# Heart Disease Prediction Using Machine Learning Techniques

P. Mohana Vamsi Shardul Thakare Ashish Mishra Suryansh Raj B. Sangeetha

***Abstract*—Heart diseases are the number one cause of death of people around the world. They are caused by the disorders related to the heart and blood vessels. Huge amount of patient related data is maintained in the health industry. This stored data can be useful for predicting future diseases. In this research paper, we try to focus on various machine learning algorithms that efficiently predict the heart diseases. The complex task of finding heart diseases requires lots of experience and knowledge. Some of the many ways of predicting it are stress test, ECG and heart MRI etc. Here, our model uses 13 parameters for predicting heart disease that include blood pressure, chest pain, cholesterol level, heart rate. These attributes are used to improve the accuracy levels and make accurate predictions. The main aim this paper is to provide analysis of various machine learning models like SVM, k-NN, Decision tree, Random Forest, XG-Boost, Cat-Boost, Light GBM, Artificial Neural Networks (ANN) and using stacking to create models for efficient prediction of Heart diseases.***

Keywords— ANN, Catboost, Classification, Decision Tree, Ensemble Models, Heart Disease, KNN, LightGBM, Machine Learning, Neural Networks, Random Forest, SVM, XGBoost

## I. Introduction

The heart is one of the most important organs in the human body. Life is dependent on the heart functioning well. Cardiovascular diseases are various diseases that affect the functioning of the heart negatively. They are the number one cause of deaths around the world [1]. Each year, around 17.9 million people die due to heart diseases.

People with cardiovascular diseases or those who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management using counselling and medicines, this is why prediction is so important when it comes to these diseases. We plan to achieve that by using machine learning algorithms. This is a tough task as it poses a lot of challenges such as the high risk in the event of a false negative, which could even be fatal, or not having relevant data provided for. Due to digital technologies growing rapidly, healthcare centres store huge amount of data in their database that is very complex and challenging to analyse.

Selecting the relevant features and working on them is very important to find good prediction models. The common attributes used include age, sex, chest pain type, resting blood pressure, cholesterol levels, electrocardiograph results(resting), number of major vessels coloured by fluoroscopy, and so on. Various algorithms were tried, including KNN, Random Forest, Decision Trees, SVM, XGBoost, Cat-Boost, LightGBM and artificial neural networks. All of these approaches were tried since the most effective model was to be found. Least amount of log loss was needed as the Stacking and ensemble modelling was eventually used to find the best models possible.

## II. Related Work

K. Polaraju et. al [2], proposed the usage of Multiple Regression Models to predict Heart Diseases. The datasat used, the training data, consisted of 3000 instances. It had 13 features and those were used as the basis for training the model. Training data made up for 70% of the whole dataset whereas the rest, i.e., 30% was sent into the testing data. Analysis of the results leads us to believe that regression is better than majority of other models while predicting cardiovascular diseases.

Das et. al [3] used an ensemble method consisting of neural networks based on SAS software 9.1.3 for diagnosing of the heart disease. It combined posterior probabilities or the predicted values from multiple predecessor models. The classification accuracy obtained by them using this ensemble model was 89.01%. This was trained and tested on the Cleveland heart disease dataset.

Anbarasi et. al [4] used almost a 2 step approach where first they used 3 classifiers, namely, decision trees, Naïve Bayes and classification by clustering using thirteen attributes. In the next step, they applied feature subselection using genetic algorithm and obtained almost similar accuracy. Their results showed that the highest accuracy was attained by decision tree classifier, with an accuracy of 99.2% for binary classification. This was followed by the Naïve Bayes classifier with an accuracy of 96.5% and then by the classification clustering, with 88.3%.

Zhang et. al [5] used SVM (Support Vector Machines) as the model for cardiovascular disease prediction. PCA (Principal Component Analysis) was used by them to find the relevant and important features and different kernel functions were used as classifier. Radial Basis Function (RBF) gave the highest

classification accuracy. Grid search method was employed to find the optimal parameters values, and optimal values were found to be c=1 and g=0.0909. The highest classification accuracy reached in binary classification was 88.6364%.

Noura Ajam [6] experimented with artificial neural network for heart disease diagnosis. Feed-forward Back propagation learning algorithms have been used to train and test the model based on the requirements. On finding the appropriate parameters, classification accuracy reached to 88% with 20 neurons in hidden layer. ANN shows significantly optimistic results for cardiovascular prediction.

