ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse227952020.pdf
https://doi.org/10.30534/ijatcse/2020/227952020

Analysis and Prediction of Chronic Heart Diseases Using Machine Learning Classification Models

Abhishek Anand[1], Harshit Anand[2], Siddharth S. Rautaray[3], Manjusha Pandey[4], Mahendra Kumar Gourisaria[5]
[1]School of Computer Engineering, KIIT DU., Bhubaneswar, India, 1706288@kiit.ac.in
[2] School of Computer Engineering, KIIT DU., Bhubaneswar, India, 1705238@kiit.ac.in
[3]School of Computer Engineering, KIIT DU., Bhubaneswar, India, siddharthfcs@kiit.ac.in
[4]School of Computer Engineering, KIIT DU., Bhubaneswar, India, manjushafcs@kiit.ac.in
[5]School of Computer Engineering, KIIT DU., Bhubaneswar, India, mkgourisariafcs@kiit.ac.in

ABSTRACT

The continuous development in science and technology is touching skies. Advancement in technology is making humankind to enjoy all ease. It brings all aspects of a peaceful living down to the finger-tips of a man. But even in this fast-moving world, it is observed that getting detections for chronic diseases and disorders is tedious and time-taking. Further, seeking a doctor is accompanied by certain issues ranging from quality concern, staffing issues, patient safety, and human negligence. Above all, the bills associated with healthcare are sufficient enough to eat up all savings of a common man. Henceforth, to expand the productivity and pace of this process, utilizing the concepts of Machine Learning and Data Science, a modal is proposed for the prediction and classification of the occurrence of chronic heart diseases based on the user entered values. The heart is one of the most vital organs is supposed to be monitored at regular intervals of time, consequently, for this, a classifier is developed which can anticipate if a patient is suffering from any heart disorder or not. The idea in this research work was to capture patterns and find out any hidden traits from different patients to ease the process of prediction, detection, and classification using algorithms. Algorithms of Machine Learning such as SVM, Decision Trees and Random Forest were implemented along with the libraries of Numpy, Pandas, Seaborn and Scikit Learn for the prediction of the output label. Further, to enhance the accuracy Hyperparameter Tuning is done which increased the model efficiency with an accuracy of around 98%.

Key words: Data Science, Django, Heart Disease Prediction, Machine Learning, Python, Pandas, Random Forest.

1. INTRODUCTION

Heart, the powerhouse of our body, being one of the most vital and requisite organs is responsible for our existence. It not only pumps up blood with a rhythm but also manages the removal of wastes concerned with various metamorphic processes of our body. As the quote “With great powers, comes great responsibilities” exclaims, the fundamental and biological architecture of the heart is much complex in order to support its functionalities. The fact that the wall of the heart is made up of three tissue layers-Epicardium, Myocardium, and Endocardium and it is contained in a double-walled sac-like structure-Pericardium, is in itself making evident how well organized and complex the heart is.

With greater complexity, the heart comes with a greater chance of diseases and disorders. Starting from 2008, cardiovascular disorders and diseases are one of the most broadly perceived explanations behind death, representing 30% of the total. According to the CDC, every 37 seconds an individual dies from cardiovascular diseases in the United States. A Survey says someone encounters a heart attack every 40 seconds in America. Heart diseases refer to improper functioning, weakening, and getting damaged of the heart which might occur as an outcome of bad lifestyle, ageing or inherent qualities. Heart diseases include the unhealthy development of fat deposits on the inner sides of the arteries, hypertension, uneven flow of blood in the vessels, increasing stiffness, or hardening of the arteries et cetera.

Identification of cardiovascular diseases have always been a tough job. Doctors and medical practitioners are asked to search every bit of their degrees to get to a conclusion about a patient. It takes a series of tremendous and rigorous processes to conclude by the process of medical science. Doctors are next to God, but “With all pros, comes their cons”. Government commands, staffing worries, patient safety along with the task of maintaining a minimum desired level of quality have always been a matter of concern for the hospitals. High hospital bills are always tension. Procrastination towards access to care and increasing human negligence give reasons for reducing faiths in doctors. At this time, the fist-sized organ of our body, Heart needs a healthy chronology. We all are so and so busy making our future that

we have completely forgotten, what we would do with such a good time in the future if we lose our existence in the present only. Actions like smoking, drug abuse, alcohol, or caffeine make us more susceptible to heart disorders. The situation is so vulnerable that 1 in every 5 heart attacks happens to be silent i.e. the damage is faced without the individual monitoring it.

