Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

An Efficient Deep Learning-based Model for Heart Diseases Prediction

Balaji Venkateswaran[1], Dr Deepak Dagar[2]
[1]Research scholar (Computer Science), School of Engineering and Technology
Shri Venkateshwara University, Gajraula, UP, INDIA
Email: balaji.venkateswaran@gmail.com
[2]Research Guide (Computer Science), School of Engineering and Technology
Shri Venkateshwara University, Gajraula, UP, INDIA
Email: deepakdagar.faculty@maims.ac.in

Cite this paper as: Balaji Venkateswaran, Dr Deepak Dagar (2024) " An Efficient Deep Learning-based Model for Heart Diseases Prediction". *Frontiers in Health Informatics,* (8), 5693-5702

Abstract: Heart disease remains one of the leading causes of mortality worldwide, necessitating effective predictive models for timely diagnosis and intervention. This study explores the application of deep learning techniques in predicting heart disease, leveraging the power of data mining and machine learning. Initially, electrocardiogram (ECG) numeric datasets are preprocessed to extract relevant features crucial for classification. Convolutional Neural Network (CNN) is employed as a classification technique due to its superior performance compared to traditional methods. Evaluation metrics including accuracy, precision, and F-measure are computed to assess the efficacy of the CNN model against the baseline K-Nearest Neighbors (KNN) classifier. Results indicate that CNN outperforms KNN, establishing its efficacy in heart disease diagnosis on the given dataset. Furthermore, a hybrid approach integrating logistic regression and neural networks is proposed for enhanced predictive accuracy. Logistic regression identifies significant risk factors contributing to heart disease based on statistical p-values, while irrelevant factors are pruned. The resultant significant factors serve as inputs to the neural network, which is trained to predict the likelihood of heart disease. This integrated approach demonstrates promising results in predicting heart disease, highlighting the potential of combining statistical analysis with deep learning techniques for improved diagnostic accuracy.

Keywords: CNN, Machine Learning, kNN, ANN, SVM

1. INTRODUCTION

The heart, a vital organ responsible for sustaining life, is susceptible to various ailments that can compromise its function. Among these are myocardial infarction, myocardial ischemia, congenital heart disease, coronary heart disease, cardiac arrest, and peripheral heart disease, each presenting unique challenges in diagnosis and treatment. Predicting the onset of these conditions is of paramount importance in ensuring timely intervention and mitigating adverse outcomes. In pursuit of effective predictive methodologies, the intersection of machine learning and predictive analytics offers promising avenues for advancement.

Machine learning, a subset of artificial intelligence, empowers machines to glean insights from data, enabling them to make informed predictions and decisions. At its core, machine learning endeavors to

discern patterns within datasets, thereby facilitating accurate prognostications. Its applicability spans diverse domains, including recommender systems, medical diagnosis, and bioinformatics. Supervised, unsupervised, and reinforcement learning represent the primary modalities through which machine learning algorithms operate, each tailored to specific use cases and objectives. Predictive analytics harnesses a gamut of statistical techniques, encompassing predictive modeling, machine learning, and data mining, to forecast outcomes based on historical or current data. Its utility extends across industries, notably in customer relationship management and healthcare, where informed decision-making is critical. Within the realm of predictive modeling, an array of methodologies such as Naive Bayes, logistic regression, neural networks, support vector machines, and classification and regression trees have emerged as indispensable tools. Among these, artificial neural networks (ANNs) stand out for their resemblance to the intricate neural architecture of the human brain.

Artificial neural networks emulate the interconnectedness of neurons within the brain, comprising layers of nodes that process and transmit information. The network architecture typically consists of input, hidden, and output layers, with each layer performing distinct computations. Input nodes convey data to hidden layer nodes, where weighted calculations occur, before transmitting processed information to output nodes for final interpretation. This iterative process of information propagation enables neural networks to discern intricate patterns and correlations within datasets, rendering them invaluable in predictive analytics.