Elshazly et. al [7] proposed a novel approach for classification, GA-SVM, for diagnosis of lymph diseases. Genetic Algorithm (GA) was used to reduce the features from 18 to 6 in the given dataset. 10-fold cross validation was used to validate while training during the experiment. Different kernel functions were employed and for each function, performance was evaluated by measures like accuracy, sensitivity, area under curve (AUC), F-measure. Linear classifier achieved the best results with an accuracy score of 83.1%, F-measure of 82.7%, AUC of 84.9% and sensitivity of 82.6%.

Dey et. al [8] used multiple models including SVM, Naive Bayes and Decision tree with and without applying PCA for attribute selection to predict heart disease. The dataset contains two types of people, those who have a heart disease and those who don't, so it's a binary classification problem. Final observations indicated that SVM outperformed the other two and was the best choice for the classifier.

Weng et. al [9] used a similar approach to Dey et. al. where they compared random forest, gradient boosting, neural networks and logistic regression. These 4 machine learning algorithms were used to try and predict cardiovascular diseases. Grid search was used for parameter optimization.

Our approach is different from these and similar in a way as well since it shares similarities in some aspects with some of these papers. We use an ensemble model like Das et al. used but other than that we are also using algorithms such as XGBoost and LightGBM which haven't really been used in this way in ensemble with an SVM so that makes our approach unique. Also, the metric that we used for evaluation is Log Loss, which heavily penalizes large deviations in incorrect predictions.

This leads to our model being different than a lot of approaches tried beforehand. Making an ensemble is the major step which led to the most improvement that we have had till now in our experiments.

## III. METHODOLOGY

Our methodology for the final model involves us taking an ensemble of the three algorithms, SVM, LightGBM and XGBoost by averaging the prediction probabilities out and then passing that to the log loss equation.

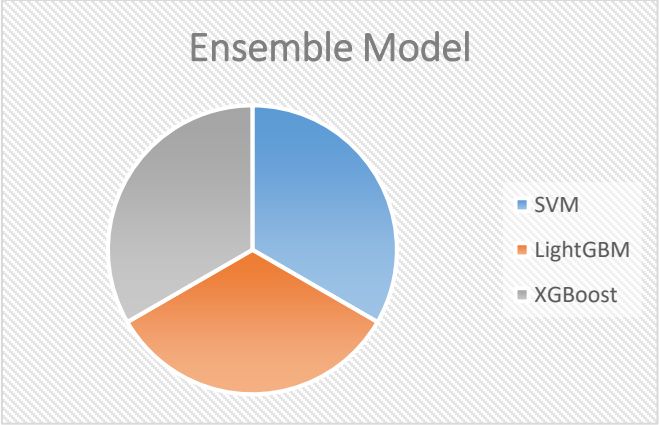


Fig 1. Ensemble model composition

As you can see in this chart, SVM, LightGBM and XGBoost were used in ensemble using the averaging method to get our model.

SVM: SVM refers to Support Vector Machines. An n-dimensional space is formed by the numerical input variables(x). Two input variables would form a two dimensional space, so to say. A hyperplane is, we can visualize it as a line that splits the input variable space. In SVM, we select a hyperplane which will best split the given space into two spaces, 1 and 0. In two-dimensions, we can visualize this as a line and let's assume that all our input points can be completely separated by this line. For example:

$$B_0 + (B_1 * X_1) + (B_2 * X_2) = 0$$

Where the coefficients ($B_1$ and $B_2$) that determine the slope of the line and the intercept ($B_0$) are found by the learning algorithm, and $X_1$ and $X_2$ are the two input variables.

On trying the various kernels, RBF was found to be the most accurate one, and the one that gave the least log loss.

LightGBM: Light GBM is a gradient boosting framework based on decision tree algorithm, which we have used for classification here. One standout feature of LightGBM is that it splits the tree leaf wise as opposed to other boosting algorithms that split it depth wise or level wise. It is based on decision tree.

XGBoost: XGBoost stands for eXtreme Gradient Boosting. XGBoost is an optimized distributed gradient boosting library. It is designed to be highly efficient and flexible. It implements machine learning algorithms under the Gradient boosting framework. XGBoost provides parallel tree boosting that solves problems in a fast and accurate way.

We used an ensemble of these three methods by averaging out the predicted probabilities and getting the log loss. Which in this case, we got as 0.28504. This was the final model that gave us the best results.

