iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46355

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com

Detection of Heart Attacks Using Machine Learning

Rohith M V[1], Dr. Ravikumar G K[2], Ms. Nandini N S[3]
[1]*Dept. of CSE, BGS Institute of Technology Adichunchanagiri University, BG Nagar, Karnataka, India-571448*
[2]*Professor & Head (R&D) Dept. of CSE, BGS Institute of Technology, Adichunchanagiri University, BG Nagar, Karnataka, India-571448*
[3]*Dept. of CSE, BGS Institute of Technology, Adichunchanagiri University, BG Nagar, Karnataka, India-571448*

Abstract: ***Heart attacks, also described as cardiac arrests, are a variety of heart-related illnesses, which are now among the main reasons of death in the globe during the recent years. Globally, CVDs are thought to be the cause of about 31% of fatalities. It represents the apex of long-lasting processes that entail intricate interactions between risk variables that can and cannot be changed. The majority of coronary heart disease symptoms can be attributed to hypertension, and the majority of cases are thought to be undesirable. Itself was selected to test a few of techniques to determine how well their anticipated outcomes replicate or enhance the outcomes acquired prior to ML becoming preferred strategy for the advancement of forecasting analytics in the medical care sector. In order to help the medical sector and experts, investigators use a variety of information mining and machine learning algorithms on a set of vast data of cardiac victims to detect heart disease earlier they happen. This study uses a variety of Supervised ML classifications, including Gradient Boosting, Decision Tree, Random Forest, and Logistic Regression, to develop a system for the forecasting of Myocardial Ischemia. It makes use of the already-existing information from the Framingham library as well as those from the UCI Heart repositories collection. This study aims to construct a forecast for the likelihood that patients will experience a cardiac event.***
Keywords: Heart Attack, Machine Learning(ML).

I. INTRODUCTION

The living body's major tissue, the heart, pumps blood into every portion of the anatomy via the cardiovascular system's blood veins. The cardiac plays the most significant role in the respiratory system [1]. The central nervous system is the most crucial component of our organism because it is in charge of moving blood that carries nutrients, oxygen, water, minerals, and other vital substances during most of the organism. If the heart's normal functioning is compromised for any reason, it may result in major health problems, possibly inevitable extinction. The phrase "cardiovascular" is used to describe illnesses that modify or influence the architecture or functioning of the cardiovascular and respiratory systems, with atherosclerosis being the most widely recognized type of cardiovascular events. The incidence of the most widespread cardiovascular diseases (CVDs) reflects the peak of incurable conditions with intricate interconnections among risk variables that can and cannot be mitigated. The majority of coronary heart disease occurrences can be attributed to modifiable risk variables, and the majority of cases are thought to be avoidable.

Conventional measures to avoid cardiovascular illnesses have centred on alterable individual behaviour [2]. Obesity, cigarette usage, poor nutrition, and insufficient physical activity are the main causes of established risk aspects for cardiovascular illnesses, including diabetes, antihypertensive, cardiomyopathy, and the improvement of plaque. As of right now, heart attacks and strokes account for the majority of the 17.9 million annual deaths caused by cardiovascular problems (CVDs). 31% of fatalities worldwide occur in the manner described above [3]. Heart and blood vessel illnesses, sometimes known as cardiovascular diseases (CVD), are a common kind of illness.

II. RELATED WORK

According to early predictions, machine learning (ML) can accelerate the development of forecasting analytics in the medical sector. It was decided to put various methods to the trial in order to assess what well respective estimate evaluations would replicate or outperform the information collected utilising the classic Massachusetts strategy. In terms of clinical practise, this framework is among the most significant cardiovascular failure risk stratification models. There are quite a number of CV-risk prediction systems proposed so far [32]. It is still possible to see performance issues with the many methods that have been employed to determine CV-risk.

