HEART DISEASE IDENTIFICATION USING ENSEMBLE LEARNING AND DEEP LEARNING

Major Project submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by:

VODELA NITYA	**221710305060**
SIROBHUSHANAM BHAVYA	**221710305054**
PAINDLA SAI GANESH	**221710305039**
SOMAYAJULA SRI LAKSHMI	**221710305055**

Under the esteemed guidance of

Dr. S. Aparna
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF TECHNOLOGY

GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT (GITAM)
(Declared as Deemed-to-be-University u/s 3 of UGC Act of 1956)

HYDERABAD CAMPUS

MAY 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF TECHNOLOGY

GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT (GITAM)
(Declared as Deemed-to-be-University u/s 3 of UGC Act of 1956)

HYDERABAD CAMPUS

DECLARATION

We hereby declare that the Major Project entitled "Heart Disease Identification using Ensemble Learning and Deep Learning" is an original work in Department of Computer science and Engineering, GITAM School of Technology, GITAM (Deemed-to-be-University), Hyderabad submitted in partial fulfillment of the award of the degree of B.Tech. in Computer Science and Engineering. The work has not been submitted to any other college or university for the award of any degree or diploma.

Date: May 2021

Registration No.	**Name**
221710305060	VODELA NITYA
221710305054	SIROBHUSHANAM BHAVYA
221710305039	PAINDLA SAIGANESH
221710305055	SOMAYAJULA SRI LAKSHMI

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF TECHNOLOGY

GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT (GITAM)
(Declared as Deemed-to-be-University u/s 3 of UGC Act of 1956)

HYDERABAD CAMPUS

CERTIFICATE

This is to certify that the Major Project Report entitled "Heart Disease Identification using Ensemble Learning and Deep Learning" is being submitted by **CSEBNUM_E10: Vodela Nitya (221710305060), Sirobhushanam Bhavya (221710305054), Paindla Saiganesh (221710305039), Somayajula Sri Lakshmi (221710305055)** in partial fulfillment of the requirement for the award of degree of Bachelor of Technology in CSE at GITAM (Deemed to Be University), Hyderabad during the academic year 2020-21. The Mini Project has been approved as it satisfies the academic requirements.

Dr. S. Aparna
Assistant Professor
Department of CSE

Prof. S. Phani Kumar
Head of the Department
Department of CSE

ACKNOWLEDGEMENT

Apart from our effort, the success of this mini project largely depends on the encouragement and guidance of our faculty. We take this opportunity to express our gratitude to the people who have helped us in the successful competition of this mini project.

We are extremely thankful to our honorable Pro-Vice Chancellor, **Prof. N. Siva Prasad** for providing necessary infrastructure and resources for the accomplishment of our seminar.

We are highly indebted to **Prof. N. Seetharamaiah**, Principal, School of Technology, for his support during the tenure of the seminar.

We are very much obliged to our beloved **Prof. S. Phani Kumar**, Head of the Department of Computer Science & Engineering for providing the opportunity to undertake this project and encouragement in completion of our mini project.

We hereby wish to express our deep sense of gratitude to **Dr. S Aparna**, Assistant Professorand our Project Guide, Department of Computer Science and Engineering, School of Technology for the esteemed guidance, moral support and invaluable advice provided by them for the success of the Major Project as well as **Mrs. Hima Bindu**, AMC and Assistant Professor, Department of Computer Science and Engineering for always motivating us.

We are also thankful to all the staff members of the Computer Science and Engineering department who have cooperated in making our Mini Project a success.

Sincerely,

VODELA NITYA
SIROBHUSHANAM BHAVYA
PAINDLA SAIGANESH
SOMAYAJULA SRI LAKSHMI

TABLE OF CONTENTS

List of Figures

List of Tables

List of Equations

1. ABSTRACT

Technology in the field of healthcare has been in constant enhancement incorporating machine learning and various other technologies for clinical decision support systems or computer-aided healthcare systems which assist in examining the complication. Heart Disease is an issue of concern for any age group and gender, which depend on innumerable factors like cholesterol, blood pressure, chest pain and more. Furthermore, initial methods of getting a check-up for the doctor to determine whether the patient is suffering or will be suffering from the disease is time consuming as well as laborious. Considering this, by using machine learning and deep learning, it is possible to overcome the complications in order to produce an effective and accurate means of detecting heart disease. This is achieved by deploying ANN, performing feature reduction by PCA as well as creating an ensemble model of Random Forest, Decision Tree and KNN.

