Heart Disease Prediction Using Machine Learning Techniques

P. Mohana Vamsi,
3rd year, CSE,
NIT Patna

Shardul Thakare,
2nd year, CSE,
NIIT University, Neemrana

Ashish Mishra,
3rd year, IT,
KIET Group of Institutions,
Ghaziabad

Suryansh Raj,
2nd year, CSE,
NIT Silchar

B. Sangeetha,
2nd year, CSE,
Sri Ramakrishna
Engg. College

Dr. Tanveer Ahmed,
Assistant Professor,
Bennett University

Abstract*—Heart diseases are the number one cause of deaths of people around the world. They refer to the disorders related to the heart and blood vessels. The health sector maintains an enormous quantity of patient-related information. This stored information may be helpful for future disease prediction. In this research paper, we attempt to concentrate on different algorithms for machine learning that effectively predict heart diseases. The complex task of finding heart diseases requires lots of experience and knowledge. Some of the many ways of predicting it are by using 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 primary objective of this research 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, Machine Learning, Heart Disease, KNN, LightGBM, SVM, XGBoost

I. INTRODUCTION

The heart is one of the most, if not 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]. Approximately 17.9 million individuals die each year from heart disease.

People with cardiac illness or those at elevated cardiac danger need early detection and management using counseling and medications, which is why prediction is so essential in these illnesses. 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.

Selecting the relevant features and working on them is very important to find good prediction models. The common specific medical attributes used include chest pain type, resting blood pressure, cholesterol levels, electrocardiograph results(resting), number of major vessels coloured by fluoroscopy, and so on. Other than these, common attributes such as age or sex are also utilized. 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 dataset used, the training data, was made up of 3000 cases. It had 13 characteristics, which were used as the grounds for model training. Training information accounted for 70% of the entire dataset, while the remainder, i.e., 30%, were sent to the test information. Results analysis leads us to think that, while predicting cardiovascular diseases, regression is better than most other models. Das et. al [3]employed the services of an ensemble method consisting of neural networks for diagnosing of the heart disease. It merged the posterior probabilities of various predecessor models or the expected values. The classification accuracy obtained by them using this ensemble model was 89.01%. The dataset that they used for training and testing was the Cleveland Heart Database. Anbarasi et. al [4] used almost a 2-step approach where first they used 3 classifiers. These classifiers that they tried are

decision trees, Naïve Bayes and classification by clustering using thirteen attributes. In the next step, they used genetic algorithm to apply feature subselection and acquired nearly comparable precision. Their findings showed that decision tree classifier performed the best at 99.2 percent for binary classification. The Naïve Bayes classifier followed with the result of 96.5% and then the classification cluster with 88.3%. Zhang et. al [5] used SVM (Support Vector Machines) as the model for cardiovascular disease prediction. They employed PCA to find the relevant and important features and then used various kernels such as rbf and linear to work on the aforementioned features. Radial Basis Function (RBF) gave the highest classification accuracy. To discover the appropriate parameter values, the grid search method was used. In binary classification, the classification precision was 88.6364 percent. Noura Ajam [6] used artificial neural networks. Feed-forward Back propagation algorithms were used to train the model based on the specified requirements. On finding the appropriate parameters, classification accuracy reached to 88%. The number of neurons that were employed in the hidden layer were 20. ANN shows good results as it is a suitable approach to use when the amount of data is large, as in large datasets. Elshazly et. al [7] proposed a novel approach for classification, GA-SVM, for diagnosis of lymph diseases. The GA part of it is just there to make the feature pool reduced. Cross validation technique used was k-fold. Different kernel features were used and performance measurements such as precision, sensitivity, area under curve (AUC), F-measurement were assessed for each feature. Final classifier that was found to be giving the most optimal results was linear, with a score of 83.1%. Dey et. al [8] used multiple models including SVM, Naive Bayes and Decision tree. They also experimented with feature selection and worked using it and without using it as well. The method used here was principal component analysis. The dataset is 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 various models such as 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. PCA was used as well and the best model based on all these parameters was found.

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.

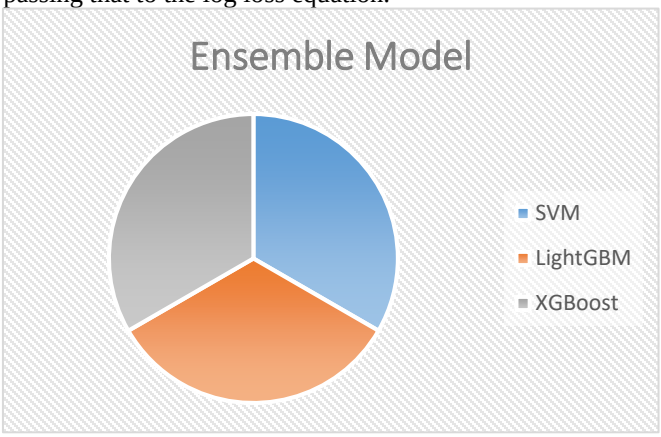

Fig 1. Ensemble model composition

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

SVM: The idea behind a support vector machine is that we have some data points that we want to label either A or B, and we can do that by plotting a line, plane or hyperplane between the two classes A and B, depending upon the problem. What an SVM does is figure out the "best" hyperplane to divide the points, which is the one that leaves the biggest margin between itself and the points. That line is called the maximum margin hyperplane, because it is a hyperplane that separates the sides while leaving the maximum margin.

