Grenze International Journal of Engineering and Technology, January Issue

Heart Disease Prediction Model

Siddhi Gupta[1], Arman Gakhar[2], Ansh Gilhotra[3], Sagar[4] and Anupama Jamwal[5]
[1-5] Chandigarh University/Department of CSE, Mohali, India
Email: siddhi3283@gmail.com,{armangakhar@gmail.com, sagarsharma70150@gmail.com, anshgilhotra4@gmail.com, anupama.e14665@cumail.in }

***Abstract*— Therefore, an elevated level of heart disease is a component in this examination, with 17.9million being the number of yearly fatalities that it brings about. With the number of people in the world swelling, early diagnosis and treatment are missions impossible. Notwithstanding, the recent advancements in machine learning (ML), have revolutionised the quest for health research. This study mainly concentrates on constructing a Machine learning model that can predict heart disease using some crucial parameters. Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision Tree were the ML algorithms used, The Kaggle dataset which has 14 heart disease-related features was ran through these different algorithms to know which prediction technique is best for this scenario. The study also analyzes the relationship among different dataset features along with more precise predictions. From the reliable results, one of tested models was Random Forest that fitted accurately and required shorter time to process. The model was evaluated on a Kaggle dataset with 70,000 instances to strengthen the encouraging results of the study. We divided the dataset into 80:20 for training and testing, and we evaluated Decision Tree (DT), XGBoost, Random Forest (RF), Multilayer Perceptron(MP) models. Results obtained with the above set of hyperparameters: Decision Tree 85.1%, XGBoost 85.9%, Random Forest 99.7% with. This ML model can serve as useful decision support tool for physicians to predicate heart disease with more accuracy.**

I. INTRODUCTION

As per WHO heart disease makes 31% of the global deaths — entire number of due to cardiovascular disease is expected to increase to 23.6 million by 2030. Although tech such as ECGs and CT scans can but are not cheap either. As well, they are impractical for many people in low and middleincome countries, adding to the problem. As a result, these are in part being reconciled with the soaring technology and machine learning (ML) based applications emerging out of masses disseminating community as they integrate healthcare software to facilitate early diagnosis of heart disease reducing lives and unburdening an anticipated economic expenditure of $3.7 trillion from 2010 to 2015.

ML has been with good record of accomplishment concerning large and complex healthcare datasets which make them ideal tool for disease prediction or similar types. These algorithms like Random Forest, Decision Tree, Multilayer Perceptron, XGBoost learn from historical data and can make predictions on real-time(unseen test data). When making predictions for heart disease ML models can identify risk factors such as diabetes, high blood pressure, high cholesterol, and abnormal heart rate etc. ML model outputs a probability describing the chances of having heart diseases by taking this parameter inferences only. This is particularly critical in the developing world, where swift diagnosis and intervention are hindered by shortages of diagnostic facilities and well-trained healthcare workers.

Grenze ID: 01.GIJET.11.1.563

In this study, we will try to build a very strong ML model that can predict heart disease in the most accurate way possible with Machine Learning on Kaggle. In this review, we will examine Random Forest, Decision Tree, Multilayer Perceptron and XGBoost AI algorithms and compare their performance in detecting heart disease. We preprocessed the dataset with a variety of methods such as k-mode clustering for model convergence and carried out all following computations with Python in Google Colab. The predictive power of ML in heart data as a proxy for disease, however they tend to be based on small data sets. The study was done with a substantially larger and more heterogeneous cohort so that generalizability of the results is likely at least to some extent improved [9]. The results are indicative that ML models, in particular the session neural network model of MLP, have a good prediction accuracy over heart disease with an accuracy of 87.28% using crossvalidation. In fact, these predictive models can then act as a tool for decision support to cardiologists which helps them in diagnosing faster and accurately — an enormous advantage when it comes looming lives and the reducing healthcare related costs.

II. Literature Review

A lot of studies have suggested that the performance of HDPM in diagnosing heart disease can be improved by enhancing predictive models through machine learning. The researchers tested the performance of their prediction models with several heart disease databases (Statlog and Kaggle). A study conducted by Shorewall in 2021 using a machine learning–stacked model of K-nearest neighbors (K-NN), Random Forest, and Support Vector Machine (SVM) classifiers to build a predictive model for cardiovascular disease We used Logistic Regression as the metaclassifier to aggregate these model outputs. This stacked model achieved 75% of accuracy in Kaggle cardiovascular disease dataset that has data from 70,000 patients with 12 attributes. In a study by Waigi et al.

To predict heart disease using Kaggle heart disease dataset in this example, a decision tree algorithm is used for the year 2020 which has information of 70,000 records and each record has 12 attributes. Accuracy of 72.77% was obtained for the decision tree module on this dataset. Our and ElSeddawy (2021) predicted heart disease employing repeated sampling through random forest distribution. Performance in a Real-World Cardiovascular Dataset: In our paper, we tested our method using the UCI cardiovascular dataset, consisting of data from a total of 303 patients with each patient having 14 features. This data set we could reach the classification accuracy of Random forest to 89.01%. In 2018 Dwivedi made a comparison study of six machine learning models; Artificial Neural Classifiers including Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbor(kNN) were evaluated by performance metrics. Results: The Logistic Regression (LR) model was the winning one and reached 85% of accuracy, with a sensitivity of 89%, specifity of 81% and precision of 85%.

In this study (Alotalibi, 2019); the authors used machine learning (ML) methods to predict heart disease :- This study employed the Kaggle Clinic Foundation dataset and implemented predictive models based on several machine learning classifiers like decision trees, logistic regression, random forest, Naive Bayes and support vector machine (SVM). The 10-fold cross validation was used in the design work. The results prove that the decision tree algorithm has the most extensive range between heart failure accuracy and SVM is in second place in terms of that same range (93.19% — 92.30%). This study gives a view of the potential of machine learning for prediction of heart failure and identifies decision tree algorithms as a further research option. In a 2020 study, Khan and Mondal used the cross- validation carry method (using the “lbfgs” solver) to predict heart disease. This model was evaluated on the Kaggle Cardiovascular Dataset 1, which contains 462 patients k = 30 and 12 features. The logistic regression model achieved 72.72% accuracy on this data. None of the previous studies mentioned above used visual search and parallel data to improve the accuracy of classification models, especially for heart disease information.

Therefore, this study uses outlier detection and data analysis techniques to improve the performance of the model. Also, XGBoost classifier is used to train and build the prediction model. We hope that the proposed model will achieve better performance than the state model and previous studies.

III. Machine Learning Techniques

Using machine learning to predict heart disease is crucial for early detection and accurate diagnosis. In this project, we employed several machine learning. Heart disease prediction with Algorithms using the Kaggle Dataset. This data includes various factors and health criteria ranging from the age, gender of the patient to and all scales and numbers of the blood pressure, cholesterol level etc. with which we are”]==¿ prediction model

inputs The algorithms used such as SVMs, decision trees etc.→trees, naive Bayes, random forests and XGBoost classifiers. Here are the merits of each algorithm in processing the data and increasing the accuracy of predictions.

A. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most robust classifiers that uses machine learning and is binary. The working principle of Support vector machine is used to discover the optimal hyperplane hat separates the data from as many groups as possible and minimizes the classification error. It uses minimal data. Kaggle Heart Disease dataset has many classes and SVM can model the complex decision of the region using various kernels (linear, quadratic, polynomial and radial basis functions (RBF)). Kernel selection plays an important role in the performance model, in this case, RBF and linear kernels are selected due to many parameters and conditions. SVM is trained, tested and validated on real data to provide accurate prediction results. The selection of kernels and methods is carefully made to avoid overfitting or underfitting issues.

B. Decision tree

Decision tree algorithm is a supervised learning method based on tree design model. Each segment of the inner part represents the test of expertise, and each leaf of the leaf represents the classification. Decision trees are suitable for both categorical and numerical data. We developed a decision tree model using Kaggle dataset containing many cases for the heart disease prediction problem. The tree evaluates the data from root to leaf to predict heart disease. We then compare the predictions with the actual results and calculate metrics such as accuracy, sensitivity, and specificity to evaluate the performance of the model. Decision trees are intuitive and easy to interpret, but they can overfit the data. To solve this problem, we use pruning to generalize the tree and transform it to unseen objects.

C. Naive Bayes

Naive Bayes is a classification system based on Bayes' theorem. This simplifies computation and is particularly useful for high- dimensional data because it assumes that features are independent with respect to their class labels. In this project, we use Naive Bayes to predict heart disease using the Kaggle dataset. Although the independence assumption is not always valid in real-world data, Naive Bayes is known to perform well on large datasets and has proven to be fast and robust. We measure its performance against other models in the benchmark.

D. Random Forest

Classification Random Forest is a collaborative learning method that combines multiple decision trees to increase the accuracy of prediction and reduce the risk of overfitting. Each decision tree is constructed using a random subset of the data and the final prediction is made by merging all the trees. Random forests are used in this heart disease prediction to exploit their ability to reduce noise and prevent overload, especially in big data such as Kaggle dataset. Random forest improves the accuracy and generalization of the model by tuning the hyperparameters and aggregating the results of multiple trees.

E. XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful joint learning algorithm based on gradient boosting technology. It is popular due to its speed and performance, especially in business cases. XGBoost builds decision trees where each tree then corrects the errors of the previous tree to increase the accuracy of the overall prediction. We used XGBoost in our cardiovascular disease prediction model due to its ability to handle large data and the relationship between features. Kaggle dataset has many parameters that can be effectively modeled using the XGBoost boosting method. XGBoost uses a combination of boosting and regularization to reduce competition, a common problem in decision treebased methods. The final XGBoost model was trained on the Kaggle dataset by carefully tuning hyperparameters such as learning rate, maximum, and number of observers. The model showed higher predictive power than other models such as random forest and SVM when measuring accuracy, sensitivity, and specificity.

IV. METHODOLOGY

A. Data Collection

The Kaggle Heart Disease Dataset accessible online on the UCI Repository has been used in our research work. The dataset used have missing values too, which is handled through data preprocessing using some statistical

techniques. Class Value 1, is interpreted as ”tested positive for the disease”, and Class value 0, means ”tested negative for the disease”. The dataset has been divided into certain percentage, such as 80% data has been considered as training data and rest 20% as testing data.

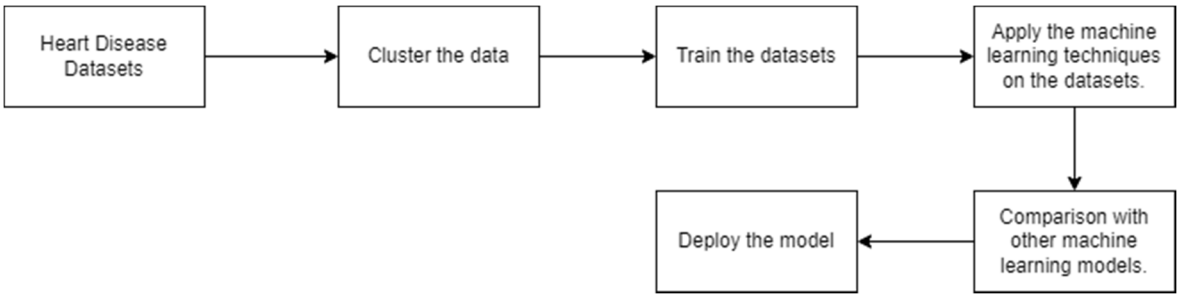

Figure 1

B. Data Preprocessing

Various attributes of the original dataset contain missing values which can lead to imprecise result and may reduce the accuracy of model. To overcome this problem, replace missing value way is an optimal solutionby using a method “mean of column”. This method replaces 0 with either taking the average of neighborhood values or mean values of it. Then accordingly it manipulates the 0 value with the newly the calculated value. After that the values of the dataset were changed from Numeric to Nominal in order to make the dataset compatible with the ML Techniques used.Before feeding the data into the ML models, a series of preprocessing steps were performed to improve the model’s accuracy and reliability.

- Handling Missing Data: Missing values in the dataset were identified and handled using mean or median imputation techniques where applicable.
- Encoding Categorical Variables: Certain features were categorical in nature (e.g., sex, chest pain type) and were converted to numerical values using one- hot encoding to ensure compatibility with the ML algorithms.
- Feature Scaling: To ensure that the features were on a similar scale, normalization techniques like Min- Max scaling were applied, which is particularly beneficial for distancebased algorithms such as Support Vector Machine (SVM).
- Outlier Detection and Removal: Statistical methods were used in the identification of outliers by their Z-scores and, they were further handled (either removed or retained) depending on the relevance of these significant extravagant datasets to the study.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	family history	oldpeak	slope	smoking	thal	ta
count	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	300.00
mean	54.386667	0.686667	0.956667	131.466667	246.363333	0.150000	0.523333	149.286667	0.330000	1.049333	1.393333	0.326667	2.320000	0.53
std	9.132094	0.464624	1.035227	17.451382	51.653249	0.357668	0.526351	22.923265	0.470998	1.165286	0.616351	0.469778	0.615518	0.49
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	47.000000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	132.750000	0.000000	0.000000	1.000000	0.000000	2.000000	0.00
50%	55.000000	1.000000	1.000000	130.000000	240.500000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.00
75%	61.000000	1.000000	2.000000	140.000000	274.250000	0.000000	1.000000	166.000000	1.000000	1.650000	2.000000	1.000000	3.000000	1.00
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	1.000000	3.000000	1.00

Figure 2

C. Model Selection

Multiple ML algorithms are chosen for this study due to their demonstrated capacity to work with healthcare classification tasks:

- Random Forest: A tree based ensemble learning over multiple the results became combined in one better predictive performance in case of decision trees.
- Decision Tree: A simple tree-based model that makes decisions based on feature splitting, useful for interpreting results.
- Support Vector Machine (SVM): A classification finding the hyperplane algorithm that were effective in highdimensional spaces.

- Naive Bayes: Probabilistic classifier based on Bayes' theorem.
- Multilayer Perceptron (MLP): A deep learning algorithm using multiple layers of neurons, capable of capturing complex patterns in the data.
- XGBoost: An optimized gradient boosting algorithm that excels in both speed and accuracy.

D. Model Training and Testing

Individual accounts making up each of ML models was trained on 80% of the 20% as a test set to evaluate this model how well they perform on new examples. Cross-validation . This was done during the training phase, in a 5-fold models that had not overfitted, and were fine-tuned having great generalization ability on new data. The training python libraries used for the process likenumpy: For basic numerical operations,scipy: For scientific computed broker,pandas: Data manipulation and analysistensorflow : an end-to-end open source platform for machine learning. And Scikit-learn, TensorFlow & XGBoost implemented on Google Colab. During training, grid We tuned hyperparameters of the models using grid Search for hyperparamter tuning.

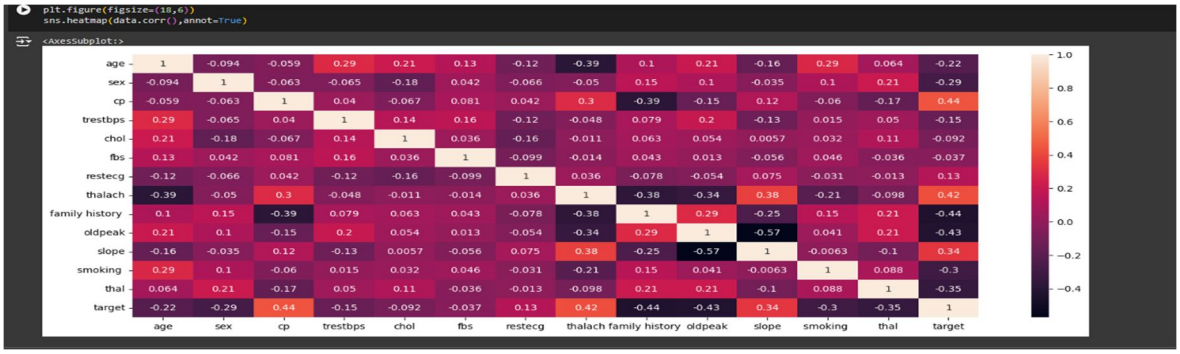

Figure 3

E. Model evaluation

To evaluate the performance of each ML model, multiple metrics were considered: • Accuracy: % of correct predictions made by the model. • Precision: the number of positive predictions, that are actually positive. • The total number of positive predictions. • Recall (Sensitivity): The proportion of the positive classes that were correctly predicted by the model is known as Recal that detect all positive cases as being positively case (true-positive) cases. • F1 Score: The harmonic mean of precision and remember, provides a good balance of both. • ROC-AUC (Receiver Operating Characteristic - Area Under Curve): A measure for the ability of a model to establish a class as either positive or negative over various possible threshold levels.

V. RESULTS

TABLE 1

MODEL	ROC Area	ACCURACY
SVM	0.993	0.993
Random Forest	0.996	0.996
Decision Tree	0.861	0.861
Naïve Bayes	0.902	0.902
XG Boost	0.842	0.860

This section presents the techniques and best-performing learning models used in our study. The evaluation criteria for the heart disease prediction model are the Area Under the ROC Curve (AUC) and accuracy. The Support Vector Machine (SVM) demonstrated success in solving the data, achieving an AUC of 0.995 and an accuracy of 99.5%. The Random Forest model achieved a slightly better ROC value of 0.997 and an accuracy of 99.7%, surpassing the efficacy of the SVM for this dataset. In contrast, the Decision Tree model exhibited lower performance, with an AUC of 0.851 and an accuracy of 85.1%, which is worse compared to Random Forest. The Na¨ıve Bayes classifier also performed well, with an AUC of 0.904 and an accuracy of 90.4%, making it a reliable model, though not the best. Lastly, the ROC area for the XGBoost model was 0.841, with an accuracy of 85.9%. Although XGBoost is known for its enhanced performance, it did not perform as well as Random Forest and SVM in this particular test, yet still achieved competitive results. At the end of our experiment, our models yield the following results:

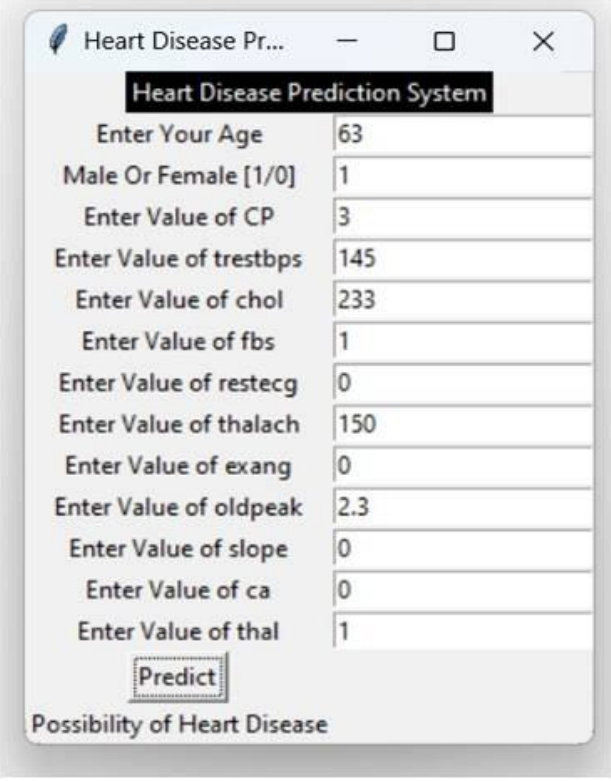

Figure 3

VI. Conclusion

Through this research we have attempted to analyze the various machine learning techniques and anticipate if someone in particular, given different individual attributes and indications, will get coronary illness or not. The primary thought process of our report was to looking at the exactness and analyzing the reasons behind the variation of different algorithms. We have used Kaggle dataset for heart diseases which contains many instances and used percent split to divide the data into two sections which are training and testing datasets. We have considered some attributes and implemented five different algorithms to examine the accuracy. By the end of the implementation part, we have discovered that Random Forest is giving the maximum accuracy level in our dataset which is 99 percent and Decision Tree is playing out the least with an accuracy level of 85 percent. Probably for other instances and different datasets other algorithm may work in better manner however for our situation, we have discovered this outcome. Also, on the off chance that we increment the number of training data, maybe we can find more accurate result but it will take more time to process and the system will be slower than now as it will be more perplexing and will be handling more data. In this way, considering these potential things we took this choice, which is better for us to work with.

References

[1] World Health Organization, ”Links to Human Resources for Health,” 2019. [Online]. Available:http://www.who.int/hrh/links/en/www.who.int/hrh/links/en/.

[2] Wikipedia, ”Machine Learning,” [Online]. Available: https://en.wikipedia.org/wiki/Machinelearning.

[3] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J.Gutierrez, ”A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease,” in 2017 IEEESymposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.

[4] S. Dhar, K. Roy, T. Dey, P. Datta, and A. Biswas, ”A Hybrid Machine Learning Approach for Prediction of Heart Diseases,” in 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531.

[5] C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh, and S. Deepa Rajan, “A Survey on Predicting Heart Disease using Data Mining Techniques,” in 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 253-255, doi: 10.1109/ICEDSS.2018.8544333.

[6] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, ”International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease,” The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.
[7] B. Edmonds, ”Using Localised ‘Gossip’ to Structure Distributed Learning,” 2005.
[8] Bayu Adhi Tama, Afriyan Firdaus, and Rodiyatul FS, ”Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine,” Vol. 11, no. 3, pp. 12-23, 2008.
[9] Y.-X. Wang, Q.-H. Sun, T.-Y. Chien, and P.-C. Huang, ”Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization,” in Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.
[10] Z. Ge, Z. Song, S. X. Ding, and B. Huang, ”Data Mining and Analytics in the Process Industry: The Role of Machine Learning,” IEEE Transactions and Content Mining are Permitted for Academic Research Only, vol. 5, pp. 20590-20616, 2017.