# Heart Disease Detection using Ensemble Learning Approach

**Pruthvirajsinh Puvar[1], Neel Patel[2], Akshay Shah[3], Ruturajsinh Solanki[4], Dhaval Rana[5]**

*[1,2,3,4] Student, Department of Computer Science and Engineering, RNGPIT, Bardoli, Surat, India*
*[5]Assistant Professor, Department of Computer Science and Engineering, RNGPIT, Bardoli, Surat, India*

---***---

**Abstract -** *With the rampant rise in the heart stroke rates at small ages, we need to put a system in place to be able to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is difficult for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of a heart disease. Thus, we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, diabetes, blood pressure, cholesterol etc. The ensemble learning approach is one of the most reliable techniques for predicting results.*

***Key Words:*** **Ensemble learning, classification algorithms, regression algorithms, heart disease, prediction**

## INTRODUCTION

The world health organization (WHO) delineated vessel diseases (CVDs) because the leading reason for death ecumenically. Coronary heart disease could be a variety of CVD that accounts for four out of 5 CVD deaths [1]. Identifying individuals in danger of cardiopathy and ensuring they receive correct treatment will stop these deaths. Other than the traditional diagnosing ways, there are many computation techniques, together with machine learning accustomed establish individuals in danger. Meanwhile, researchers have engineered many machine learning models exploitation out there heart condition risk datasets and obtained varied performances [2]–[6].

Machine learning-based strategies are adopted in several areas of life science. However, researchers are perpetually trying to find ways in which to optimize and improve these strategies. Ensemble learning is one such approach that's tried to reinforce machine learning tasks [7]. An ensemble classifier may be a set of individual classifiers beside a mechanism, like majority voting that mixes the predictions of the parts. Research has shown that ensemble classifiers typically perform higher than typical classifiers [8]. Homogeneous ensemble learning consists of members having one base learner or algorithmic program. Meanwhile, the members would possibly take issue in structure. Whereas, heterogeneous ensemble includes of members having completely different base learners.

Motivated by the event of many machine learning ways for the prediction of heart condition risk, and during a bid to enhance the classification performance, we have a tendency to propose a sort of homogeneous ensemble learning methodology. The proposed methodology involves the employment of a mean based approach to randomly partition the dataset into smaller subsets and applying the classification and regression algorithmic program to model every partition.

## RELATED WORK

This section discusses some ensemble learning strategies. Machine learning algorithms that perform classification are widely utilised in numerous fields. Hence, researchers are endlessly formulating new techniques to boost the classification performance. One such technique is ensemble learning which is either homogeneous or heterogeneous. Early samples of ensemble learning methods are bootstrap aggregating (bagging) [10], boosting [11], and random decision forests [9]. Improved classification performance is typically obtained once these ensembles ar used. However, many alternative researchers have utilized alternative ways to realize ensemble learning, as well as strategies that commit to mix several classifiers or partitions victimisation majority vote and alternative techniques.

Recently, Leon et al. [12] conducted a study to investigate the impact of voting techniques on classification tasks. The study evaluated the impact of various voting techniques on the performance of 2 algorithms applied to datasets with totally different levels of problem. Although majority voting is common in the literature, experimental results show that the only technique taught is usually different from the honest technique. In a similar study, Banfield et al. [13] carried out a comparative study of ensemble decision tree methods, including bagging and other seven randomization based methods for creating decision tree ensembles. The experiment was conducted on a public data set, and the results of random forests and boosting were better than bagging.

While some ensemble learning strategies tend to specialize in combining completely different base classifiers, it is, however, doable to make ensembles by partitioning the dataset into subsets and mixing the assorted partitions. In addition, you usually get a completely different data set without bagging, boosting or random forest structure. Ruta et al. [14] used random arrangement and division of the original data set to create various data sets. The ensuing ensemble achieved sensible generalization because of the considerably boosted compatibility of the individual models. Lastly, ensemble learning has been applied in many medical diagnostic tasks [15], [16]. In this work, we hope to build on the completed work and develop a fully functional ensemble learning model to predict the risk of heart disease.

