Applications.