2. RELATED WORK

In current healthcare research, the amalgamation of machine learning, deep learning, and data mining methodologies has become increasingly prevalent for disease prediction. Each study offers unique insights and varying levels of prediction accuracy based on their respective approaches. One study suggested a hybrid method utilizing Support Vector Machine (SVM) and Genetic Algorithm (GA), effectively combining both techniques to achieve notable results. By leveraging data mining tools like LIBSVM and WEKA on five distinct datasets from the IUC repository, their hybrid model yielded accuracies of 84.07% for heart disease, 78.26% for diabetes, 76.20% for breast cancer, and 86.12% for Hepatitis [1]. Another study advocated for data mining approaches in detecting heart diseases, employing algorithms such as J48, Naïve Bayes, and bagging through the WEKA tool. Utilizing a dataset featuring 313 attributes, they achieved accuracies of 82.31% with Naïve Bayes, 84.35% with J48, and 85.35% with Bagging for heart disease classification [2].

Emphasizing the effectiveness of the Naïve Bayes algorithm, renowned for its independence assumption, another study analyzed a dataset comprising 500 patients and achieved an accuracy of 86.419% using the WEKA tool for classification [3]. A comprehensive analysis of existing works on heart disease prediction highlighted the prevalence of data mining techniques in this domain. The study underscored the importance of hybridizing multiple algorithms for improved prediction accuracy, contrasting it with the use of single algorithms [4]. Another study evaluated a sequential feature selection approach in conjunction with a neuro-fuzzy classifier, attaining an accuracy of 88.2% on the Cleveland dataset by splitting the dataset evenly for training and testing [5].

Examining ten methods using the heart disease dataset from the UCI repository, another study found Partial Least Square Discriminant Analysis (PLS-DA) exhibiting an accuracy of 86.13% [6]. A

proposal to enhance heart disease prediction through data techniques demonstrated superior accuracy (85%) particularly in parallel fashion compared to sequential SVM [7]. Exploration of various data mining approaches for predicting heart disease achieved accuracies of 84% with Neural Network and 89% with Hybrid Systems using WEKA and MATLAB [8]. Advocacy for heart disease prediction and analysis using J48, Naïve Bayes, and Support Vector Machine techniques emphasized their potential to enhance service quality and reduce costs [9-10]. Review of machine learning based heart diseases prediction approaches are shown in table 1.

Table 1. Review of Machine Learning based heart diseases prediction approaches

Study Reference	**Risk Factors Identified**	**Methodology Used**
V. Sree Hari Rao et al., [11]	Physical inactivity	In-built imputation algorithm and particle swarm optimization
Paolo Melillo et al., [12]	Long-term heart rate variability	Automatic classifier (for risk assessment in congestive heart failure)
Minas A. Karaolis et al., [13]	Events before and after CHD (e.g., PCI, MI, CABG)	Decision tree algorithm
Carlos Ordonez et al., [14]	Relevant association rules for heart disease prediction	Association rules with search constraints
Syed Umar Amin et al., [15]	Multilayered feed-forward network initialization	Hybrid system with genetic algorithm (for neural network weight initialization)
Sikander Singh Khurl et al., [16]	Chest pain, diabetes, smoking, gender, physical inactivity, age, lipids, cholesterol, triglyceride, blood pressure	Decision trees and Apriori algorithm

3. PROPOSED SYSTEM MODEL

The primary objective is to integrate a logistic regression model with a neural network-based approach for predicting heart disease. Utilizing a heart disease dataset comprising 303 observations of individuals, 297 observations are selected for analysis. The proposed system consists of two key components. Firstly, the system aims to identify the significant risk factors crucial for predicting heart disease from the available attributes in the dataset. This is achieved by determining the p-value for each attribute, which provides insights into their statistical significance [16]. Attributes with lower p-values are deemed more significant in predicting heart disease. Secondly, the dataset is divided into training and testing subsets. The neural network is constructed using the training dataset, leveraging the identified significant risk factors. Through the training process, the neural network learns to model the relationships between the risk factors and heart disease outcomes. Subsequently, the trained neural network is utilized to predict heart disease outcomes for the testing dataset, providing insights into the

model's predictive performance and generalization ability.