This is where Machine Learning, comes into implementation. Machine Learning and is algorithms are used around in different areas. Medical science is no exception. Key insights if can be known at the right time can be of much use. Following the trends and patterns, these algorithms can foresee the occurrences of heart diseases and even classify them. In light of the clinical reports and symptoms of a patient the proposed model predicts whether a patient is having a coronary sickness or not. During the process of modeling algorithms of machine learning such as Support Vector Machine Classifier, Random Forest Classifier and Decision Trees Classifier are utilized. Alongside the libraries of Numpy, Matplotlib, Pandas, Seavorn and Scikit Learn are actualized to anticipate the output label. Also, Hyperparameter Tuning is done to make the model sensible and efficient by enhancing the accuracy ratio. Further, the model deployment is done using Django.

2. DATASET AND PARAMETERS EXPLAINED

The data which is used belongs to the area of life which is a combination of four databases of Hungary, Cleveland, VA Long Beach, and Switzerland. The dataset used is a subset containing 14 attributes out of total 76 attributes. The shape of the dataset is 1025 rows X 14 columns. It consists of multivariate characteristics that are used for ascertaining the presence of chronic heart disease in a patient or not. Some of these characteristics are age, sex, resting blood pressure, serum cholesterol, maximum heart rate achieved, fasting blood sugar, chest pain type, resting electrocardiographic results, etc. The attribute named 'target' holds the integer value 0 for invalid class and 1 for the valid class which refers to the presence or absence of heart disease. Table 1 shows the Data Dictionary associated with the chosen Dataset.

People at an elderly age are more prone than younger people to heart attacks and coronary heart diseases. Ageing is one of the key factors resulting in heart failure. As time grows from our childhood, our hearts can't beat as fast it used to be when we were young. Along with the rate of heart pumping blood, some fats start getting built upon on the inner surface of the vessels. This results in the uneven flow of blood which in turn concludes in hardening and stiffness of arteries. With time. The heart muscles become weak and end in heart failure. The column containing Age includes all the numerical values ranging from 29 years to 77 years which helped us in getting insights on a varied age group of people. Using sex as a parameter for judging, assisted us to find some beautiful patterns. The column named Sex contains categorical values with labels 0 for females and 1 for males. Science exclaims Men are more prone to heart diseases than women up to the time of Menopause. Once, the menopause is achieved, men and women stand on equal chances of getting heart attacks. However, there stands an exception to this, if a female is diabetic or undergoes any other metabolic disease, she is more vulnerable to heart diseases as compared to a male.

Chest Pains like Typical Angia, Atypical Angia, Non-anginal Pain, and Asymptomatic Pain were considered for the analysis. The attribute named 'cp' contained all the nominal values for these labels. Harvard Health Publishing says, out of all Americans admitted for the chest pain only 20% are

Table 1: Description of Data

Attribute name	Data type	Description
Age	Continuous	Age of the patient in years
Sex	Binary	Sex of the patient (1: male, 0: female)
CP	0-3	Chest pain type of patient
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
FBS	Binary	Fasting blood sugar > 120 (mg/dl) (1: true, 0:false)
Restecg	0-2	Resting electrocardiographic results
Thalach	Continuous	Maximum heart rate achieved
Exang	Binary	Exercise induced angina (1: yes, 0: no)
Oldpeak	Continuous	ST depression induced by exercise relative to rest
Slope	0-2	The slope of the peak exercise ST segment
CA	0-3	Number of major vessels (0-3) coloured by fluoroscopy
Thal	0-2	(0:normal, 1: fixed defect, 2: reversible defect)
Target	Binary	(0: Patient does not have a heart problem, 1: Patient has a heart problem)

susceptible to heart attacks. However, Chest Pain is treated as one of the important symptoms for heart attacks if there is a gradual outbreak of pain and the pain is in the diffused area. Blood Pressure is a result of the stiffness of the arteries, which creates hypertension. Because of high blood pressure, plagues grow inside the arteries which hardens and narrow the space for oxygen-rich blood flow. This, in turn, limits the transportation of oxygen to different parts of our body. The attribute 'trestbps' contains numeric data for resting blood pressure in a range of 94 mm Hg to 200 mm Hg. Column 'chol' shows the parameter Serum Cholesterol containing numeric values in a range of 126 mg/dl to 564 mg/dl. Low-Density Cholesterol narrows the arteries therefore, increasing the risks of the heart whereas High-Density Cholesterol is less harmful comparatively.

High Fasting blood sugar can damage veins and arteries and hence can obstruct the smooth flow of blood. People with diabetes are more prone to health diseases and tend to acquire heart diseases at an early age. The ST-T curve variations of the electrocardiographic results are monitored for predicting trends. Column 'restecg' contains 3 ordinal values for the same. The increase in the probability of heart attack is closely related to the Maximum Heart Rate achieved. It is observed that an increase in the heartbeat of 10 beats per minute results in increasing the chances of heart diseases by 20%. 'Exang' contains nominal values of exercise-induced angina which is in turn related to different types of chest pains. 'Oldpeak' and 'slope' represent the ST depression by exercise as well as the slope of the ST depression curve respectively. An abnormal treadmill ECG test is when the ST curve is horizontal or down-sloping, on the other hand, it is treated as normal when the curve is sloping upward. The horizontal and down-sloping curve individuals are more prone to heart diseases and disorders as compared to up-sloping one.

