

Prediction of Ischemic Heart Disease in Patients Diagnosed with Coronavirus Disease (Covid-19) or Showing Symptoms of the Same using Machine Learning Algorithms and Ensemble Techniques

Aalapi Bhave¹ and Siddharth Gaikwad²

¹⁻²College of Engineering, Pune, Department of Computer Engineering, Pune, India
Email: bhaveac20.comp@coep.ac.in, skg.comp@coep.ac.in

Abstract—Ischemic Heart Disease (IHD) refers to the problems of the heart occurring due to narrowed heart arteries. Ischemic Heart Disease is also called as Coronary Heart Disease (CHD) or Coronary Artery Disease (CAD). Coronary artery is responsible for supplying blood to the heart muscle. IHD is caused by blockages in the coronary artery due to fat deposits or thrombus. Early diagnosis of IHD is critical. This research involves prediction of IHD in patients diagnosed with Coronavirus disease (Covid-19) or showing symptoms of the same. The dataset used for this research has been collected from nine hospitals across the Indian state of Maharashtra. It is a dataset of patients diagnosed with Coronavirus disease (Covid-19) or showing symptoms of the same. Machine Learning algorithms and ensemble methods have been used for the prediction of IHD. The algorithms used are Decision Tree, Random Forest, AdaBoost, CatBoost, LightGBM and XGBoost. The ensemble methods used are the Majority Voting Ensemble and Stacking. Maximum accuracy of 90.14% is obtained by using Majority Voting Ensemble which is an ensemble of Random Forest, XGBoost and AdaBoost algorithms using voting type as soft voting. Google colab has been used for the implementation of the project and the programming language used is Python.

Index Terms— AdaBoost, CatBoost, Coronavirus disease (COVID-19), Decision Tree, Ischemic heart disease, LightGBM, Majority Voting Classifier, Random Forest, XGBoost.

I. INTRODUCTION

Ischemic Heart Disease refers to the problems of the heart occurring due to narrowed heart arteries. Ischemic Heart Disease is also called as coronary heart disease (CHD) or coronary artery disease (CAD). It is caused by blockages in coronary arteries which are responsible for the blood supply to the heart muscle due to fat deposits or thrombus. The dataset used for this research has been collected from nine hospitals across the Indian state of Maharashtra. It is a dataset of patients diagnosed with Coronavirus disease (Covid-19) or showing symptoms of the same. Coronary heart disease is a major cause of death worldwide [1]. Early diagnosis of coronary heart disease is critical so that damage to the heart muscle can be minimized.

Machine Learning algorithms can be used to predict the possibility of Coronary Heart Disease. Use of Machine learning automates the process of prediction and decision making. Machine Learning algorithms process large

datasets, learn from them, detect patterns and the model can be employed for accurate and efficient prediction. There are large quantities of patient data in the medical field and Machine Learning can be used effectively in the healthcare industry. It can be used for the purpose of classification on a labelled dataset, association rule mining, for clustering or grouping of similar cases, for anomaly detection in case of finding out unusual patterns and to extract relevant information or features.

For the purpose of this research, the dataset used has 3,500 records of patients either diagnosed with Coronavirus disease (COVID-19) or showing symptoms of the same [2]. A total of 18 attributes are considered which will be useful in the prediction of Ischemic Heart Disease in these patients. It is a labelled dataset. In this research, the aim is to determine the best approach based on the accuracy values returned by the machine learning algorithms and ensemble methods which have been implemented.

II. RELATED WORK

Several models have been implemented by ML researchers for the prediction of heart diseases on various datasets.

Apurva Garg et al have worked on the UCI Cleveland dataset [3]. This is a dataset available in the UCI Machine Learning repository. This paper uses K-Nearest Neighbour and Random Forest Algorithms. The K-Nearest Neighbour algorithm was the algorithm which returned the highest accuracy of 86.88%. The value of k was 12 when this accuracy was obtained. Random Forest gave an accuracy of 81.96%. The UCI Cleveland dataset used is a small dataset which has 303 instances and 14 attributes. By comparison, the dataset used in the proposed research is larger with 3500 samples and so the model can be more reliable. Additionally, the dataset used in the proposed research is of patients suffering from or showing symptoms of Coronavirus disease (Covid-19).

W.M Jinjri et al have worked on the cardiovascular disease dataset available on Kaggle [4]. Decision Tree, Naïve Bayes, Logistic Regression, K-NN and SVM algorithms have been used in this paper. SVM gave a value of accuracy of 72.6% which was highest value of accuracy obtained.

Apurb Rajdhan et al have worked on the UCI Cleveland dataset [5]. This paper presents a comparative study by comparing the performances of Decision Tree, Random Forest, Logistic Regression and Naïve Bayes algorithms. Maximum accuracy of 90.16% was obtained using Random Forest.

Yu Lin has also worked on the UCI Cleveland dataset of heart disease patients [6]. The algorithms used are Logistic Regression, K-NN, AdaBoost, CART, Random Forest and XGBoost. The algorithm giving the highest value of accuracy was Random Forest with a value of 84.8%.

J. Zhou has used the Cardiovascular Disease Dataset of Kaggle [7]. Stacking Technology has been implemented. A fusion model has been created that uses BP Neural network, SVM and Logistic Regression by using Stacking Technology. The stacked model gave a better accuracy than the single algorithms when run individually. The fusion model gave an accuracy of 72.4%. The paper proposes the combination of blending technology and XGBoost as part of future work for better results.

Vardhan Shorewala has worked on the cardiovascular disease dataset on Kaggle [8]. Logistic Regression, SVM, Decision Tree, K-NN, Naïve Bayes and Neural Network were used as the base models. Then techniques of Bagging, Boosting and Stacking were applied and the performance was compared with that of the base models. The base models had a capped accuracy at 75% and an average accuracy of 71.92% while the Neural Network gave an accuracy of 73.97%. Bagging technique showed an average accuracy change of +1.9% whereas Stacking was the most effective giving an accuracy of 75.1%.

Sumit Bhatia et al used SVM and Inter Coded Genetic algorithm for Feature selection [9]. The UCI Cleveland dataset has been used for this study. The accuracy obtained for the five-class classification problem was 72.55% whereas for a two-class problem the accuracy obtained was 90.57%.

III. DATASET DESCRIPTION AND ANALYSIS OF DATA

A. Dataset Attributes or Features

This is a dataset of 3500 patients' records with 18 features considered. 17 features or attributes are used as input while IHD is the target attribute (output). Table I describes the various features or attributes of the dataset. The last column also lists the value of the datatype after pre-processing has been applied to the dataset.

As part of the process of analysing the data, correlation between each feature of the dataset and the target feature IHD is determined. Diabetes Mellitus, Hypertension and Chronic Kidney disease were found to have maximum positive correlation with the target variable IHD. Table II depicts the values.

TABLE I. FEATURES OF THE DATASET

Serial No	Feature Name	Feature Description	Datatype	Datatype value After Pre-processing
1	Age	Age	object	int
2	Sex	Gender	object	Categorical Code
3	Fever	Presence or Absence of Fever	object	Categorical Code
4	Cough	Presence or Absence of Cough	object	Categorical Code
5	Breathlessness	Presence or Absence of Symptoms of Breathlessness	object	Categorical Code
6	DM	Presence or Absence of Diabetes Mellitus	object	Categorical Code
7	HTN	Presence or Absence of Hypertension	object	Categorical Code
8	CKD	Presence or Absence of Chronic Kidney Disease	object	Categorical Code
9	Respiratory Disorders	Presence or Absence of Respiratory system related disorders	object	Categorical Code
10	Hypothyroidism	Presence or Absence of Hypothyroidism	object	Categorical Code
11	Hb	Haemoglobin level	object	float
12	D Dimer	D Dimer level	object	float
13	Creatinine	Creatinine level	object	float
14	Ferritin	Ferritin level	object	float
15	Steroid	Intake of Steroid by the patient	object	Categorical Code
16	LMWH Heparin	Use of Low Molecular Weight Heparin and dose	object	Categorical Code
17	Covid19 Positive Negative	Covid-19 test result	object	Categorical Code
18	IHD	Presence or absence of ischemic heart disease (Target variable)	object	Categorical Code

TABLE II. MAXIMUM POSITIVE CORRELATION VALUES

Attribute Name	Correlation with Target variable
DM	0.164374
HTN	0.305016
CKD	0.084774

B. Data Pre-Processing and Analysis

The dataset pre-processing involved handling of missing values. The dataset contains no null values. For the numeric columns, the missing values are replaced with the zero. For the target variable, the rows with missing values have been removed from the dataset. For categorical columns, label encoding has been performed as part of pre-processing. Also, in the original dataset, all the attributes were of datatype object. During the pre-processing, they have been converted to int, float or categorical code as mentioned in Table I.

The phase of data analysis also involves calculating the entropy and Information gain (IG). The dataset entropy and the IG for each feature of the dataset with respect to the target feature IHD is computed. Table III depicts the computed values. IG values for all the features lie within the boundary of 0 to 1.

ENTROPY OF THE DATASET: 0.5860540781380684

IV. MACHINE LEARNING TECHNIQUES AND IMPLEMENTATION METHODOLOGY

A. Machine Learning Techniques

The following machine learning algorithms have been used for the prediction of ischemic heart disease:

- Decision Tree
- Random Forest
- CatBoost
- AdaBoost
- LightGBM
- XGBoost

Criteria used is entropy in case of Decision Tree and Random Forest algorithms. Hyperparameter tuning is performed using optuna framework.

The following ensemble methods have been used:

- Majority voting ensemble using hard voting
- Majority voting ensemble using soft voting
- Stacking

TABLE III. IG VALUES

Feature Name	IG w.r.t target feature (IHD)
Age	0.052691403835504524
Sex	0.0010756127779946167
Fever	5.252415637169161e-05
Cough	0.000933169042024562
Breathlessness	0.0037270973861173173
DM	0.014824054354283134
HTN	0.06407880731179594
CKD	0.013977574336736964
Respiratory Disorders	0.008286341525661878
Hypothyroidism	0.009832425412663492
Hb	0.03670815772373048
D Dimer	0.235151840971485
Creatinine	0.11475089702894065
Ferritin	0.34273926845464076
Steroid	0.017782864865361736
LMWH Heparin	0.00024252181365136405
Covid19 Positive Negative	0.0010971972452297196

The majority voting ensemble used in this research consists of three algorithms namely Random Forest, XGBoost and AdaBoost.

The results returned by the machine learning algorithms and ensemble methods are analysed and compared.

B. Implementation Methodology

As explained before, Pre-processing of the dataset and analysis of data is carried out. Further, the dataset entropy and IG values are computed as seen in Table III. Splitting of the dataset is performed with size of test data as 20% while size of training data is 80% and Machine learning algorithms are applied. Hyperparameter tuning is performed using optuna framework. Majority Voting ensemble consisting of Random Forest, XGBoost and AdaBoost is then applied with voting type as hard and soft. Stacking has also been implemented. The results of the Machine Learning algorithms and the ensembles are analyzed and compared.

V. RESULTS AND ANALYSIS

This research aims to determine the approach which gives the highest accuracy for the different Machine Learning algorithms and ensemble methods implemented on the dataset. The size of test data is 20% and the size of training data is 80%. sklearn library, LightGBM, XGBoost and catboost python packages were used during the implementation. The code has been written in python and executed in Google colab.

The Accuracy is a measure of the correctly classified samples over the total number of samples and is calculated as:

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative}) / (\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{True Negative})$$

The values for accuracy given by the different machine learning algorithms implemented and the ensemble methods are as shown in Table IV, Table V and Table VI.

TABLE IV. ACCURACY VALUES FOR DIFFERENT ALGORITHMS

Algorithm	Accuracy (%)
Decision Tree	88.50
Random Forest	89.65
AdaBoost	88.66
LightGBM	88.17
XGBoost	89.98
CatBoost	88.99

TABLE V. ACCURACY OF MAJORITY VOTING ENSEMBLE

Type of Ensemble	Algorithms used in ensemble	Type of voting	Accuracy (%)
Majority Voting Ensemble	Random Forest, XGBoost, AdaBoost	Soft	90.14
Majority Voting Ensemble	Random Forest, XGBoost, AdaBoost	Hard	89.98

TABLE VI. ACCURACY OF STACKING

Type of Ensemble	Algorithms used in ensemble	Accuracy (%)
Stacking	Random Forest, XGBoost, AdaBoost	89.49

Table IV shows us that out of the individual algorithms, the accuracy given by XGBoost is the highest with a value of 89.98%. The accuracy returned by XGBoost is much greater than the values of accuracies seen in the related work section for the Cardiovascular dataset of Kaggle. The increase in accuracy is higher than 10%. Also, the number of records in the dataset used in this project are much larger than the UCI Cleveland dataset reviewed in the related work section. Therefore, the model is more reliable and robust. Additionally, this dataset is of patients suffering from or showing symptoms of Coronavirus disease (Covid-19).

Table V and Table VI gives us the values for Ensemble methods. Majority voting ensemble of Random Forest, XGBoost and AdaBoost gives an accuracy of 90.14% using soft voting and 89.98% using hard voting. The Majority Voting ensemble with voting type as soft voting improves over the accuracies given by the individual algorithms. With voting type as hard voting, the accuracy value is higher than that of Random Forest and AdaBoost but same as that of XGBoost when run individually. Table VI shows that using Stacking, the accuracy obtained is 89.49%.

Thus, with an accuracy of 90.14%, we can see that a Majority Voting ensemble of Random Forest, XGBoost and AdaBoost using soft voting gives us the highest accuracy on this dataset.

VI. CONCLUSION

The purpose of this research is to determine the machine learning technique which gives us the highest accuracy on this dataset. This is a dataset of patients diagnosed with Coronavirus disease (Covid-19) or showing symptoms of the same collected from nine hospitals across the Indian state of Maharashtra.

As seen in the Results and Analysis section, amongst individual algorithms, the accuracy given by XGBoost is the highest with a value of 89.98%. The accuracy returned by XGBoost is much greater than the values of accuracies seen in the related work section for the Cardiovascular dataset of Kaggle. The increase in accuracy is higher than 10%. The dataset used in this project consists of 3500 samples. Thus, the number of records in the dataset used in this project are much larger than the UCI Cleveland dataset reviewed in the related work section which has only 303 samples. Therefore, the model is more reliable and robust. Additionally, this dataset is of patients suffering from or showing symptoms of Coronavirus disease (Covid-19) and the aim is prediction of Ischemic Heart Disease in these patients.

The Majority voting ensemble of Random Forest, XGBoost and AdaBoost gives an accuracy of 90.14% using soft voting and 89.98% using hard voting. The Majority Voting ensemble with voting type as soft voting improves over the accuracies given by the individual algorithms. With voting type as hard voting, the accuracy value is higher than that of Random Forest and AdaBoost but same as that of XGBoost when run individually. Using Stacking, the accuracy obtained is 89.49%.

Thus, with an accuracy of 90.14%, we can see that a Majority Voting ensemble of Random Forest, XGBoost and AdaBoost using soft voting gives us the highest accuracy on this dataset.

In the future, we plan to implement association rule mining.

REFERENCES

- [1] <https://www.nhs.uk/conditions/coronary-heart-disease/>
- [2] Dr S.B Dixit, M.D. Covid-19 Dataset_Ischemic Heart Disease. (2021).
- [3] Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012046). IOP Publishing.
- [4] W. M. Jinjri, P. Keikhosrokiani and N. L. Abdullah, "Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study," 2021 International Conference on Information Technology(ICIT), 2021, pp. 132-138, doi: 10.1109/ICIT52682.2021.9491677.
- [5] Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020).
- [6] Y. Lin, "Prediction and Analysis of Heart Disease Using Machine Learning," 2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI), 2021, pp. 53-58, doi: 10.1109/RAAI52226.2021.9507928.
- [7] J. Zhou, "Cardiovascular Disease Diagnosis Based on Stacking Technology," 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2021, pp. 706-710, doi: 10.1109/IPEC51340.2021.9421128.

- [8] Vardhan Shorewala, Early detection of coronary heart disease using ensemble techniques, Informatics in Medicine Unlocked, Volume 26, 2021, 100655, ISSN 2352-9148, <https://doi.org/10.1016/j.imu.2021.100655>. (<https://www.sciencedirect.com/science/article/pii/S235291482100143X>)
- [9] Bhatia, S., Prakash, P., & Pillai, G. N. (2008, October). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. In Proceedings of the world congress on engineering and computer science (pp. 34-38).
- [10] <https://lightgbm.readthedocs.io/en/latest>
- [11] <https://xgboost.readthedocs.io/en/stable/index.html>
- [12] Bhawe, A. and Gaikwad, S., " Prediction of Cardiovascular Disease using Machine Learning Algorithms and Ensemble Techniques", International Journal of Computer Science & Communication (ISSN: 0973-7391), Volume 13, Issue 1 pp. 14-23 Sept 2021 - March 2022 www.csjournals.com