We began our experimentations by working with the dataset, cleaning it up, one-hot encoding the categorical variables, normalizing with standardscaler and splitting it into 70% training and 30% testing dataset. Following that began experimentations with SVM and Neural networks at first. They were tried separately, with SVM giving the least log loss among the two at first, optimized using grid search. After optimizing the hyper parameters of NN, we got it close to SVM but it could never cross it. So we decided to go with SVM. After that we tried various algorithms, random forest classifier, decision tree, KNN, stacking of various models and logistic regression. None of them gave results close to SVM even after grid searching. This was followed by LightGBM, which improved upon SVM by quite a bit so we kept working with that. XGBoost was tried but couldn't come close to LightGBM's results, the log loss. After that an ensemble of various algorithms was tried, at the end of which the combination of XGBoost, LightGBM and SVM gave the least log loss.

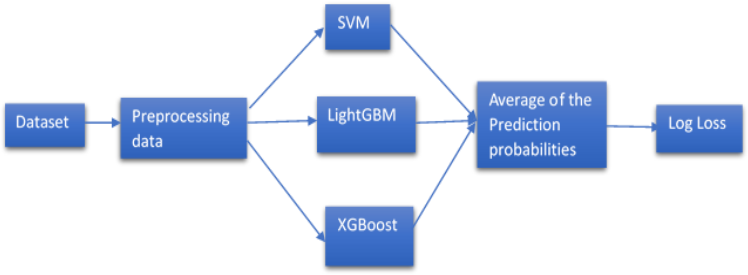


Fig 2. Model architecture

## IV. EXPERIMENTAL RESULTS

The dataset used was the dataset given over at the competition on predicting heart diseases using machine learning at drivendata.

The dataset has 13 essential features, and 180 instances. Patient id is a feature as well but it must be dropped.There features are not dependent on each other.

The features include:

1. Patient_id
2. Slope of the peak exercise ST segment
3. Thal
4. Resting Blood Pressure
5. Chest pain type
6. Number of major vessels colored by fluoroscopy
7. Fasting Blood Sugar
8. Resting Ekg Results
9. Serum Cholesterol in mg/dl
10. ST depression induced by exercise relative to rest
11. Sex
12. Age
13. Max Heart Rate achieved
14. Exercise induced angina

The metric used for the competition is logarithmic loss.

$Log\ loss = \sum_{i=1}^{n} [y_i\, log(y`_i)+(1-y_i)log(1-y`_i)]$

This is the formula for log loss. $y`$ is the probability that $y=1$. Logarithmic loss provides a steep penalty for predictions that are both confident and wrong. The goal is to minimize the log loss.

The testing procedure involved us passing the data through the model, noting down the accuracy and log loss and repeating it for the various models.

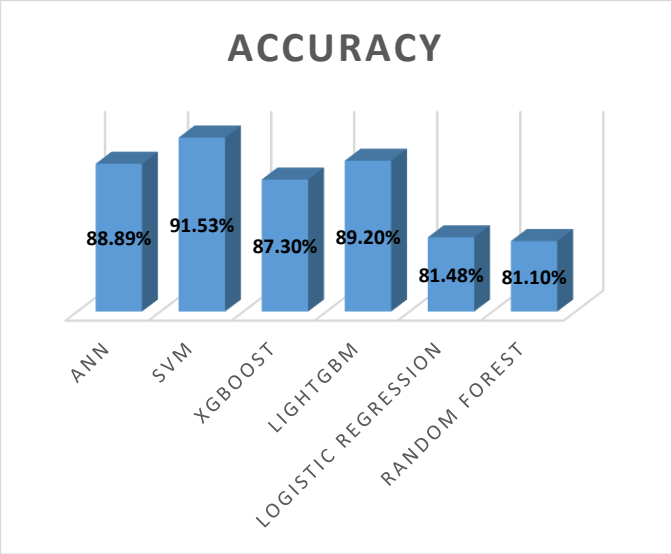


Fig 3. Accuracy results

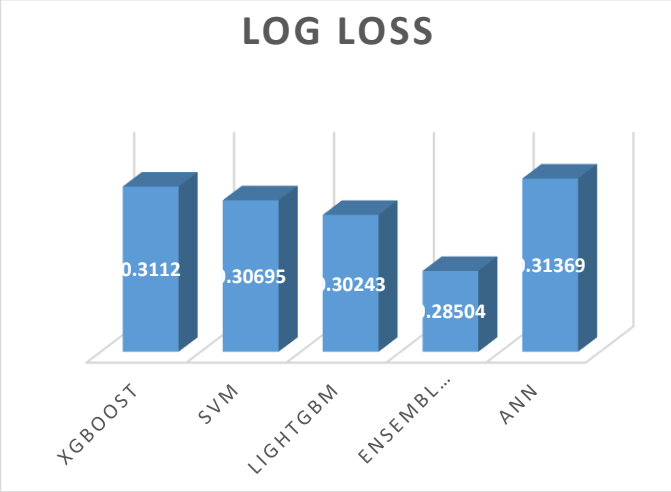


Fig 4. Log loss results

These two graphs show us the best metrics that we have gotten from our models. Log loss is the defining metric but we included the top 6 performers on basis of accuracy as well.

## V. CONCLUSION AND FUTURE WORK

This experiment provides the deep insight into machine learning techniques for classification of heart diseases. After

applying numerous models, and defining log loss as the evaluation metric, we have found the ensemble of SVM, LightGBM and XGBoost to be the best model for achieving the metric laid down. The models that we tried are, XGBoost, LightGBM, SVM, ANNs, Random Forest, Logistic Regression and Decision Trees. Due to the small size of the dataset given in the competition, Neural Networks could overfit quite easily, and the regularization applied to counter that would bring down the accuracy by quite a bit, resulting in relatively simpler machine learning algos like SVM outperforming them. Averaging the probabilities out was the best way to get the least log loss even if the individual models didn't give the best log loss by themselves.

In the future, we plan on trying to apply various other algorithms such as Recurrent Fuzzy Neural Network, Genetic Algorithm, Cascading Neural Network, and other such algorithms. Trying to deploy this in real life is a priority as well.

## REFERENCES


[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] K. Polaraju, D. Durga Prasad, "Prediction of Heart Disease using Multiple Linear Regression Model", International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017. )

[3] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36.4 (2009): 7675-7680.

[4] Anbarasi, M., E. Anupriya, and N. C. S. N. Iyengar. "Enhanced prediction of heart disease with feature subset selection using genetic algorithm." *International Journal of Engineering Science and Technology* 2.10 (2010): 5370-5376.

[5] Zhang, Yan, et al. "Studies on application of Support Vector Machine in diagnose of coronary heart disease." *2012 Sixth International Conference on Electromagnetic Field Problems and Applications*. IEEE, 2012.

[6] Ajam, Noura. "Heart Diseases Diagnoses using Artificial Neural Network." *Network and Complex Systems* 5.4 (2015): 7-10.

[7] Elshazly, Hanaa Ismail, Abeer Mohamed Elkorany, and Aboul Ella Hassanien. "Lymph diseases diagnosis approach based on support vector machines with different kernel functions." *2014 9th International Conference on Computer Engineering & Systems (ICCES)*. IEEE, 2014

[8] Dey, A., Singh, J. and Singh, N., 2016. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis.Analysis,140(2), pp. 27-31

[9] Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data?.PloS one,12(4), p.e0174944

[10] Mr.P.Sai Chandrasekhar Reddy, Mr.Puneet Palagi, S.Jaya, "Heart Disease Prediction using ANN Algorithm in Data Mining", International Journal of Computer Science and Mobile Computing, April 2017, pp. 168-172

[11] https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

[12] https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[13] David Opitz, Richard Maclin, "Popular Ensemble Methods: An Empirical Study", Journal of Artificial Intelligence Research 11 (1999), pp. 169-198

[14] Grove, A., & Schuurmans, D. (1998). Boosting in the limit: Maximizing the margin of learned ensembles. In Proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 692–699 Madison, WI

[15] Hinton, G.E., Krizhevsky, A., Srivastava, N., Sutskever, I., & Salakhutdinov, R. (2014). "Dropout: a simple way to prevent neural networks from overfitting" Journal of Machine Learning Research, 15, 1929-1958

[16] S. Palaniappan, Rafiah Awang, 2008, "Intelligent Heart Disease Prediction using Data Mining Techniques", JCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008

[17] https://machinelearningmastery.com/ensemble-machine-learning-algorithms-python-scikit-learn/

[18] David Haussler, Michael Kearns, and Robert E. Schapire. *Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension*. Machine Learning, 14:83–113, 1994

[19] Ho, T., Random Decision Forests, *Proceedings of the Third International Conference on Document Analysis and Recognition*, pp. 278-282, 1995

[20] https://www.drivendata.org/competitions/54/machine-learning-with-a-heart/page/107/

[21] https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

[22] *scikit-learn.org/stable/modules/ensemble.html*