3. LITERATURE SURVEY

In the past few times, a lot of research works have been carried out in this domain of medical detection and classification of cardiovascular disease using various concepts and algorithms of Machine Learning and Data Science. An intensive study on the research of such papers which focuses on the similar domain is done under this section.

Reference [1] (Luis Polero et al., 2020) found that there was no conspicuous confirmation of a demonstrative approach using objective and subjective information about the characteristics of chest pain. Further, they assessed the performance of a Random Forest classifier to predict the chance for non-ST segment elevation intense coronary disorder (ACS) in a patient. Further, they found that it was helpful to foresee acute coronary syndrome but the general estimation took around 30-days and also the accuracy could have likewise be improved. Reference [2] (Wei et al., 2020) suggested a nomogram for the prediction of heart disease in patients having rheumatoid arthritis by assessing the F1-score of machine learning models such as decision tree, logistic regression, XG gradient-boosting, K-nearest-neighbors, random forest and support vector machine with increasingly stable execution. Reference [3] (Ye et al., 2020) exclaimed a risk assessment tool by using machine learning-based algorithm such as XG Boost and tested on the information gathered from electronic health records comprising details of older patients and found that the tool accomplished an improved discriminative capacity that can promptly be deployed in the health care system to provide automatic early admonition to older patients with increased fall risk. Reference [4] (Nusinovici et al., 2020) used the five different machine learning algorithms such as single-hidden-layer neural network, support vector machine, random forest, gradient boosting and k-nearest neighbour which were compared to logistic regression based on performance for the prediction of cardiovascular diseases, chronic kidney disease, diabetes and hypertension with simple clinical parameters. Further, they found that the best performing models were neural network and logistic regression.

Reference [5] (Wang et al., 2020) worked on an improved bagging algorithm which joined a resample strategy, a neural network, and a support vector machine. It was utilized to predict the risk factor for heart failure patients using electronic health record data. But the accuracy acquired by them was quite low so as to be deployed as a valuable tool. Reference [6] (Fu and Y., 2020) found that the predictors associated with economic condition plays a significant role in the prediction of heart disease by applying random forest and extreme gradient boosting to overall 89 predictors. They observed that the extreme gradient boosting outperforms all other machine learning algorithms with the accuracy higher than 90%. Reference [7] (Thapa et al., 2020) suggested two computational methods RF-MaloSite and DL-MaloSite based on machine learning algorithm for site prediction in proteins based on their primary amino acid sequences and found that the DL-MaloSite algorithm outperforms RF-Malosite algorithm by having an MCC score of 0.51 and 0.49. But the model had a high scope of improvement in the prediction system taking the size of the data into consideration. Also, multi-windows input can be used for upgradation. Reference [8] (Rajkumar et al., 2019) found that the grasshopper optimization algorithm with twin SVM classifier, when used for the prediction of CAHD, performs efficiently and accurately where the effectiveness of the classifier was measured by the parameters like performance metrics sensitivity, accuracy and elapsed time. But, the optimization algorithm utilized resulted in slow convergence speed. Reference [9] (Ashraf et al., 2019) exclaimed that the machine learning techniques do not reflect the genuine

capability of any algorithm when tested on the single dataset. So for the heart prediction, they suggested an automated system using deep neural network methods which were tested on multiple datasets but it is observed the results could have been improved more in terms of accuracy.

Reference [10] (Maini et al., 2019) suggested an approach to remove redundant or irrelevant attributes and to distinguish the significant attributes from the Cleveland heart disease dataset by using few feature selection algorithms such as filter feature selection algorithm (MRMR, Relief), wrapper (genetic algorithms) and embedded feature selection algorithms. Further, the prediction model was built by applying algorithms of machine learning such as logistic regression, k-nearest neighbors, random forest, Naive Byes and support vector machine. But the performance of the system could have been progressed by actualizing other feature selection techniques as well, in order to further increase the accuracy. Reference [11] (D. Panda et al., 2019) have shown the comparison of accuracies of the various classification algorithms such as Logistic regression, Random Forest, Extra Trees and Naive Bayes, before and after using feature selection methods such as least absolute shrinkage, selection operator (LASSO) and Ridge regression. Better outcomes would have been accomplished by including optimization techniques. Reference [12] (A. Bhardwaj et al., 2019) proposed an approach of identifying the risk factors and predicting the overall risk of having a heart disease using the machine learning algorithm- logistic regression on the Framingham heart study dataset. Further, the model outcomes with an accuracy of 87%. Improvement in the prediction model could have been carried out by including optimization and feature selection techniques with a need for model deployment framework for necessary utilization. Reference [13] (Yekkala et al., 2018) exclaimed a hybrid approach including more than one techniques such as Random Forest algorithm and Feature Selection using rough sets on the heart statlog dataset from the UCI repository for prediction of heart diseases. Further, they found that utilizing the Random forest algorithm with the Feature selection method results in high accuracy prediction of a patient having heart disease but, better recall values would have been achieved by utilizing and comparing other feature selection methods and machine learning algorithms.