Particularly, several assessments like the Framingham Rating and the Comprehensive Cardiovascular Vulnerability Assessments have a tendency to underestimate patients' cardiovascular riskUtilizing well-known ML methods as k-nearest neighbour, support vector machine, classification, gradient boosting, regression analysis, regression models, and random forest, a comprehensive analysis has been conducted for the prediction of CV risk[3]. Naive Bayes, SVM, and KNN were the most pessimistic classifiers for heart attack prediction when comparing various feature selection ML techniques. Another method for predicting the risk of myocardial infarction involves randomly splitting the information and using outdated data mining techniques like J48, REPTREE, Naive Bayes, Bayes Net, and CART. In terms of myocardial infarction forecasting, the implemented system was capable of responding to more complicated questions.

The study was conducted in February 2021 to develop a model that makes use of an optimization approach as the inadequate sampling-clustering- oversampling technique, that also uses sample from the population under sample selection, grouping, and frame interpolation procedures (shortly, UCO algorithm). The training data for machine learning techniques were almost perfectly distributed, which made this technique different from others. With an efficiency of 70.29 percent, specificity of 70.05 percent, 1-Recall of 75.59 percent, and 0-Recall of 63.95 percent of the random forest, this approach was excellent at information extraction that were then evaluated on several classifications. Analysis has been done to determine how earlier heart attacks can be predicted by accounting for chest pain along with 24 other characteristics. Decision tree and random forest classification machine learning algorithms could be utilised to examine the cardiac event information. A clustering technique technique was employed for the deep categorization, and random forest was utilized to classify the objectives. However one method for predicting the likelihood of developing heart disease involved spontaneously going to split the set of numbers into number of partitions using a mean-based clustering methodology, and then employing uniformly distributed combination built using different regression and categorization tree models utilising an accuracy- based weighted ageing learning algorithm combination.

There must have been two databases, Strength and durability properties and Cuyahoga, that had classification performance rates of 93% and 91%, correspondingly. A study was conducted in July 2020 to develop a model that uses two separate approaches to estimate the prevalence of coronary heart disease. The support vector machine (SVM) was originally worked and precisely adjusted for its specifications, and after learning the Svm classifier 1000 magnitudes, the accuracy obtained achieved for the model's ability to anticipate cardio-vascular condition precisely was up to 96.5 percent with its median recall rate 89.8 percent while the detection precision utilising K- nearest neighbours achieves to 92.9 percent.

III. RESEARCH METHODOLOGY

A dataset is typically defined as a group of data that has been organised with a goal in mind. In this study, two separate databases were employed. The first information we employed is the Clinicopathological information, which was released by Georgetown University, the National Institute of Heart (NIH), and the Norwegian Cardiovascular, Pulmonary, and Blood Institute (NHLBI). This dataset is solely focused on finding markers for cardiovascular disorders, particularly heart assaults and heart conditions. The second sample we utilized is the UCI Machine Learning Repository's Heart dataset. The "Hungarian Department of Cardiovascular, Hungary," "University Hospital, Berne, Germany," "University Hospital, Geneva, Luxembourg," contributed information to the creation of this information. Additionally, the websites for both of the aforementioned databases are Kaggle.

A. Framingham Dataset

The Harvard dataset's characteristics are broken down into four categories: biographical, behavioural, preceding healthcare history-based, and existing healthcare condition-based. Demographics Characteristics: Sex: Classified as 0 or 1, with 0 denoting female and 1 denoting male. Age: The participant's older at the time of the assessment • Schooling: This is an irrelevant piece of information even though a person's standard of education has no bearing on any given medical problem. Behavioral: Current Smoker: A client is classified into either 0 or 1 based on whether they presently smoked or not; 1 is for yes and 0 is for no.

Cigarettes Smuggled Per Day: The average number of cigarettes smoked by a person per day depends on how frequently he smokes. Information based on prior medical histories: • Diabetes: Defined as either 0 or 1, with 1 denoting presence of diabetes and 0 denoting absence. • BP Meds: Patients are categorised as either 0 or 1 depending on whether they are taking blood pressure medicine or not. A score of 1 indicates that the patient is taking medication, while a score of 0 indicates that they are not. • Prevalent Stroke: Regardless on for certain if the sufferer has ever experienced a stroke, they are categorised as either 0 or 1, where 1 means they have, and 0 means they have not. • Prevailing Hyp: Regardless on how much the sufferer had hypertensive, the classification was either 0 or 1. (abnormally high blood pressure).