3.1 Data Collection

The data utilized in this project is sourced from the Cleveland Heart Disease database. It comprises a total of 297 records, each containing 14 medical attributes [7], which are employed for the prediction of heart disease. This table provides an overview of the various attributes present in the dataset, including demographic information, medical measurements, and diagnostic indicators, all of which are utilized for predicting the presence or absence of heart disease. A detailed description of the dataset is provided in Table 1 below:

Table 2. Description of Heart Diseases prediction dataset

S. No	Attribute	Description
1	Age	Age of the individual in years
2	Sex	Gender of the individual (0 = female, 1 = male)
3	Chest Pain Type (CP)	Type of chest pain experienced
4	Resting Blood Pressure (RBP)	Resting blood pressure measurement in mm Hg
5	Serum Cholesterol (Chol)	Serum cholesterol measurement in mg/dl
6	Fasting Blood Sugar (FBS)	Fasting blood sugar level > 120 mg/dl (1 = true, 0 = false)
7	Resting Electrocardiographic Results (Restecg)	Resting electrocardiographic measurement
8	Thalach	Maximum heart rate achieved during stress test
9	Exercise Induced Angina (Exang)	Exercise-induced angina (1 = yes, 0 = no)
10	ST Depression (Oldpeak)	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Number of Major Vessels (Ca)	Number of major vessels colored by fluoroscopy
13	Target (Num)	Presence of heart disease (0 = no, 1 = yes)
14	Thal	3= normal, 6= fixed defect, 7= reversible effect

3.2 Logistic Regression Model

The logistic regression model serves as a powerful statistical tool for quantifying the relationship between a categorical dependent variable and one or more independent variables. In the context of predicting heart disease, the independent variables encompass factors such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrographic results, maximum heart rate, exercise-induced angina, ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise, number of major vessels affected, and thalassemia defect. The dependent variable, representing the presence or absence of heart disease, is coded as "1" for individuals with heart disease

and "0" for those without [5]. Through logistic regression analysis, the model calculates the probability of an individual having heart disease based on these risk factors. By computing the conditional probability X represents the set of risk factors associated with the disease, the likelihood of disease occurrence can be estimated. A cutoff value of 0.5 is typically used, whereby probabilities greater than 0.5 indicate the presence of heart disease, while probabilities below 0.5 suggest its absence.

Moreover, the logistic regression model possesses the capability to identify significant risk factors that strongly influence heart disease, as indicated by their statistical significance, or p-value. The p-value, representing the probability associated with the logistic regression model's summary, is a crucial statistical metric utilized in hypothesis testing across various disciplines such as political science and economics [6]. In the context of heart disease prediction, the logistic regression model employs p-values to select the most influential risk factors from the dataset's attributes. Variables with p-values less than 0.05 ($p<0.05$) are considered statistically significant and are thus included in the predictive model. Conversely, variables with p-values exceeding 0.1 ($p>0.1$) are deemed statistically insignificant and are excluded from the model.

A higher p-value suggests that changes in the independent variable are not strongly associated with changes in the dependent variable. Following the logistic regression analysis, the significant risk factors identified based on their statistical significance include sex, chest pain type, resting blood pressure, fasting blood sugar, exercise-induced angina, slope of the peak exercise, number of major vessels affected, and thalassemia defect [8]. These variables play pivotal roles in predicting the likelihood of heart disease and are instrumental in developing effective predictive models for early detection and intervention.

3.3 Training and Testing Dataset

The dataset consisting of 297 records [16] is partitioned into training and testing datasets to facilitate the development and evaluation of predictive models. In this partitioning scheme, the training dataset is utilized to establish a predictive relationship, serving as the set of examples employed for learning and adjusting the weights of the classifier. On the other hand, the testing dataset is utilized to assess the performance of the fully-specified classifier. The division of the dataset into training and testing subsets typically follows a ratio where the training dataset constitutes 75% of the total records, while the testing dataset comprises the remaining 25%. This partitioning ratio ensures that an adequate amount of data is available for model training, while still providing a sufficient number of samples for rigorous evaluation of the trained model's performance. By employing this partitioning strategy, predictive models can be trained on a representative subset of the data and subsequently evaluated on unseen data to gauge their generalization ability and effectiveness in predicting heart disease outcomes. This approach helps to ensure that the developed models are robust and reliable, capable of accurately predicting heart disease in real-world scenarios beyond the data used for training.

3.4 Proposed Model

The neural network, inspired by biological neural networks, serves as a computational model mimicking the intricate architecture of the human brain [17]. Artificial neural networks (ANNs) replicate this structure, comprising interconnected units known as input, hidden, and output layers. In the context of medical diagnosis, patient risk factors or attributes serve as inputs to the neural network

The effectiveness of artificial neural networks in medicine, particularly in predicting coronary heart disease, has been well-established. In a typical setup, the input layer consists of neurons corresponding to significant attributes, with an output layer representing the presence or absence of heart disease (0 for absence, 1 for presence). Meanwhile, the hidden layer, containing a predetermined number of nodes, facilitates the complex processing of input data. A sample Artificial Neural Network architecture is depicted in Fig 1. One of the primary advantages of neural networks lies in their high accuracy, making them invaluable in various domains such as accounting, medicine [14], and fraud detection. Leveraging the learned network from the training dataset, the neural network can effectively predict the presence or absence of heart disease for the testing dataset, thereby aiding in early detection and intervention.