A study by [14] (H.M. Le et al., 2018) suggests an automatic Heart Disease prediction method which includes Infinite Latent Feature Selection method, Support Vector Machine Classification and SMOTE (Synthetic Minority Over-sampling Technique) performed on UCI machine learning repository. The accuracy accomplished was 97.87%. Reference [15] (D Jain and V Singh, 2018) inspected a model identified with various feature selection and classification techniques for the presence of chronic heart diseases. Further, in their review of feature selection algorithms, they found that the filter method generally outperforms other algorithms and a hybrid approach is computationally progressively proficient as it integrates the merits of more than one algorithm. Reference [16] (Usman et al., 2018) proposed two cuckoo-inspired algorithms- cuckoo search algorithm (CSA) and cuckoo optimization algorithm (COA) for the feature selection. Further, they found that CSA filter-based feature selection outperformed the COA filter-based feature selection in terms of accuracy. Further scope of improvement can be seen by implementing a hybrid approach of feature selection. Reference [17] (Anand H. et al., 2020) has suggested a prototype for chronic heart disease detection and classification based on Machine Learning and Data Science algorithms such as Decision Tree classifier and Random Forest classifier. Further, Hyperparameters are tuned so as to improve the model. The approach lacked more efficient Feature Selection techniques like Wrapper methods, also considering few more algorithms for training would have been better. Reference [18] (Weng et al., 2017) proposed a cardiovascular disease risk prediction approach using machine learning and deep learning algorithms like gradient boosting and neural networks on Clinical Practice Research Datalink (CPRD). But, the proposed model needed a platform for implementation and practical approach.

Reference [19] (Sibghatullah I. Khan et al., 2019) worked on the detection of Coronary Artery Disease (CAD) using Spectral and Spectro Temporal Techniques. The different features of heart sounds are extracted using an electronic 3M Littmann 3200 stethoscope. Stockwell transformation is performed followed by bandwise spectral kurtosis. Finally, the classifying bandwidth is found using the Minimum Distance Classifier. Reference [20] (Porkodi1 et al., 2019) came up with a system implementing Gabor Filtration along with Random Forest Classification(RGRFC). The adopted techniques has been tested for their specificity, efficiency, precision and sensitivity on COPD Datasets. However, Gabor Filter comes with few demerits of being unsteered in nature, which means it requires to be done from different angles in order to gain a superimposed image for improving accuracy. Also, the attribute weights generated by Random Forest on the data seems not to be credible. Henceforth, the accuracy accomplished has more scope for improvement.

4. PROPOSED METHODOLOGY

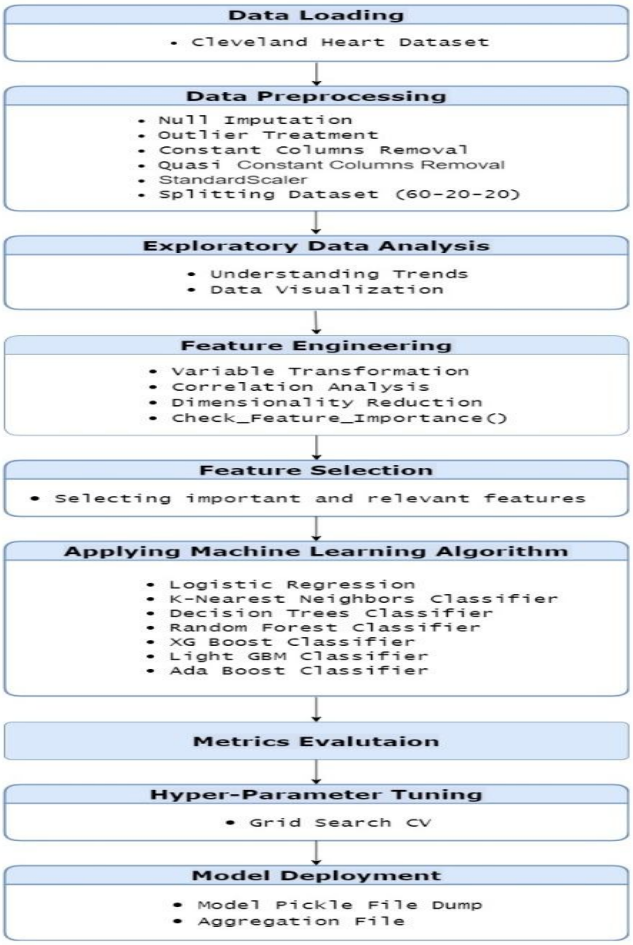

Figure 1: Implemented ML Pipeline

The model is built on the methodologies of Data Science and Machine Learning, which has a role to predict the occurrence of heart disease in a patient by learning from the previously diagnosed cases. Further, the model is deployed which can preferably be used by the doctors and researchers for an efficient and accurate prediction of heart disease on the basis of user-entered data. A detailed description of the Machine Learning pipeline being followed is stated below and in Figure 1.

A. Data Loading:

The dataset which is implemented for building this model is explained under the 'Dataset and Parameters Explained' section. The Dataset was loaded using the libraries of Pandas and Numpy. Figure 2 shows the head of the dataset. The Dataset contained missing and redundant values along with constant and quasi constant columns, removal of which is handled in the Data Preprocessing stage.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Figure 2: Head of the Dataset

B. Data Preprocessing:

This is one of the most crucial processes which makes changes and alters the dataset into the most useable and desired form by performing outlier treatment, null imputation, constant columns removal, quasi constant columns removal, re-scaling, and standardization. The presence of these results in inaccurate predictions, so they are needed to be removed and treated accordingly. Further, by using (60-20-20) rule the entire dataset is divided into 60% for training, 20% for validation, and rest 20% for testing. It is observed that the above distribution of dataset into training, validation, and testing was not having any class imbalance. Rather the data in training, validation, and testing have almost 50%-50% distribution of 1 & 0.

C. Exploratory Data Analysis:

Exploring the dataset provided us with important variables, some underlying trends, and assumptions. It helped in finding insights and hidden patterns about the dataset which in turn helped in understanding the problem and the dataset precisely. A lot of trends are captured in this experiment using the python packages of Matplotlib, Seaborn, Pandas, Scipy & Numpy. Analysis exclaims a high percentage of the young generation is affected by chronic heart disease which is shown in Figure 3. Also, it is observed that the Type 0 chest pain had the most frequency of a chronic heart disease followed by chest pain of Type 2 which is shown in Figure 4. Moreover, the resting blood pressure seemed to vary inversely with the occurrence of a chronic heart disorder as in Figure 5. Further, Figure 6. makes it evident that Type 0 and 4 blood vessels has more chances of having a chronic heart disease. Patients with resting ECG of Type 1 are more prone to heart borne diseases, followed by patients with Type 0 and then Type 2 which is obvious from Figure 7. Further, the patients with defects (Thal) of Type 0 and Type 2 are more likely to have the occurrence of heart diseases.

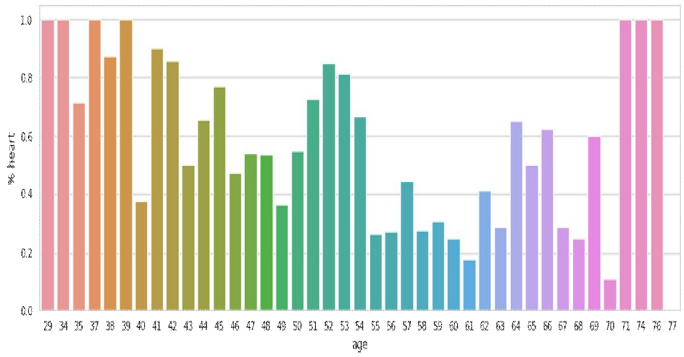

Figure 3: Age Wise Chronic Heart Disease Distribution

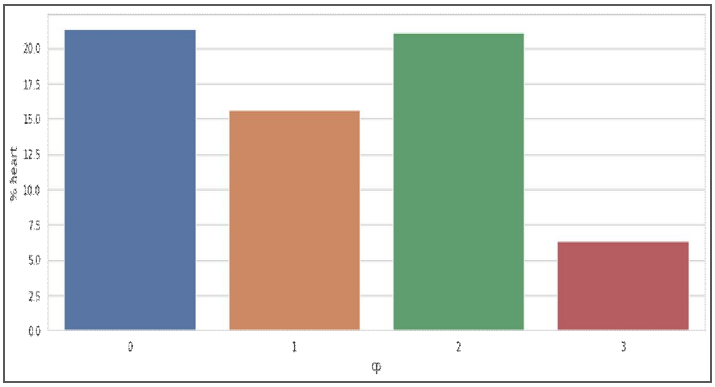

Figure 4: Percentage of Chronic Heart Disease Varying Over Types of Chest Pains

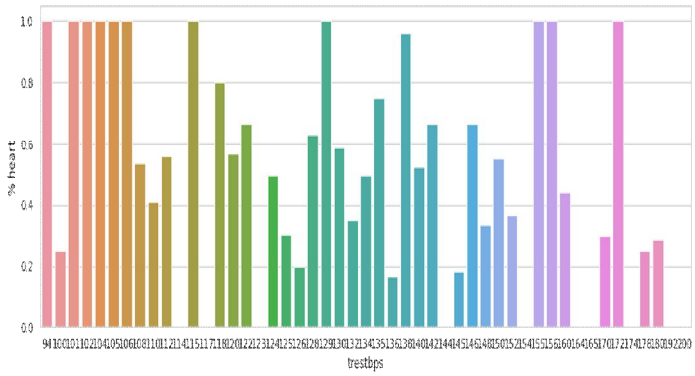

Figure 5: Percentage of Chronic Heart Disorders for Different Types of Resting Blood Pressures

D. Feature Engineering:

To increase the accuracy and the overall predictive power of the machine learning model, feature engineering is done. Under this stage, Variable Transformation along with Correlation Analysis is performed which also helped in reducing the dimensionality of the dataset. Important features are selected based on feature importance which is shown in Figure 8. After combining different useful characteristics from the dataset, multiple features are developed that are transformed on the basis of gaussianity.

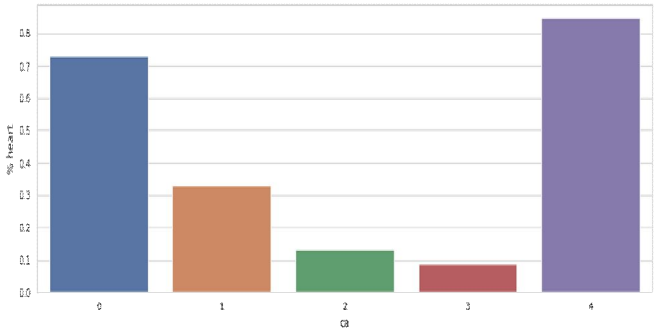

Figure 6: Percentage of Chronic Heart Disorder over Blood Vessel Types

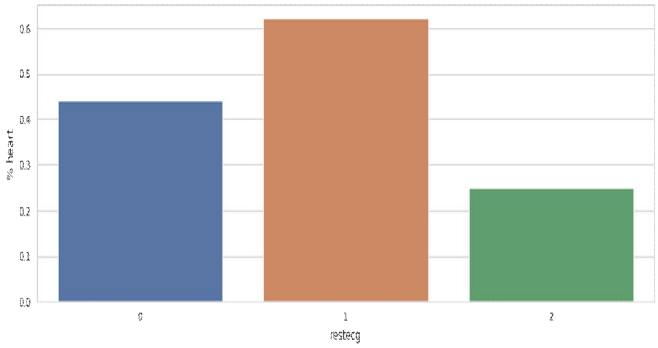

Figure 7: Variation of Chronic Heart Disease for Resting ECG Types

E. Feature Selection:

Important and relevant feature which contributed in the prediction of heart diseases are selected in this phase. They include Age, Chest pain, Exercise induced Angina, Resting blood Pressure, Serum Cholesterol, Slope of ST segment, Sex, Fasting Blood Sugar, Resting Electrocardiography, Maximum Heart Rate, ST depression relative to rest, , Number of vessels, and Defect Type. Further, the selected features are passed via Machine Learning algorithms.

F. Applying Machine learning algorithm:

Machine learning algorithms such as Naive Bayes Classifier, K-Nearest Neighbors Classifier, SVM Classifier, Random Forest Classifier, Decision Trees Classifier, XG Boost Classifier, Cat Boost, Light GBM Classifier and Ada Boost Classifier are used for training the data. Further, on comparing the performance of different machine learning algorithms it is observed that the Random Forest Classifier outperformed all other algorithms. The Random Forest classification model showed a recall of 0.97 for valid class and 0.93 for the invalid class for the testing set. The outputs and its aspects are well explained under the 'Implementation and Results Analysis' section. The algorithm associated with Random Forest classification is shown in Algorithm 1.

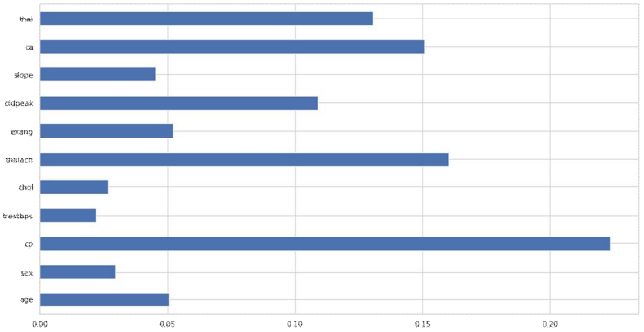

Figure 8: Feature Importance

G. *Metrics Evaluation:*

The performance of all the machine learning algorithms are evaluated using the confusion matrix. Further, the classification report which contained precision, recall, f1-score and accuracy result is calculated. The model was evaluated using more than one evaluation metric which ensured an optimal and efficient system.

H. Hyperparameter Tuning:

To improve the accuracy of the Machine learning model the parameters are optimized using Grid Search CV. A grid of parameters was defined and the cross-validation scores were improved. This optimization technique showed a remarkable change in the final accuracy.

Algorithm 1: Pseudo Code for Random Forest

To generate c classifiers:
for i = 1 to c **do**
  Randomly sample the training data D with replacement to produce D_i
  Create a root node, N_i containing D_i
  Call BuildTree(N_i)
end for

BuildTree(N):
if N contains instances of only one class **then**
  return
else
  Randomly select x% of the possible splitting features in N
  Select the feature F with the highest information gain to split on
  Create f child nodes of N, N_1,..., N_f, where F has f possible values (F_1, ... , F_f)
  for i = 1 to f **do**
    Set the contents of N_i to D_i, where D_i is all instances in N that match F_i
    Call BuildTree(N_i)
  end for
end if

I. Model deployment:

The last step of this pipeline- Model Deployment is crucial to understand the practical implications of this proposed approach. It can preferably be used by doctors for distinguishing the occurrence of heart disease in a patient more easily. Django, a web framework of python is used for this purpose. The final model is dumped and stored in the cloud using Joblib, Scikit Learn so as to make it accessible from the web services. From here, it can be utilized whenever and wherever there is a need of an efficient and immediate chronic heart disease detection just by filling a form. The form would get the results within fraction of seconds. Figure 9 shows a screen capture of the Deployed model.

Figure 9: Deployed Model

4. RESULT ANALYSIS

After the Exploratory Data Analysis, which helped in capturing hidden patterns and traits, Feature Engineering and Feature Selection are carried out. Variable Transformations are done, followed by Correlation Analysis. The heat map showed the interdependence of various attributes on each other which aided in the selection of better features. From the map, Chest pain (cp), Maximum value of heart rate achieved (thalach), and Slope of the ST depression curve (Slope) which contained cardinal values (1 = Up, 2 = flat, 3 = down) turned out to be better featured.

Dimensionality Reduction is done using a High Correlation Filter followed by Forward Feature Selection adding to a rich set of features for optimum and efficient performance. The feature importance is estimated using Scikit-learns's 'feature_importance_'. Figure 8 is bar graph showing importance of various features. As it is evident from the graph, Chest Pain (cp), Maximum value of heart rate achieved (thalach) and Number of major blood vessels (ca) are few of the features which were more important in explaining the target variable.

After training the data over various machine algorithms which included Naive Bayes Classifier, K-Nearest Neighbors Classifier, SVM Classifier, Random Forest Classifier, Decision Trees Classifier, XG Boost Classifier, Cat Boost, Light GBM Classifier and Ada Boost Classifier, it was discovered that the Random Forest Classifier worked best on

the training dataset. For the Validation set, the model predicted around 97% of the valid class and 96% of the invalid class correctly. And for the Testing set, 97% of the valid class and 93% of the invalid class was predicted correctly. Hyperparameter Tuning was performed using Grid Search CV. Finally, on combining the training set and validation set and training on the combined, it was observed that the final accuracy reached to a recall of 98% on valid class and a recall of 96% on invalid class.

5. CONCLUSION

In this present world, heart diseases are one of the most evident and undeniable reasons for death on a global level. According to WHO, cardiovascular diseases take an average of 17.9 million lives per year which accounts for 31% of all deaths worldwide. Apart from ageing, they occur as an outcome of poor lifestyle and activities like junk eating, smoking, drug addiction, negligence of resting diabetes level, and more. Doctors and medical processes are always there as a means to accomplish the need of the hour. But they are blessed with certain cons as discussed under Introduction. Here, the glamour of Machine Learning and Data Science can be introduced which is quite evident from our analysis. A study says, even the most skilled doctors can predict the occurrence of cardiovascular diseases with a precision of 67% only. On the contrary, Machine Learning algorithms can provide results with accuracy ranging in the 90s.

Machine Learning can act as an advantage over the limitations of human decision making and negligence towards access to care. They can be a way of getting rid of long pathological bills that too within time lesser than standing in long queues in front of hospital registration counters. Algorithms can provide outcomes whenever and wherever required right at the point of care. Just by simply asking for a few symptoms on a form, and passing them over a model to calculate predictions can spare a lot of time and most importantly will be in the budget of common people as well. Applying such algorithms in the field of healthcare would result in solving problems with enhanced productivity that too within an efficient time limit. The use of Data Science and Machine Learning as a means of chronic heart diseases detection and classification is an inexpensive, dynamic, and robust approach.

REFERENCES

1. Polero, L., Garmendia, C.M., Echegoyen, R.E., Alves de Lima, A., Bertón, F., Lambardi, F., Ariznavarreta, P., Campos, R. and Costabel, J.P., **A Machine Learning Algorithm for Risk Prediction of Acute Coronary Syndrome.** *Argentine Journal of Cardiology*, *88*(1), pp.9-13.
2. Wei, T., Yang, B., Liu, H., Xin, F. and Fu, L. **Development and validation of a nomogram to predict coronary heart disease in patients with rheumatoid arthritis in northern China.** *Aging (Albany NY)*, *12*(4), p.3190, 2020.
3. Ye, C., Li, J., Hao, S., Liu, M., Jin, H., Le, Z., Xia, M., Jin, B., Zhu, C., Alfreds, S.T. and Stearns, F. **Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm.** *International journal of medical informatics*, p.104105, 2020.
4. Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., Wong, T.Y. and Cheng, C.Y. **Logistic regression was as good as machine learning for predicting major chronic diseases.** *Journal of Clinical Epidemiology*, *2020.*
5. Wang, B., Ma, X., Wang, Y., Dong, W., Liu, C., Bai, Y., Bian, S., Ying, J., Hu, X., Wan, S. and Xue, W. **In-Hospital Mortality Prediction for Heart Failure Patients Using Electronic Health Records and an Improved Bagging Algorithm.** *Journal of Medical Imaging and Health Informatics*, *10*(5), pp.998-1004, 2020.
6. Fu, Y. **Assess the heart disease risk of the Chinese elderly using a predictive model.** *Advances in Social Sciences Research Journal*, *7*(2), pp.251-262, 2020.
7. Thapa, N., Hiroto, S., Roy, K., Newman, R.H. and Dukka, K.C. **RF-MaloSite and DL-Malosite: Methods based on random forest and deep learning to identify malonylation sites.** *Computational and Structural Biotechnology Journal*, *2020.*
8. Rajkumar, R. **Grasshopper Optimization Algorithm based Feature Selection with Twin Support Vector Machine Classifier for Coronary Artery Heart Disease Prediction.** *International Journal of Control and Automation*, *12*(6), pp.256-267, 2019.
9. Ashraf, M., Rizvi, M.A. and Sharma, H. **Improved Heart Disease Prediction Using Deep Neural Network.** *Asian Journal of Computer Science and Technology*, *8*(2), pp.49-54, 2019.
10. Maini, E., Venkateswarlu, B. and Gupta, A. **Role of Feature Selection in Building High Performance Heart Disease Prediction Systems.** *ADBU Journal of Engineering Technology*, *8*(2), 2019.
11. Panda, D., Ray, R., Abdullah, A.A. and Dash, S.R. November. **Predictive Systems: Role of Feature Selection in Prediction of Heart Disease.** In *Journal of Physics: Conference Series* (Vol. 1372, No. 1, p. 012074). IOP Publishing, 2019.
12. Bhardwaj, A., Kundra, A., Gandhi, B., Kumar, S., Rehalia, A. and Gupta, M. **Prediction of Heart Attack Using Machine Learning.** *IITM Journal of Management and IT*, *10*(1), pp.20-24, 2019.
13. Yekkala, I. and Dixit, S. **Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection.** *International Journal of Big Data and Analytics in Healthcare (IJBDAH)*, *3*(1), pp.1-12, 2018.
14. Le, H.M., Tran, T.D. and Van Tran, L.A.N.G. **Automatic heart disease prediction using feature**

selection and data mining technique. *Journal of Computer Science and Cybernetics*, *34*(1), pp.33-48, 2018.

15. Jain, D. and Singh, V. **Feature selection and classification systems for chronic disease prediction: A review**. *Egyptian Informatics Journal*, *19*(3), pp.179-189, 2018.
16. Usman, A.M., Yusof, U.K. and Naim, S. **Cuckoo inspired algorithms for feature selection in heart disease prediction.** *International Journal of Advances in Intelligent Informatics*, *4*(2), pp.95-106, 2018.
17. Anand H., Anand A., Das I., Rautaray S.S., Pandey M. (2021) **Hridaya Kalp: A Prototype for Second Generation Chronic Heart Disease Detection and Classification**. In: Gupta D., Khanna A., Bhattacharyya S., Hassanien A., Anand S., Jaiswal A. (eds) *International Conference on Innovative Computing and Communications.* Advances in Intelligent Systems and Computing, vol 1166. Springer
18. Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N. **Can machine-learning improve cardiovascular risk prediction using routine clinical data?.** *PloS one*, *12*(4), 2017.
19. Kumar, G Ganesh & Khan, Sibghatullah & Basha, Mohammed. **Preliminary Diagnosis of Coronary Artery Disease from Human Heart Sounds: A Signal Processing Prospective.** *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE) (Vol.8, No. 3),* 2019.
20. Porkodi1, V & Karuppusamy, S. **Preliminary Diagnosis of Coronary Artery Disease from Human Heart Sounds: A Signal Processing Prospective.** *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE) (Vol.8, No. 5),* 2019.