B. UCI Dataset

There are a total of 13 factors in the development in the database. "target" stands for the target value. Age (age): The participant's age at the moment of the assessment. Sex (sex): 0 or 1, with 1 designating a male and 0 a female, respectively. Chest Pain (cp): Classified into four sections, with scores ranging from 0 to 3, where scores below 0 indicate classic angina, those above 1 indicate abnormal myocardial infarction, scores below 2 indicate non-anginal irritation, and those below 3 indicate aspirational suffering. mmHg reading taken by the patient when at rest (trestbps) (unit). chol: The patient's triglyceride score in milligrammes per deciliter (unit). Fasting blood sugar (fbs) is categorised as 0 or 1, with instances such as 1 = if fbs > 120 mg/dl (true) else 0 (false). From 0 to 2, a resting ECG (restecg) can take one of three different shapes: benign, an inconsistent ST-T pattern, or left cardiac vascularity. Max Heart Rate (thalach): The highest heart rate any patient has ever attained. Exercise-induced angina (exang) is categorised as 0 or 1, where 0 means it doesn't exist and 1 means it does. Oldpeak: demonstrates the importance of ST depressed brought on by activity compared to repose in any discipline(float values). Slope: This term represents the maximum amount of activity during the ST phase. It has three intervals: 0 for an upslope, 1 for a level slope, and 2 for a downslope. Number of significant arteries (ca): Based on fluorescence-based colouring, it is categorised in the range of 0 to 4. From 1 to 3, there are three categories for thalasemia (thal), and 1 denotes normality, 2 indicates a fixed problem, and 3 denotes reversibility.

C. Preprocessing

Data pre-processing is the act of altering or encrypting data in such a way that it can be quickly and accurately interpreted by computers. In other words, material should be changed in a way that allows various algorithms to quickly understand it and produce results that are more accurate. Each dataset does not have to contain all pure data in its entirety. Almost any dataset contains some incomplete data in "NULL" form, thus causes the information to appear repetitious and causes the models to produce predictions with low accuracy.

Data pre-processing emerged as a solution to these low accuracy issues in order to achieve more and superior accuracy levels. We often remove the tuples with incomplete data from the information, impute the mean or median elements of the relevant column, or use another hyperparametric optimizing technique to obtain the imputable quantities to replace the missing entries. Due to the fact that our model solely uses numeric data from the two databases utilized. Therefore, in order to maintain the provides data integrity collection and produce improved accuracy, we are employing mean and average restoration procedures in our developed framework to impute incomplete data.

Mean restoration is a technique for substituting incomplete data in datasets (i.e., "NA" or "NULL") with the parameter's average. And average replacement is the process of substituting the parameter's average for incomplete data (i.e., "NA" or "NULL") in a dataset. There is constant uncertainty on when to employ mean and median interpolation, even in the case of mean and median allegation. It can be said that anytime a variable exhibits a normally distributed, we can impute either the mean or the median. However, the median restoration is recommended over mean restoration if the variable indicates a positive skewness rather than a normally distributed.

IV. PROPOSED APPROACH

Two significant datasets from the domain of cardiology assessment are used in our suggested approach. The Framingham information is really the first database, and the UCI cardiac data source is the second. Regression Analysis, Clustering Algorithm Classifier, Random Forests, and Logistic Regression Classification algorithm are the four machine learning classifications that we are currently using. Here, we are applying an iterative methodology to the deployment of our model over the two datasets. Recursively, both of the initial databases are imported, verified for information loss, and from there pre-processed using interpolation of the incomplete data. The information is then divided into two sections: training data and instance data for evaluation. The database is divided between training and retention parts in a 9:1 ratio, meaning that 90% of the information will be employed to train the algorithm and the leftover 10% will be utilized to evaluate it. textual headings; the templates will take care of that automatically for you.