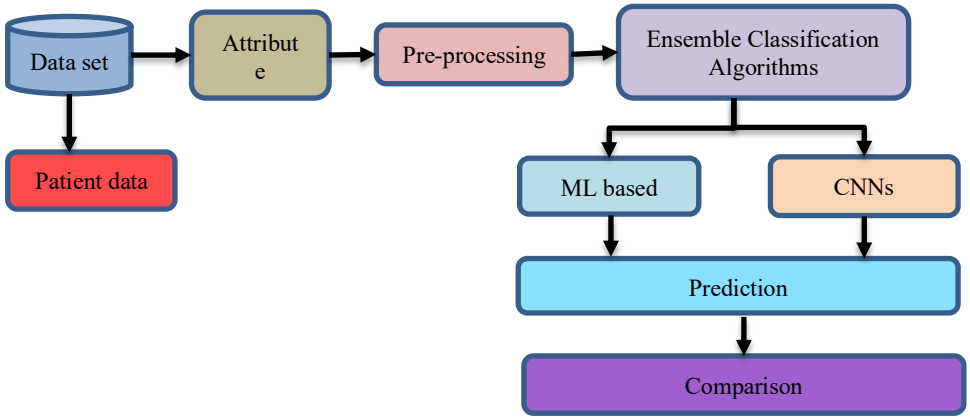

Figure 1. Proposed research methodology for heart diseases prediction

3.5 Performance Measure

The performance of the neural network is evaluated using various measures, including accuracy, specificity, and sensitivity. These metrics provide insights into the model's ability to correctly classify instances of heart disease. Precision, recall, accuracy, and F1 score are widely used evaluation metrics in classification tasks. Each metric provides a different aspect of model performance. These metrics are valuable in evaluating the performance of a classification model and can provide insights into its effectiveness in correctly predicting positive and negative instances [12-13] as depicted in Table 3.

Table 3. Performance evaluation metrics

Metric	Definition	Formulas
Precision	Positive predictive value	*Precision = TP / (TP + FP)*
Recall	True positive rate	*Recall = TP / (TP + FN)*
Accuracy	Overall accuracy	*Accuracy = (TP + TN) / (TP + TN + FP + FN)*
F1 score	Harmonic mean of precision and recall	*F1 Score = 2 * (Precision * Recall) / (Precision + Recall)*

4. RESEARCH METHODOLOGY:

The proposed research methodology for deep learning-based heart disease prediction encompasses several key steps aimed at leveraging advanced computational techniques to enhance diagnostic accuracy and patient care (Figure 2 and 3).

- **Data Collection and Preprocessing:** The process begins with data collection and preprocessing. A comprehensive dataset comprising medical records of individuals, including demographic details, clinical measurements, and relevant diagnostic indicators related to heart disease, is gathered. Subsequently, rigorous data preprocessing techniques are applied to ensure the dataset's integrity. This involves addressing issues such as missing values, outliers, and inconsistencies. Furthermore, numerical features are normalized or standardized, while categorical variables are appropriately encoded for numerical representation [11].
- **Model Selection:** This involves identifying and selecting suitable deep learning architectures tailored to the specific characteristics of the dataset and the complexity of the prediction task. Various models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), or hybrid architectures, are evaluated based on considerations such as interpretability, computational efficiency, and their ability to capture intricate patterns within the data.

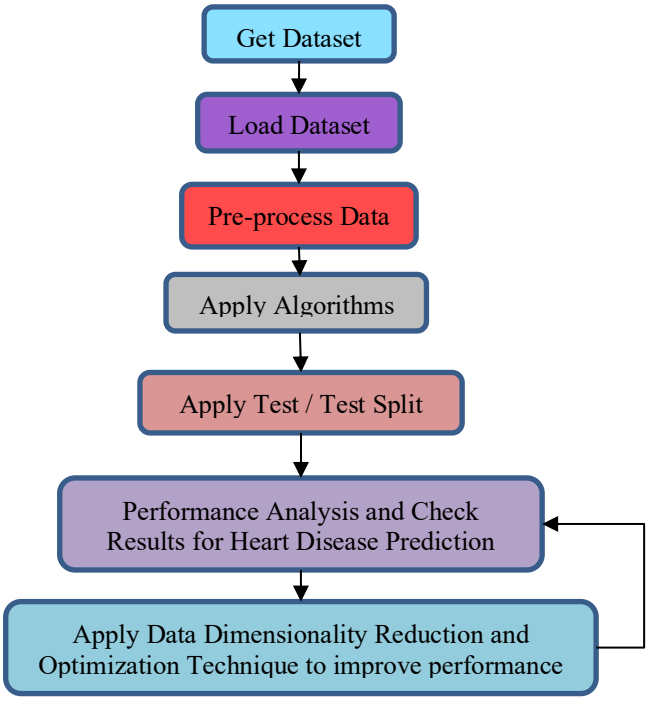

Figure 2. Flow of algorithm for heart diseases prediction

- **Model Development:** Deep learning architecture is implemented using popular frameworks such as TensorFlow or PyTorch. The model's architecture is meticulously designed, specifying the arrangement of input features, hidden layers, activation functions, and output layers to maximize predictive performance. Furthermore, hyperparameters such as learning rate, batch size, and layer configuration are fine-tuned to optimize the model's efficacy.

- **Training and Evaluation:** The model is trained iteratively using the prepared dataset, with parameters adjusted to minimize prediction errors and enhance performance. Evaluation metrics such as accuracy, sensitivity, and specificity are employed to assess the model's effectiveness in predicting heart disease. Additionally, rigorous testing on separate validation datasets is conducted to gauge the model's ability to generalize to unseen data and ensure robustness. Evaluate the trained model using the validation set. Calculate metrics such as accuracy, precision, recall, and F1-score to assess the model's performance on classifying different diseases accurately.

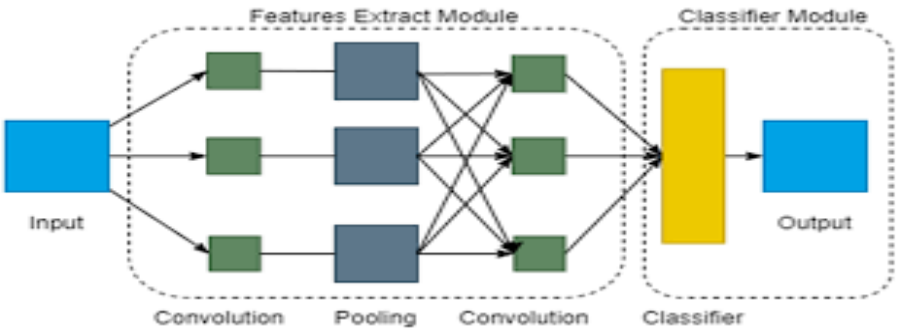

Figure 3. CNN based architecture for heart diseases prediction

5. RESULT AND ANALYSIS

The table 3, provides a detailed overview of the performance metrics for logistic regression and CNN based heart diseases prediction as shown in table

Table 4. Classwise accuracy of proposed system

Class	Precision (%)	Recall (%)	F-measure (%)
Logistic Regression	94.2	90.1	92.5
Proposed CNN based	97.2	96.4	96.2

The results from the comparative analysis between Logistic Regression and the Proposed CNN-based model provide valuable insights into their performance across key metrics: precision, recall, and F-measure. For Logistic Regression, the precision is reported at 94.2%, indicating that out of all instances classified as positive, 94.2% were correctly identified. The recall stands at 90.1%, signifying that the model successfully captured 90.1% of all actual positive instances. The harmonic means of precision and recall, known as the F-measure, is calculated at 92.5%, representing a balanced performance between precision and recall. On the other hand, the Proposed CNN-based model exhibits notably higher performance metrics. With a precision of 97.2%, the CNN-based approach showcases a higher accuracy in classifying positive instances compared to Logistic Regression. Moreover, the recall rate of 96.4% demonstrates the model's effectiveness in capturing a larger proportion of actual positive instances. Consequently, the F-measure of 96.2% reflects a robust balance between precision and recall, reaffirming the CNN-based model's superior performance over Logistic Regression.