## PROPOSED METHODOLOGY

### 3.1 Classification and ensemble algorithms

Classification is a supervised learning process used to predict outcomes based on available data. This article proposes a method for diagnosing heart disease using classification algorithms and using many classifiers to improve classification accuracy. Dataset is divided into two parts: training set and test set. Every classifier is trained using training data set. Use the test data set to check the efficiency of the classifier.

### 3.1.1 Neural Networks

Convolutional Neural Networks (CNN) are used to develop early medical diagnosis and prediction systems. Convolutional neural network algorithm is a tool that uses structured data to determine the risk of heart disease early. Neural networks can increase the accuracy through training. You can continue after the training. As mentioned above, neural networks can be used in parallel to improve performance. The error rate is low therefore it gives higher accuracy with proper training. [19]

### 3.1.2 Logistic regression

Logistic regression is a statistical and machine learning technique that can classify data sets in a data set based on the values of input fields. Predict the dependent variable based on one or more sets of explanatory variables to predict the outcome. It can be used for both binary classifications and multi-class classification. Logistic regression is one of the machine learning algorithms, which is relatively widely used in research to assess the risk of complex diseases. Therefore, the purpose of this study is to determine the most important predictors of cardiovascular disease and use logistic regression analysis to predict the overall risk. [20]

### 3.2 Ensemble Techniques

The ensemble method is most suitable for improving the accuracy of the classifier. This is an impressive meta-classification approach that combines weak learners with strong learners in order to improve the performance of weak learners. Ensemble learning approach is used to improve the accuracy of various algorithms to predict heart disease. The purpose of combining multiple classifiers is to improve the performance. This gives better results than individual classifiers. [18] The ensemble learning process is shown in Fig. 1.

### 3.2.1 Boosting

Boosting is an algorithm used to build an ensemble learning model. The real data set is divided into multiple datasets called subsets.The classifier is trained on a subset to create a series of moderately effective models. Create a new subset based on elements that were not correctly classified by the previous model. Then, the ensembling process improves the performance of the model by integrating the weak model with the cost function. [18] The boosting algorithm is shown in Fig. 2.

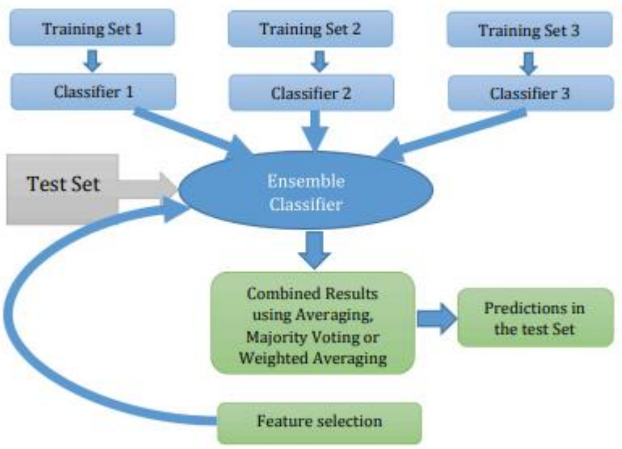
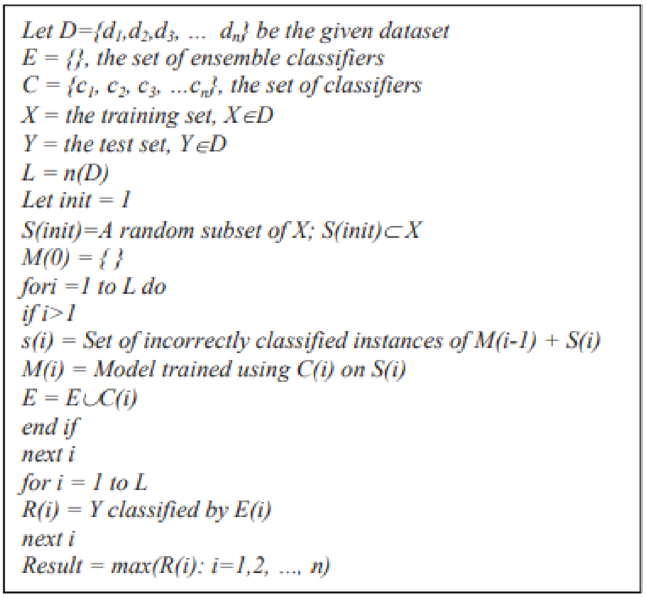


**Fig. 1 –** Ensemble Classifier

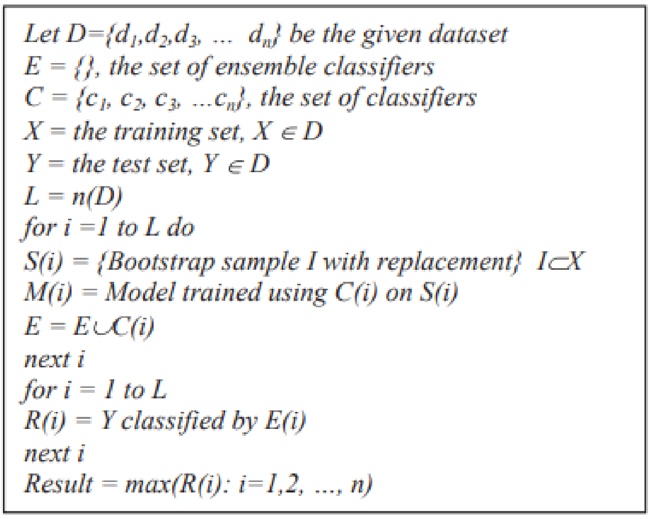


```
Let $D=\{d_1,d_2,d_3, \dots d_n\}$ be the given dataset
$E = \{\}$, the set of ensemble classifiers
$C = \{c_1, c_2, c_3, \dots c_n\}$, the set of classifiers
$X$ = the training set, $X \in D$
$Y$ = the test set, $Y \in D$
$L = n(D)$
Let init = 1
$S(init)$=A random subset of X; $S(init) \subset X$
$M(0) = \{ \}$
fori =1 to L do
if i>1
$s(i)$ = Set of incorrectly classified instances of $M(i-1) + S(i)$
$M(i)$ = Model trained using $C(i)$ on $S(i)$
$E = E \cup C(i)$
end if
next i
for i = 1 to L
$R(i)$ = Y classified by $E(i)$
next i
Result = $max(R(i): i=1,2, \dots, n)$
```

**Fig. 2 –** Boosting Algorithm

### 3.2.2 Bagging

Bagging is also called as bootstrap aggregation. Random bagging considers multiple patterns from the replacement training set. The newly created training set contains the same number of patterns as the original training set, but there are some gaps and repetitions. The new training set is called a bootstrap copy. During bagging, bootstrap samples are collected from the data, and the training of classifier is performed on individual samples. Combine the votes of each classifier and select the classification results based on the majority voting. Optimally improve the performance of weak classifiers. Bagging generates multiple records by replacing a random sample of the original records, thereby reducing variance of prediction. [18] The bagging algorithm is shown in Fig 3.

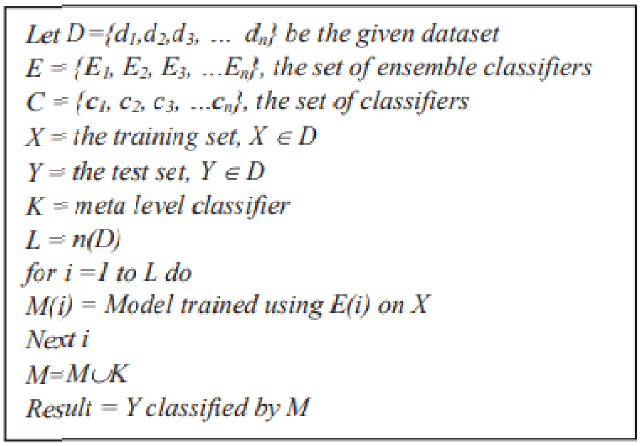


```
Let $D=\{d_1,d_2,d_3, \dots\ d_n\}$ be the given dataset
$E = \{\}$, the set of ensemble classifiers
$C = \{c_1, c_2, c_3, \dots c_n\}$, the set of classifiers
$X$ = the training set, $X \in D$
$Y$ = the test set, $Y \in D$
$L = n(D)$
for $i = 1$ to $L$ do
$S(i)$ = {Bootstrap sample $I$ with replacement} $I \subset X$
$M(i)$ = Model trained using $C(i)$ on $S(i)$
$E = E \cup C(i)$
next $i$
for $i = 1$ to $L$
$R(i)$ = $Y$ classified by $E(i)$
next $i$
Result = $\max(R(i)$: $i=1,2, \dots, n)$
```