A. Processes of the Developed Framework

They are continuously applying each classification to the sample following dividing and pre-processing to check for its discrete probability distribution and achieved accuracy results. We are creating four refineries for every classification by improving and optimising certain characteristics from the prior phase.

A refinery is a method for writing code and organizing the workflows of the model-building process in computer vision. Among a combined amount of sixteen pipework, each classification algorithm has four transmission lines, the first without even any enhancement, the second with combination of input enhancement (i.e., marked as "hpo-1"), the third with preceding predicted optimal enhancement and algorithm development (i.e., labelled as "hpo-1 + fe"), and the fourth with recent combination of input enhancement, model evaluation, and ultimate modifying once more with combination of input enhancement (i.e., characterised as "hp Following these three sessions on every classifier's workflow.

The best four pathways from each classification are then combined for analysis based on effectiveness and accuracy. The completed algorithm is then deployed for future forecasts using the sixteen pipelines that performed the best overall. The flowchart describing the entire process used to carry out this research

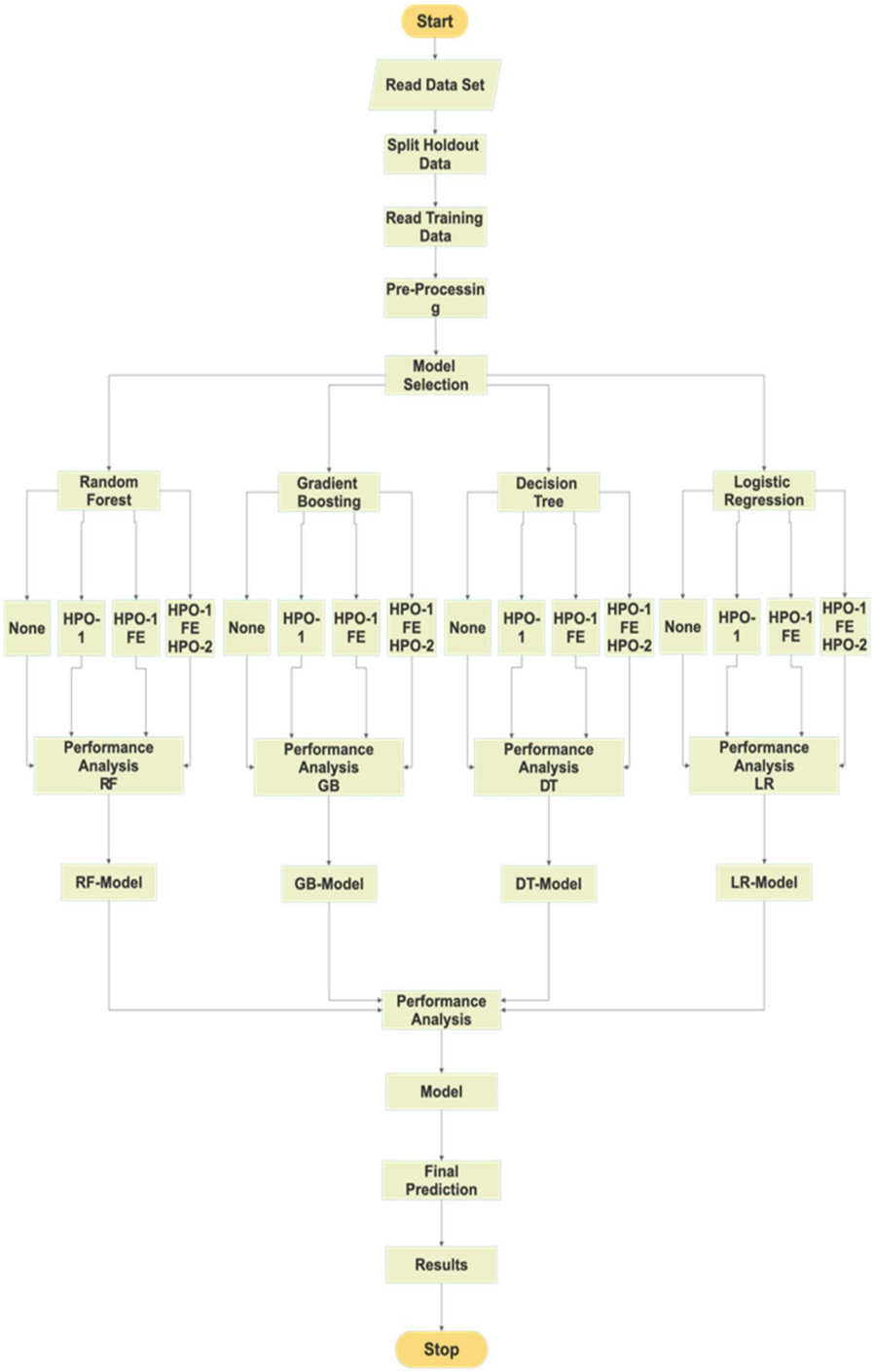


Figure 1: Flowchart for Proposed Model

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com

B. Algorithm

Flowchart-based algorithm for the developed framework

```
Set t to no of datasets
Set data[t]
    Set clf[ ]
    Set enhance[ ]
    Set x = 0
      Repeat while x < t:
      Repeat for i  in clf:
         Repeat for j in enhance:
         Deploy clf[i] with enhance[j] over data[x]
          Set best_enhance[i] = max(clf_accuracy)
          Set x = x+1
         Set best_clf = max(best_enhance)
      Set td to test data
      Use best_clf to predict td
      Return result
```

Preliminary sources include a list of data sources, a list of classifications, a list of data points, and a list of modifications to be made in every succeeding period. Following that, a few variables are declared for the model's work flow, such as x=0 when approaches each database in the collection info[] in the dowhile loop. Factors clf precision to be efficiency predictions foreach repetition of a classification improvements, good enhance[] is a dictionary of those improvements, and last good clf the maximum of those good enhance quantities for the projection of the best classification & improvements for developed and delivered.