Keywords— Heart Disease, Ensemble Learning, Artificial Neural Networks,Feature Reduction

2. INTRODUCTION

The heart is a well-known muscular organ that is notably important for blood circulation, filtering the wastes, supplying the proper amount of oxygen and nutrients for living. While many conditions affect heart functioning, the most common type of constraint is blood vessel disease which results in heart attacks, heart failures, coronary heart disease, and many more cardiovascular diseases. Heart disease that commonly occurs in elders (i.e) in the age group of 50+ is now occurring in the younger age group, affecting babies too. The main cause of heart disease points out to anomalies in the blood vessels, heart defects by birth or later in their lifetime as it was a hereditary problem.

Acceding to this, the computer aided healthcare system developed has become popular which can help people who often hesitate to go for the checkup for minor symptoms. Sometimes these minor symptoms can become a major issue. The healthcare computer aided technology is simple and easily accessible for users and it can be a good opportunity for the users to analyse the presence of disease. Sometimes there can be a problem in computer aided technology about the accuracy of prediction. Hence this project aims to detect heart disease in an effective and accurate way with the assistance of machine learning and deep learning.

The use of Machine Learning can help prevent the errors caused as well as detect anomalies early.
In the field of healthcare, Machine Learning can be applied in various ways which provides record keeping to handle large numbers of records, predictive analysis of diseases with high accuracy. Some examples being - Heart Disease, Kidney Disease, Breast Cancer, etc.

3. LITERATURE SURVEY

Machine learning in the healthcare industry is extensively used to aid patients and support many specialists. Deep learning which is a part of Artificial intelligence provides metamorphic potential and inaugurates all the way to medical computing solutions.

Zhixu Hu *et al.*(2020)[2] Proposed a stacking ensemble model to predict seven-days ahead Hospital admissions for cardiovascular diseases ,using hospital admission data, meteorological data, and air pollution data Support Vector Machine, Decision tree, linear regression ,Random forest, gradient boosting decision tree, XG boost were trained and their predictions are combined with necessary features. The proposed stacking model has the best performance over base learners. Especially, in Cardiovascular disease dataset, the parameters for evaluation compared with Random forest , the mean absolute error (MAE),root mean square error(RMSE), and mean absolute percentage error (MAPE) of proposed model decreased by 6.3%, 7.4%, and 6.3%, and the coefficient of determination (R^2) improved by 1.7%. It indicates that the proposed model can successfully predict the number of hospitalizations.

Zhanquan sun *et al.*(2020)[7] proposed a multi label classifier to analyze ECG signals. Ensemble multi label classifiers- multi label KNN and multi label SVM are presented where the classifiers are considered as basic classifiers to increase the performance of signal categorization. The result shows that Ensemble of multi label classification methods (0.752) is better than the individual methods.

Hui yang *et al.*(2021)[5] proposed a ensemble classification algorithm which combines the classification algorithms such as Neural networks, Support Vector Machine, K-Nearest Neighbor, Adaboost, XG Boost based on ECG morphological features to detect irregularity in heart arteries and veins. The accuracy has been improved with an increase of 13.32%, 13.26% ,12.04%, 11.47%, 10.22%, and 8.43% for multi-layer feed-forward perceptrons (MLP), SVM, KNN, AdaBoost and XGBoost. The overall accuracy was 98.68% for the proposed ensemble model.

Pronab ghosh *et al.*(2021)[4] Proposed distinct supervised models such as Adaboost, decision tree, gradient boosting, KNN, and random forest. hybrid classifiers are unified with regular classifiers by using bagging and boosting methods. The final result shows that the random forest bagging method works well with selected features and achieved 99.05% accuracy.

Archana Singh *et al.*(2020)[13] used different machine learning models -SVM, Decision tree , Linear regression, Knn. The accuracy has been calculated with the support of confusion matrix and concluded that K- nearest neighbour is best among them with 87% accuracy.

4. PROBLEM IDENTIFICATION & OBJECTIVES

The increasing cases of heart disease, affecting every age group and gender unbiased as well as causing a great decline in the mortality rate has generated an overall distress. Heart disease is a complication that when the situation is time bound, it is important that the complication be eliminated sooner as a prevention from the conditions that might affect it if not addressed to. Moreover, the lack of assistance might cause an effect so severe that could make it arduous.