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 the algorithm of the decision tree that we use to classify. One of LightGBM's unique features is that it divides the tree leaf wise as compared to other boosting algorithms that divide it depth or levelwise. It's based on decision tree.

XGBoost: XGBoost stands for eXtreme Gradient Boosting. XGBoost is an optimized library for boosting distributed gradients. It is intended to be flexible and extremely effective. It uses algorithms for machine learning under the framework for gradient boosting. XGBoost offers a parallel tree boost that quickly and accurately solves issues. 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.

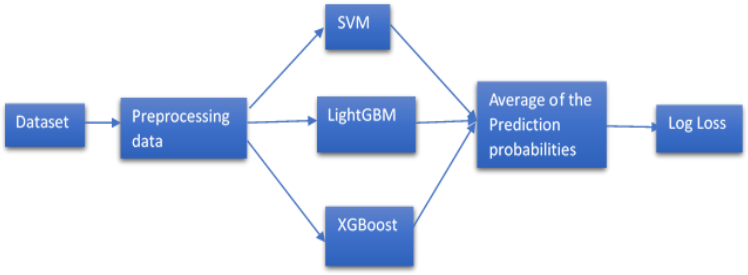

Fig 2. Model architecture

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 metric used for the competition we took part in is logarithmic loss. The formula for log loss is given as:

$$Log\ loss\ = \sum_{i=1}^{n} \quad [y_i\, log(y`_i)+(1-y_i)log(1-y`_i)] \qquad (2)$$

$y`$ is the probability that $y=1$.

Our goal was to reduce log loss. Log loss provides a penalty for too deviant of predictions. 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.

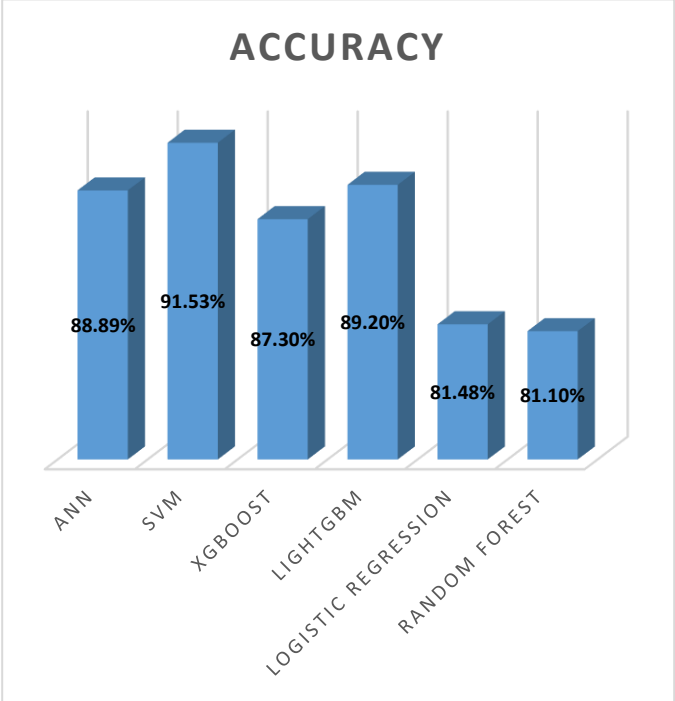

Fig 3. Accuracy results

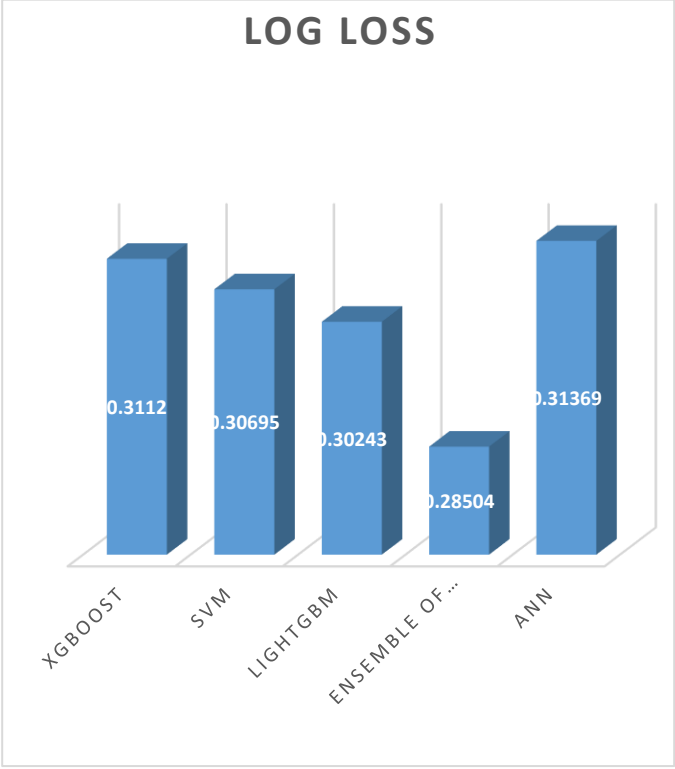

Fig 4. Log loss results

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