V. RESULTS

Ibm watson machine learning, autoai-libs, scikit-learn, and other packages are installed for an exploratory configuration via IBM cloud Watson Studio. Following that, the pipeline is generated, and variables are set using the get params() method of the network optimization. Information about training processes and analysis methods is listed using the summary () method and is presented as a Pandas Data Frame. For each classification model, we also implemented a Scikit Learn ML Transmission framework to achieve the best precision possible after implementing all methodologies, i.e., without enhancement or extraction of features and with supervised learning, pattern discovery, or both consistently accompanied by predictive algorithms. Because the most difficult part of ML implementations has consistently been network optimization. Ibm watson ML, autoai-libs, scikit-learn, and other packages are installed for an exploratory configuration via IBM cloud Watson Studio. Following that, the pipeline is generated, and variables are set utilising the get params() method of the network optimization. Information about training processes and analysis methods is listed using the summary () technique and is presented as a Numpy arrays Data Frame. For each classification model, we also implemented a Scikit Learn ML Transmission framework to achieve the best precision possible after implementing all methodologies, i.e., without enhancement or extraction of features and with supervised learning, pattern discovery, or both consistently accompanied by predictive algorithms. Because the most difficult part of ML implementations has consistently been network optimization.

The method of training a system by choosing a set of ideal hyperparameters is known as supervised learning. In this portion of this research, all research methods for each classification and database are illustrated. Here, according to this graph plot, our article's forecast across a validate set of roughly 300 validate instances revealed though roughly 160 validate instances currently experience heart attacks, whereas roughly 140 test cases do not currently experience myocardial infarction in the coming years.