**Fig. 3 –** Bagging Algorithm

### 3.2.3 Stacking

Stacking is an ensemble technique that uses a meta-classifier to combine multiple classification models. Multiple levels are placed in sequence, and each model communicates its prediction information to the above model, and the top model makes decisions based on the models below it. The low-level model receives input features from the original data set. The top model gets the output from the lower layer and makes predictions. The stacking algorithm is shown in the Fig. 4.Original data is provided as input in stacking for several separate models. Then, the meta-classifier is used to evaluate the input and output of each model, and the weights are estimated. Choose the most effective model and discard the others. Stacking combines several basic classifiers trained with different learning algorithms L into a single data set S using meta classifier. [18] The stacking algorithm is shown in Fig 2.



```
Let $D=\{d_1,d_2,d_3, \dots\ d_n\}$ be the given dataset
$E = \{E_1, E_2, E_3, \dots E_n\}$, the set of ensemble classifiers
$C = \{c_1, c_2, c_3, \dots c_n\}$, the set of classifiers
$X$ = the training set, $X \in D$
$Y$ = the test set, $Y \in D$
$K$ = meta level classifier
$L = n(D)$
for $i = 1$ to $L$ do
$M(i)$ = Model trained using $E(i)$ on $X$
Next $i$
$M = M \cup K$
Result = $Y$ classified by $M$
```

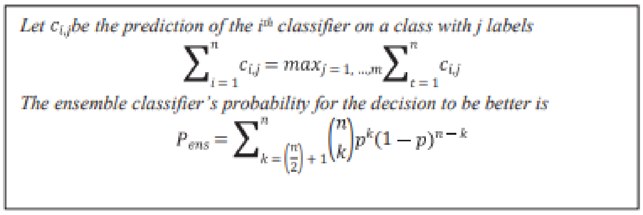
**Fig. 4 –** Stacking Algorithm

### 3.2.4 Majority Vote

The majority voting classifier is a meta-classifier that combines each majority voting classifier with it. The final class label is the class label predicted by most classifiers The final class label $d_J$ is defined as

$$d_J = \text{mode}\ \{C_1, C_2, \dots, C_n\}$$

Where $\{C_1, C_2, \dots, C_n\}$ represents the individual classifiers that participate in the voting. The majority voting algorithm is shown in Fig. 5. [18]



Let $c_{i,j}$ be the prediction of the $i^{th}$ classifier on a class with $j$ labels

$$\sum_{i=1}^{n} c_{i,j} = max_{j=1,\dots,m} \sum_{t=1}^{n} c_{i,j}$$

The ensemble classifier's probability for the decision to be better is

$$P_{ens} = \sum_{k=\left(\frac{n}{2}\right)+1}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

**Fig. 5 –** Majority vote Algorithms

## DATASET AND PERFORMANCE METRICES

In this paper, one heart disease dataset is used, the Framingham dataset obtained from the Kaggle web site [17]. The previous contains 303 instances and 14 attributes, whereas the latter consists of 4238 instances and 16 attributes. The Framingham dataset contains missing attributes, and it's preprocessed to form it appropriate for machine learning. This dataset embody demographic and health records like age, sex, cholesterol level, blood pressure, diabetes, etc. For our experiments, the 75-25 train-test holdout validation scheme is used; this is often to enable us to form a good and higher comparison between our proposed technique and former studies that used an analogous dataset. To adequately assess the performance of the proposed methodology, numerous performance indices are used, including accuracy, precision, sensitivity, specificity and F1 score. The accuracy is the proportion of the total number of predictions that were correct, and precision is the magnitude relation of correct positive predictions to the number of positive results predicted. At the similar time, Sensitivity (True Positive rate) measures the proportion of positives that are correctly identified, while Specificity (True Negative rate) measures the proportion of negatives that are correctly identified and F-score is the harmonic mean between precision and sensitivity. The mathematical representations of these performance metrics are:

*Accuracy = (TP+TN) / (TP+FP+FN+TN)* (1)
*Precision = TP/(TP+FP)* (2)
Sensitivity = *TP/(TP+FN)* (3)
*Specificity = TN/(TN+FP)* (4)
*F1 Score=2*(*Sensitivity**Precision)/(*Sensitivity + *Precision)*
*= 2TP/(2TP+FP+FN)* (5)

Where, TP represents the number of true positives,
TN represents the number of true negatives,

FP represents the number of false positives and
FN represents the number of false negatives.

## RESULTS AND DISCUSSION

To validate the effectiveness of the proposed method, a comparative study is conducted with other well-known machine learning methods. The methods include k-nearest neighbor (KNN), Logistic regression (LR), linear discriminant analysis (LDA), support vector machine (SVM), classification and regression tree (CART), gradient boosting (GB), and random forest (RF). Fig. 6 summarize the test results of the various methods on the Framingham test sets.

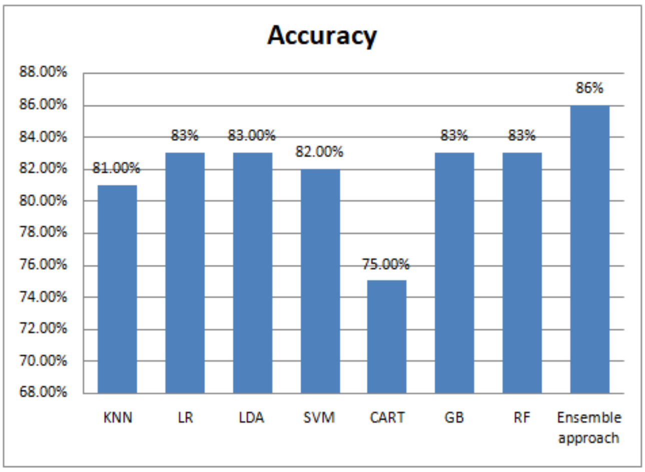
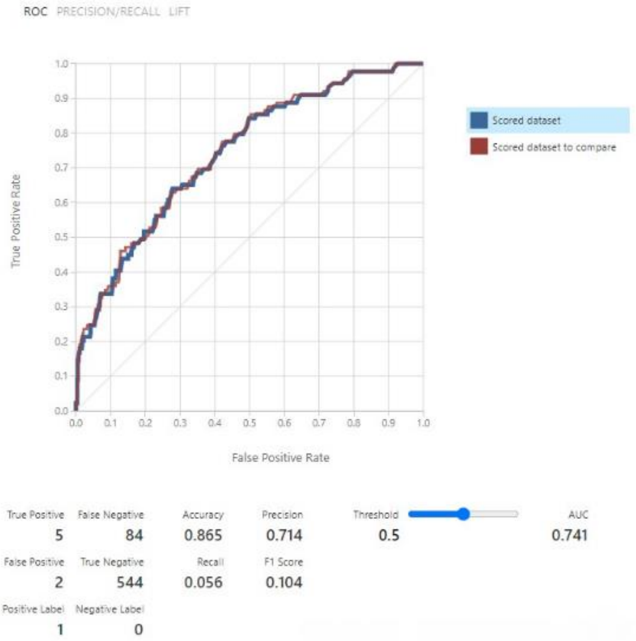


**Fig. 6 –** Accuracy Graph

From Fig. 6, it is evident that the proposed method achieved superior classification performance on the Framingham test sets with accuracy of 86%. Also, from the results, it can be observed that the ensemble learning methods, i.e., the Gradient Boosting and Random Forest, performed better than the other algorithms. The performance of these ensembles, together with the proposed method, is further studied with the receiver operating characteristic (ROC) curves. The ROC curves are useful for evaluating the predictive ability of the various ensemble models. They are created by plotting the true positive rate against the false positive rate at various threshold settings. The ROC curves for the ensemble approach is shown in Fig. 7.



True Positive 5
False Negative 84
Accuracy 0.865
Precision 0.714
Threshold 0.5
AUC 0.741
False Positive 2
True Negative 544
Recall 0.056
F1 Score 0.104
Positive Label 1
Negative Label 0

**Fig. 7 –** ROC Curve

## CONCLUSION

Heart disease is one of the leading causes of death worldwide. Early diagnosis can help prevent the disease from getting worse. An ensemble learning approach was proposed to predict heart disease effectively. The proposed ensemble achieved accuracy of 86.32% on Framingham test sets using heroku cloud services. Furthermore, the proposed method can be used to predict heart disease risk and aid in clinical advising efficiently.

## ACKNOWLEDGMENT

We thank Mr. Dhaval J. Rana for his assistance with paper and for helping out as our mentor/guide throughout the process.

## REFERENCES

[1] ”cardiovascular disease” https://www.who.int/westernpacific/healthtopics/cardiovascular-diseases (accessed Apr. 10, 2020).

[2] Maalej, Ramzi & Parchur, Abdul. (2017). Editorial Special Section on Engineering Sciences in Biology and Medicine. IEEE Transactions on NanoBioscience. 16. 647-649. 10.1109/TNB.2017.2772378.

[3] Jin, Bo & Che, Chao & Liu, Zhen & Zhang, Shulong & Yin, Xiaomeng & Wei, X.. (2018). Predicting the Risk of Heart Failure With HER Sequential Data Modeling. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2789324.

[4] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, “An Automated Diagnostic System for Heart Disease Prediction Based on $\chi^2$ Statistical Model and Optimally Configured Deep Neural Network,” IEEE Access, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

[5] C. B. C. Latha and S. C. Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,” Informatics in Medicine Unlocked,

vol. 16, p. 100203, Jan. 2019, doi: 10.1016/j.imu.2019.100203.

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

[7] R. K. Sevakula and N. K. Verma, "Assessing Generalization Ability of Majority Vote Point Classifiers," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 12, pp. 2985–2997, Dec. 2017, doi: 10.1109/TNNLS.2016.2609466.

[8] H. Li et al., "Ensemble Learning for Overall Power Conversion Efficiency of the AllOrganic Dye-Sensitized Solar Cells," IEEE Access, vol. 6, pp. 34118–34126, 2018, doi: 10.1109/ACCESS.2018.2850048.

[9] Tin Kam Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.

[10] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1023/A:1018054314350.

[11] R. E. Schapire and Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions," Machine Learning, vol. 37, no. 3, pp. 297–336, Dec. 1999, doi: 10.1023/A:1007614523901.

[12] F. Leon, S.-A. Floria, and C. Bădică, "Evaluating the effect of voting methods on ensemble-based classification," in 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Jul. 2017, pp. 1–6, doi: 10.1109/INISTA.2017.8001122.

[13] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A Comparison of Decision Tree Ensemble Creation Techniques," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 173–180, Jan. 2007, doi: 10.1109/TPAMI.2007.250609.

[14] D. Ruta, B. Gabrys, and C. Lemke, "A Generic Multilevel Architecture for Time Series Prediction," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 3, pp. 350–359, Mar. 2011, doi: 10.1109/TKDE.2010.137.

[15] B. Zhang et al., "Ensemble Learners of Multiple Deep CNNs for Pulmonary Nodules Classification Using CT Images," IEEE Access, vol. 7, pp. 110358–110371, 2019, doi: 10.1109/ACCESS.2019.2933670.

[16] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes," IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 2, pp. 728–734, Mar. 2015, doi: 10.1109/JBHI.2014.2325615.

[17] "Framingham Heart study dataset." https://kaggle.com/amanajmera1/framingham-heartstudy-dataset (accessed Jan. 24, 2020).

[18] C. Beulah Christalin Latha, S. Carolin Jeeva," Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Informatics in Medicine Unlocked, Volume 16, 2019, 100203

[19] Singhal, Shubhanshi & Kumar, Harish & Passricha, Vishal. (2018). Prediction of Heart Disease using CNN.

[20] https://medium.com/@babatolatemi/heart-disease-prediction-a-logistic-regression-implementation-from-python-scikit-learn-c4eb391a873f