These results highlight the efficacy of deep learning methodologies, particularly CNNs, in handling classification tasks, especially those involving intricate patterns within the data. The CNN-based

model's ability to extract hierarchical features from input data enables it to achieve higher accuracy and capture more relevant information compared to the traditional Logistic Regression approach. This suggests that in scenarios where accuracy and reliability are paramount, such as medical diagnosis or image recognition, adopting CNN-based models could offer significant advantages over conventional statistical methods like Logistic Regression.

6. CONCLUSION

In conclusion, machine learning continues to be a dynamic and rapidly evolving field, particularly in healthcare, where researchers are actively engaged in identifying disease risks. Logistic regression offers the advantage of interpretability of model parameters and ease of use, making it a valuable tool in disease prediction. Conversely, neural networks require less formal statistical training and have the capability to detect complex non-linear relationships between dependent and independent variables, enhancing their utility in healthcare applications. The integration of logistic regression and neural network techniques presents a novel approach to predicting heart disease in individuals. By combining the strengths of both methods, researchers can leverage the interpretability of logistic regression alongside the non-linear modeling capabilities of neural networks, resulting in more accurate and comprehensive predictive models. Looking ahead, future research can focus on extending these methodologies to longitudinal studies of patients, allowing for the continuous monitoring and prediction of heart disease progression over time. Additionally, efforts can be directed towards further improving the accuracy of heart disease prediction models through the incorporation of additional data sources and advanced machine learning techniques.

References:

[1] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, "A Hybrid Evolutionary Algorithm for Attribute Selection in Data Dining", Expert Systems with Applications, Vol.36, No.4, pp.8616-8630, 2009.

[2] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology, Vol.2, No.4, pp.56-66, 2014.

[3] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET-International Journal of Innovative Science, Engineering & Technology, Vol.2, pp.441-444, 2015.

[4] A.Sahaya Arthy, G. Murugeshwari “A Survey on Heart Disease Prediction using Data Mining Techniques” (April 2018).

[5] Hamid Reza Marateb and Sobhan Goudarzi, “A Non-invasive Method for Coronary Artery Diseases Diagnosis using a Clinically Interpretable Fuzzy Rule-based System,” Journal of Research in Medical Sciences, Vol. 20, Issue 3, pp.214-223, March 2015.

[6] K.R. Lakshmi, M.Veera Krishna, and S.Prem Kumar, “Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability,” International Journal of Scientific and Research Publications, Vol.3, Issue 6, pp.1-10, June 2013.

[7] R. Sharmila, S. Chellammal, “A Conceptual Method to Enhance the Prediction of Heart Diseases using the Data Techniques”, International Journal of Computer Science and Engineering, May 2018.

[8] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, “Efficient Heart Disease Prediction System”, 2016, pp.962-969.

[9] Ashwini Shetty A, Chandra Naik, “Different Data Mining Approaches for Predicting Heart Disease”, International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277-281.

[10] Wani MA, Bhat FA, Afzal S, Khan AI. Training Supervised Deep Learning Networks. In Advances in Deep Learning 2020 (pp. 31-52). Springer, Singapore.

[11] V. Sree Hari Rao, M. Naresh Kumar, “Novel Approaches for Predicting Risk Factors of Atherosclerosis,” IEEE Journal of Biomedical and Health Informatics ., vol. 17, No. 1, Jan 2013.

[12] Paolo Melillo, Nicola De Luca, Marcello Bracale and Leandro Pecchia ,"Classification Tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability", IEEE Journal of Biomedical and Health Informatics., Vol. 17, No. 3, May 2013.

[13] Minas A. Karaolis,Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis, “Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees,” IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 3, May 2010

[14] Carlos Ordonez, “Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, April 2006.

[15] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", Proceedings of IEEE Conference on Information & Communication Technologies, 2013.

[16] Sikander Singh Khurl, Gurpreet Singh, “Ranking Early Signs of Coronary Heart Disease Among Indian Patients", IEEE International Conference on Computing for Sustainable Global Development, 2015

[17] Beigh, M. A, Quadri Javeed Ahmad Peer, Kher , S. K. and Ganai, N. A, “Disease and pest management in apple: Farmers' perception and adoption in J&K state”, Journal of Applied and Natural Science 7 (1): 293 – 297 (2015)