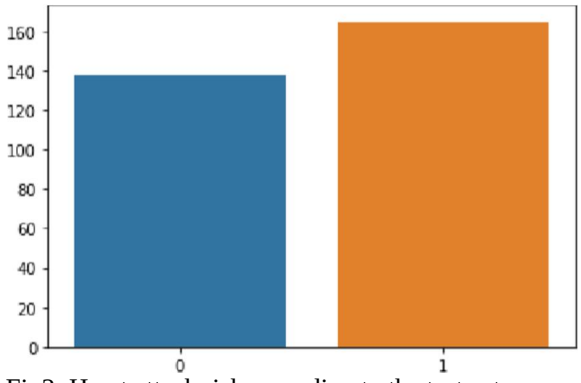


Fig2: Heart attack risk according to the test-set

Greater blood pressure, an elevated heart rate, chest discomfort, and higher serum cholesterol are all significant risk factors for cardiovascular diseases.

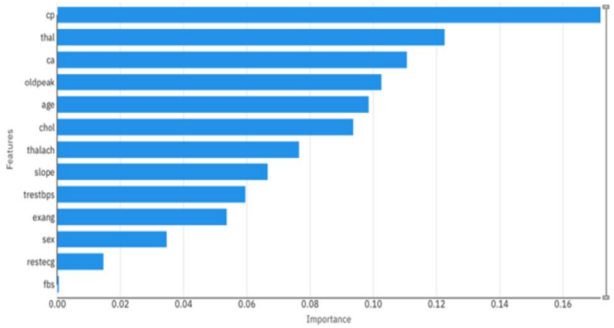


Fig3: significant determinants of heart attacks

Higher cholesterol levels (>200) and faster heart rates (>150) are associated with an increased risk of heart attacks.

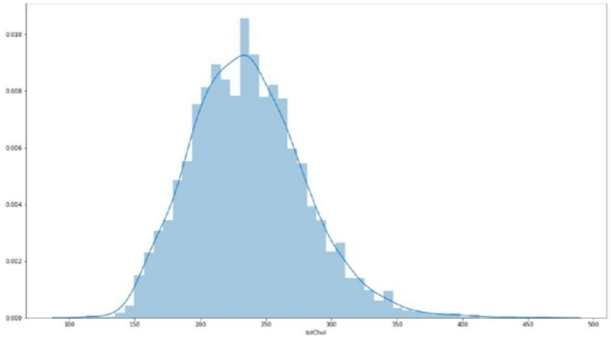


Fig4: A visual illustration of cholesterol levels in relation to the likelihood of a heart attack

According to the above visual analysis of total cholesterol (total saturated fat extent) in this research, The risk of experiencing a cardiac arrest due to high saturated fat starts to surpass at about >200mg/dl, with the maximum predicament having occurred at hyperlipidemia 250mg/dl, which is below the most especially vulnerable range of sustaining a heart attack (i.e. >200mg/dl), according to a bar chart that whether demonstrates dietary cholesterol variety vs. heart condition possible scenarios (as X-Y graph). People who experience frequent chest pain are more likely to suffer a heart attack. Traditional angina has a lower risk of heart attack than other forms of heart problems. Here, distinct forms (i.e., 0, 1, 2, 3) are divided into four categories, with numbers 0 to 3, denoting classic angina, unconventional angina, non- anginal pain, and asymptotic angina. According to this graph-plot, people who experience type-0 chest pain, or classic angina, possess a higher likelihood of heart attack than other people, whereas people who experience type-2 difficulty breathing, or non-anginal pain, are just slightly more likely to have a heart disease.

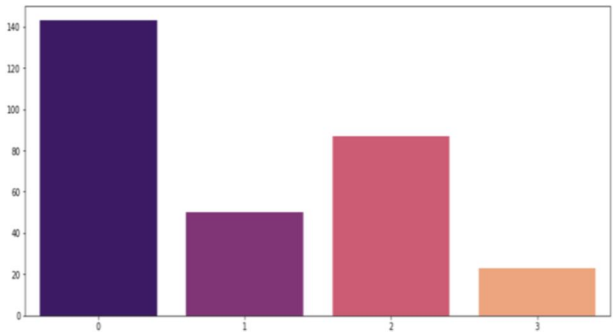


Fig5: Heart attack-prone kinds of chest pain

Everybody should regularly visit a doctor to have their blood pressure, heart rate, and cholesterol levels checked for prevention purposes. Any cardiac patient should also practise frequent mindfulness. And besides these preventative measures, one must speak with a doctor and routinely take the prescribed medications to reduce the likelihood of recurrent heart attacks.