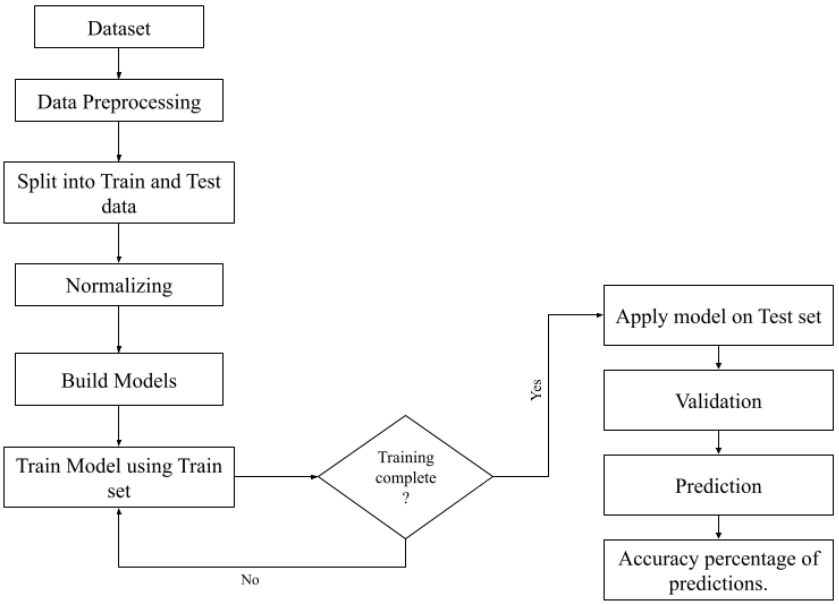
4.1. OBJECTIVES

This project is to predict if the patient suffers from Coronary Heart Disease, due to an increase in the cases of such it is difficult to get immediate assistance from the medical experts which is the motivation for this project, by developing models that can predict the occurrence and to reduce the risk of mortality rate and provide accuracy for the precisely predicted affected patients for this disease by the use of Artificial Intelligence methods which include Machine Learning and Deep Learning. By doing so, the accuracy level can be determined error free as well as help overcome the time constraint. Supervised Learning and PCA which is a Unsupervised Learning method as well as ANN has been implemented.

5. SYSTEM METHODOLOGY

The structure of the project can be better explained with the help of a flowchart. By creating a flowchart, there are higher chances of executing a project without error, furthermore, having a structured plan of execution from initiation to completion. Below portrays the flowchart:

Figure 5.1: Flowchart depicting the structure of the project



As viewed in the flowchart, the structure of the projects is initiated by importing the dataset acquired. Data cleaning is performed if there exists null values moving onto the next step which is Data Preprocessing which consists of visualising the correlation of data, Splitting it in a ratio and finally normalizing it to better fit into the models. After the models are developed it is trained using the train split of data and is validated. Once validated the models can be tested with the test split of data further validating with the help of accuracy metric. This completes the project.

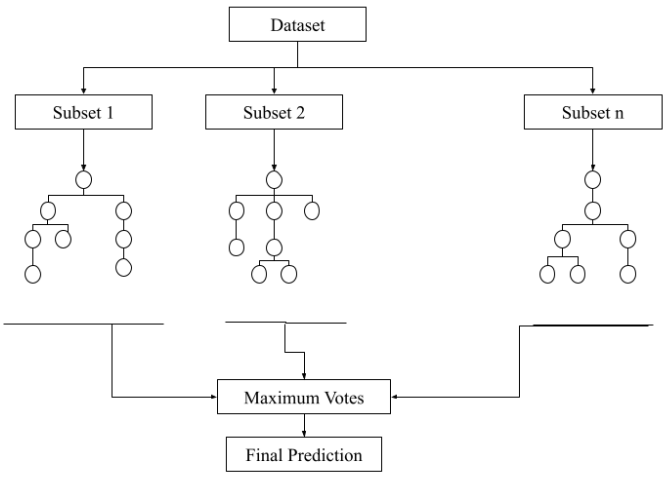
6. OVERVIEW OF TECHNOLOGIES

To briefly describe the technologies and models used, highlighting the main points are :

6.1. RANDOM FOREST

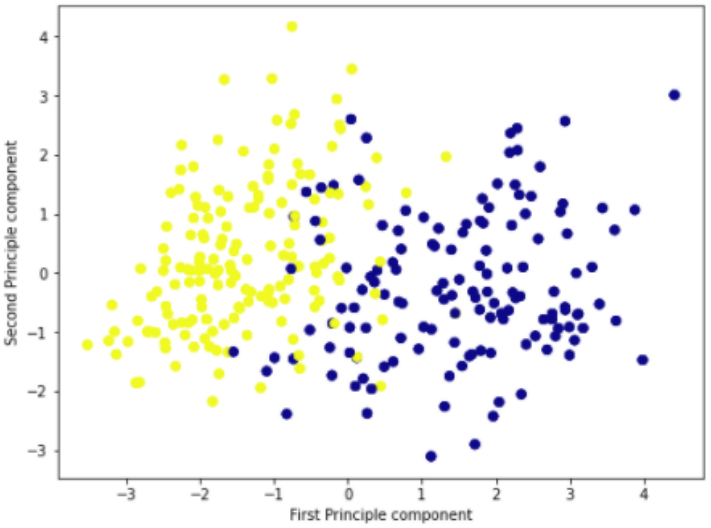
Random Forest, a supervised machine learning model which is created using an Ensemble of DecisionTrees using various subsets of the dataset. These decision trees are unique in nature. The more decision trees are formed, the better since it eliminates the constraint of overfitting.

Figure 6.1: Random Forest Structure



PCA also known as Principal Component Analysis, an Unsupervised Machine Learning Model is used for dimensionality reduction. This is performed on large datasets so as to eliminate features which would not affect the model prediction drastically. Before performing PCA, the data must be normalised to help clear the features and not obtain errors.

Figure 6.2: Graph depicting the PCA based on heart disease dataset



This method can be implemented in 2 steps :

1. **Fit** : to input the dataset in to the method
2. **Transform** : transform data based on model

By applying PCA, principal components are formed which determine the features to be selected and implemented, this is based on the eigenvalues of the said data points in the principal components. Features with maximum eigen values are considered to be best for implementing models.

Hyperparameter Tuning performed on random forest is to determine the best number of decision trees that can be obtained.

6.2. ENSEMBLE LEARNING

Ensemble Learning is the combination of 2 or more machine learning models, inheriting the features of each model and grouping them together to help enhance the accuracies of the basic machine learning models used.

The methods used in this project are:

- ➢ Random Forest - ensemble of decision tree
- ➢ Decision Tree - tree form of target value dataset
- ➢ K Nearest Neighbors - based on the value of K (odd number of nearest neighbors front aken data point) determining if the disease is present or not.

After grouping these 3 models, implement VotingClassifier which is part of the ensemble package to understand whether the patient is suffering or not based on the number of votes achieved.

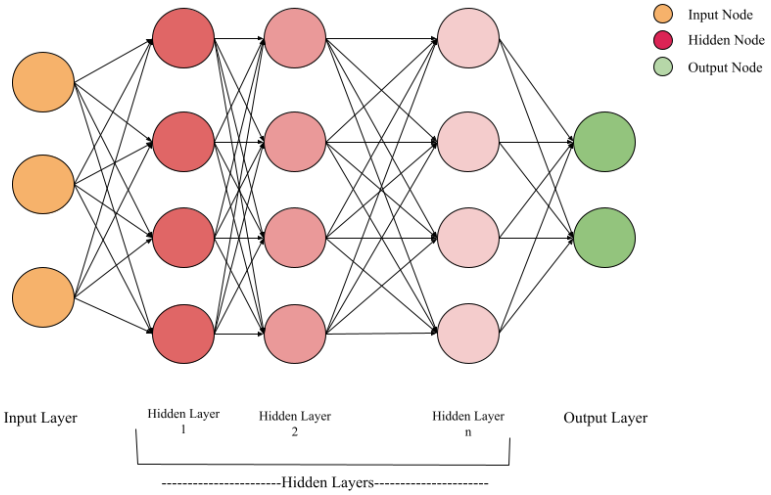
6.3. ARTIFICIAL NEURAL NETWORKS

ANN also known as Artificial Neural Network is an imitation of the human brain. Just as the brain receives signals through the synapses, ANN makes use of weights to gather input from the user. Split ratio of 80:20 for the dataset is preferred in case of ANN.

ANN has 3 layers in them:

1. Input Layer - this layer takes the input from the user.
2. Hidden Layer - this layer can contain N number of hidden layers, in these layers the input is processed and selected.
3. Output Layer - the selected inputs are then delivered as output to the user.

Figure 6.3: Layers of ANN and flow process



The deployed models in ANN are :

1. Sequential Model & Dense Layer

This is a stack of layers in the linear form which takes an input and provides the output simultaneously after processing through the layers. Dense Layer is used along with Sequential to obtain output based on activation function of input, kernel and bias.

The formula for Dense Layer is :

$$output = activation(dot(input, kernel) + bias) \quad \textbf{Equation (6.1).}$$

2. Rectified Linear Activation Function (ReLU)

Is the most used activation function in the field of neural networks as it is simple with no complex mathematics. To transform the input it uses the formula :

$$R(z) = max(0, z) \quad \textbf{Equation (6.2).}$$

By using this formula, any input in the negative form gives 0 as output and any input of the positive form gives z as the output which can range from 0 to infinity.

3. Adam Optimiser

Adam optimizer is a substitute to Stochastic Gradient Descent which has the advantage of Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). By doing so, this gives results quickly and is not complex when configuring the problems.

7. IMPLEMENTATION

7.1. DATASET

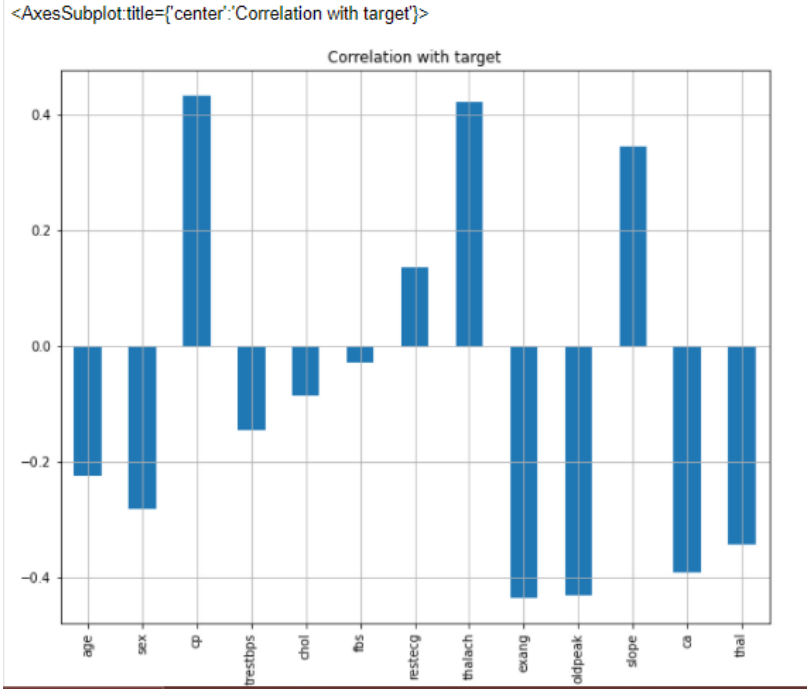
This dataset consists of 303 instances and 14 features, as mentioned in the table below:

Table 7.1: Table describing each feature in dataset

S.no	Feature Name	Description
1	Age	Age in years
2	Sex	Gender of the person
3	Cp	Type of Chest pain
4	Trestbps	Resting blood pressure
5	Chol	Serum cholesterol in mg/dl
6	fbs	Fast blood sugar
7	Restecg	Resting electrocardiographic results
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	St depression induced by exercise relative to rest looks at stress of heart
11	Slope	Slope of the peak exercise ST segment
12	Ca	Number of major vessels colored in fluoroscopy
13	Thal	Thallium stress result
14	Target	Predicted disease or No disease

Visualising the target feature is also part of preprocessing to understand the dataset better. Each feature in correlation with the target values.

Figure 7.1: Correlation graph of target with other features



Dataset is split in a 60:40 ratio consequently also normalized after splitting in order to have an organised dataset without duplicate values . The next step is to build models, train it and finally test them.

8. RESULTS & DISCUSSIONS

The results generated were higher compared to traditional machine learning models deployed individually, deeming that by developing models that are conglomerated will be progressive as well as effective. In doing so, further developments can be done by deeper research. Below is the table comparing our proposed models to previously deployed models:

Table 8.1: Comparison table with previous work

	Model Names	**Accuracy Percentage**
Archana *et al.*	Linear Regression	81.42%
	SVM	83.61%
	KNN	82.30%
	Decision Tree	76.99%
S. Mohan *et al.*	HRFLM- Hybrid Random Forest with Linear Model	88.40%
Proposed Models	Random Forest + PCA + Hyperparameter tuning	92.98%

	ANN	86.89%
	Ensemble	88.52%

As compared, our proposed models have performed better than the machine learning models. The best performing model is Random Forest with PCA and hyperparameter Tuning which in comparison with HRFLM[1] has achieved 5% higher.

9. CONCLUSION & FUTURE SCOPE

In conclusion, by applying the said proposed models, it has enhanced the accuracies as compared to traditional machine learning models. Random forest classifier is as effective as the machine learning models, achieving 84.07 %, furthermore, upon application of PCA and hyperparameter tuning, the results obtained is 91.2%. Ensemble model obtained 88.52% and ANN has achieved 86.89%. The average of these proposed models 87.70% .

For future work, further implementation of hybrid and ensemble of models and use of Security and Data Mining techniques will be progressive. Inclusions of IOT techniques do make the work easier while assisting the healthcare industry as a technological revolution.

10. REFERENCES

[1] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, **vol. 7, pp. 81542-81554, 2019**, doi: 10.1109/ACCESS.2019.2923707.

[2] Z. Hu, H. Qiu, Z. Su, M. Shen and Z. Chen, "A Stacking Ensemble Model to Predict Daily Number of Hospital Admissions for Cardiovascular Diseases," in *IEEE Access*, **vol. 8, pp. 138719-138729, 2020**, doi: 10.1109/ACCESS.2020.3012143.

[3] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, **vol. 8, pp. 133034-133050, 2020**, doi: 10.1109/ACCESS.2020.3010511.

[4] P. Ghosh *et al*., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in *IEEE Access*, **vol. 9, pp. 19304-19326, 2021**, doi: 10.1109/ACCESS.2021.3053759.

[5] H. Yang and Z. Wei, "A Novel Approach for Heart Ventricular and Atrial Abnormalities Detection via an Ensemble Classification Algorithm Based on ECG Morphological Features," in *IEEE Access*, **vol. 9, pp. 54757-54774, 2021**, doi: 10.1109/ACCESS.2021.3071273.

[6] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest With

an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, **vol. 8, pp. 59247-59256, 2020,** doi: 10.1109/ACCESS.2020.2981159.

[7] Z. Sun, C. Wang, Y. Zhao and C. Yan, "Multi-Label ECG Signal Classification Based on Ensemble Classifier," in *IEEE Access*, **vol. 8, pp. 117986-117996, 2020**, doi: 10.1109/ACCESS.2020.3004908.

[8] J. Zhang et al., "Coupling a Fast Fourier Transformation With a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment," in IEEE Access, **vol. 5, pp. 10674-10685, 2017,** doi: 10.1109/ACCESS.2017.2706318.

[9] S. A. Ali et al., "An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzo–Tompa and Stacked Genetic Algorithm," in IEEE Access, **vol. 8, pp. 65947-65958, 2020,** doi: 10.1109/ACCESS.2020.2985646.

[10] Kathleen H. Miao, Julia H. Miao and George J. Miao, "Diagnosing Coronary Heart Disease using Ensemble Machine Learning" *International Journal of Advanced Computer Science and Applications (IJACSA)*, **(2016)**. http://dx.doi.org/10.14569/IJACSA.2016.071004

[11] R. Sateesh Kumar, S. Sameen Fatima, Anna Thomas, "Heart Disease Prediction using Ensemble Learning Method", *in International Journal of Recent Technology and Engineering (IJRTE)*,**Volume-9 Issue-1 (May 2020), pp. 2612-2616**, doi: 10.35940/ijrte. A2997.059120

[12] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," *in IEEE Access*, **vol. 7, pp. 180235-180243 (2019)**, doi: 10.1109/ACCESS.2019.2952107.

[13] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, **2020, pp. 452-457,** doi: 10.1109/ICE348803.2020.9122958.

[14] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *in IEEE Access*, **vol. 8, pp. 107562-107582 (2020)**, doi: 10.1109/ACCESS.2020.3001149.

[15] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," *in IEEE Access*, **vol. 7, pp. 54007-54014 (2019)**, doi: 10.1109/ACCESS.2019.2909969.