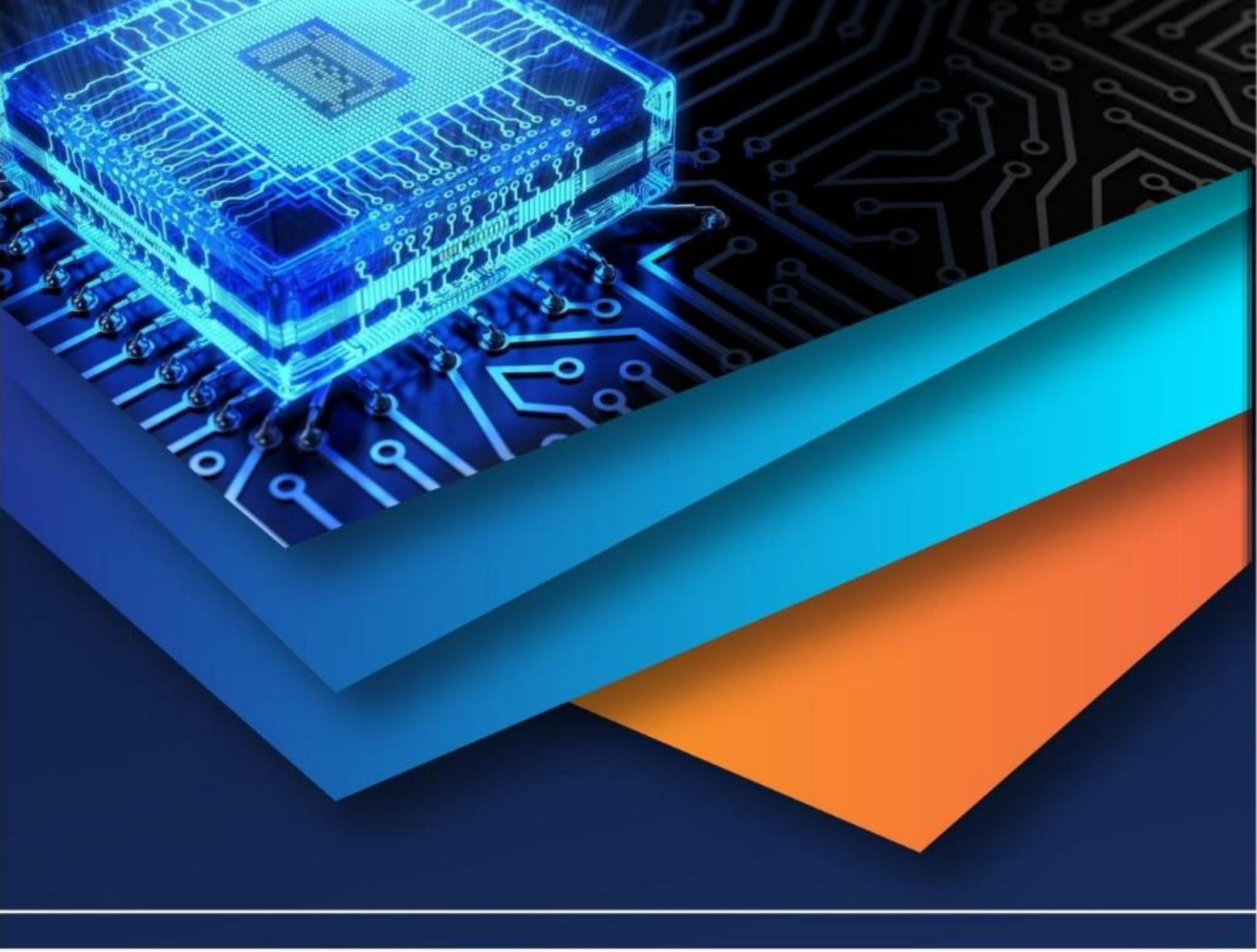
VI. CONCLUSION AND FUTURE WORK

In this study, a framework for the forecasting of myocardial ischemia has been deployed using a variety of trained machine learning techniques, including Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression. Despite the fact that both data sources contained a lot of inconsistencies, different feature transmission lines were employed to keep improving the consistency of the and to attain an average precision of 85,5% or a number of responses of 82% for the information using the Prospective cohort study dataset and a Support Vector classification algorithm, and an even higher recall rate of 89.1 percent when using the UCI Heart dataset. Gradient Boosting Classifier is used to deploy our final model for cardiopulmonary arrest predictions after these various approaches. We can further broaden this investigation by using semi- supervised and deep learning methodologies.

The data demonstrate that our approach predictions in digital digits, with 1 signifying a possibility of cardiac arrest and 0 representing no risk, and that the results reveal that the Gradient Boosting categorization is achieving the greatest accuracy score. Increasing heart rate, thal, & age are a few of the least crucial variables. Chest pain type—typical nitroglycerin is the greatest frequent and exponentially cardiac discomfort is indeed the lowest frequently level—levels exceeding 200 mg/dl seem to be more probable. Conclusion: By maintaining a healthy weight, getting regular exercise, and avoiding tobacco products, untimely heart strokes are prevented in 80% of cases. Additionally, those who consume upwards of 5 liters of water daily are less likely to experience outbreaks.

REFERENCES

[1] Heart Attack Prediction System Using Fuzzy C Means Classifier by R. Chitra, IOSR Journal of Computer Engineering, vol. 14, no. 2, 2013, pp. 23–31, doi: 10.9790/0661-1422331.
[2] H. S. Buttar, T. Li, and N. Ravi, “Prevention of cardiovascular diseases: Role of exercise, dietary Clin. Cardiol., vol. 10, no. 4, pp. 229–249, 2005.
[3] I. D. Mienye, Y. Sun, and Z. Wang, “An improved ensemble learning approach for the prediction of heart disease risk,” Informatics Med. Unlocked, vol. 20, p. 100402, Jan. 2020.
[4] M. Hortmann et al., “The mitochondria-targeting peptide elamipretide diminishes circulating HtrA2 in ST-segment elevation myocardial infarction,” Eur. Hear. J. Acute Cardiovasc. Care, vol. 8, no. 8, pp. 695–702, 2019, doi: 10.1177/2048872617710789.
[5] Cost-effective gradient boosting, Advanced Neural Information Processing Systems, vol. 2017- Decem, no. Nips 2017, pp. 1552–1562, 2017. S. Peter, F. Diego, F. A. Hamprecht, and B. Nadler.
[6] P. Severino et al., “Ischemic heart disease pathophysiology paradigms overview: From plaque activation to microvascular dysfunction,” Int. J. Mol. Sci., vol. 21, no. 21, pp. 1–30, 2020, doi: 10.3390/ijms21218118.
[7] A. Segura-Galindo, F. Javier Del Cañizo-Gómez, I. Martín-Timón, C. Sevillano-Collantes, and F. Javier Del Cañizo Gómez, “Type 2 diabetes and cardiovascular disease: Have all risk factors the same strength?,” 2014, doi: 10.4239/wjd.v5.i4.444.
[8] P. B. Lockhart and Y.-P. Sun, “Diseases of the Cardiovascular System,” in Burket’s Oral Medicine, John Wiley & Sons, Ltd, 2021, pp. 505–552.
[9] T. Mü nzel, M. R. Miller, M. Sørensen, J. Lelieveld, A. Daiber, and S. Rajagopalan, “Reduction of environmental pollutants for prevention of cardiovascular disease: it’s time to act,” doi: 10.1093/eurheartj/ehaa745.
[10] M. Ferrante et al., “Air Pollution in High-Risk Sites–Risk Analysis and Health Impact,” in Current Air Quality Issues, InTech, 2015 and Strategy”, IEEEE conference on Consumer Electronics, Communications and Networks, April-2012, pp. 1216 – 1219.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY