iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59266

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 12 Issue III Mar 2024- Available at www.ijraset.com

Heart Disease Detection Using Machine Learning

Arpit Verma[1], Mayank Pathak[2], Akhilesh Kumar Prajapati[3], Harsh Srivastava[4]
Babu Banarasi Das Northern India Institute of Technology, Lucknow

Abstract: ***One of the most common tasks in machine learning is to classify data. Machine learning is a key feature to derive information from corporate operating datasets from large databases. Machine Learning in Medical Health Care is an essential emerging field for delivering prognosis and a deeper understanding of medical data. Most methods of machine learning depend on several features defining the behavior of the algorithm and influencing the output and the complexity of the resulting models directly or indirectly. In the last ten years, heart disease is the world's leading cause of death. Many machine learning methods have been used in the past to detect heart disease. Neural Network and Logistic Regression are some of the few popular machine learning methods used in heart disease diagnosis. They analyze multiple algorithms such as Neural Network, KNearest Neighbors, Naive Bayes, and Logistic Regression along with composite approaches incorporating the above-mentioned heart disease diagnostic algorithms. The system was implemented and trained in the Python platform using the UCI machine learning repository benchmark dataset. For the new data collection, the framework can be extended.***

I. INTRODUCTION

Heart disease (HD) or cardiovascular disease is the most common human disease in the world that imposes a threat to life. In this state, the heart is usually unable to pump the required amount of blood to various parts of the human body during order to meet the regular functionalities of the body, resulting in irreversible heart failure. The United States has an extremely high rate of heart disease. Symptoms such as breathlessness, physical body weakness swollen ankles and tiredness with the associated sign, such as increased vein stress and peripheral edema from functional or non-cardiac disorders, are symptomatic of cardiac illnesses. Early-stage testing strategies for the diagnosis of HD have been difficult and their complexity is one of the main reasons for impacting the quality of living. Diagnosis and treatment of cardiovascular diseases are extremely complex, especially in developing countries, since diagnostic instruments are rarely available and medical practitioners and other services are missing that affect correct prediction and care of cardiac patients . A lot of research and study has been done over the last few years on better and reliable data sets for heart disease.

[4] offers a strategy based on knowledge. Initially, Dr.Robert Detrano used logistic regression to achieve a precision of 77%. Newton Cheung used algorithms with Naive Bayes and obtained a classification accuracy of 81.48% [6]. In treating patients with heart disease, Palaniappan and Awang explored the analysis of various data mining techniques. Naive Bayes, decision trees, and neural networks were used in these techniques. The results showed that Naive Bayes was able to achieve the highest precision in diagnosing patients with heart disease. Indira [8] uses the BF network of probabilistic artificial neural networks. This is a radial-based algorithm class RB approach is useful for automated pattern recognition,nonlinear modeling, and class composition probability estimation and probability ratios. With a total of 576 records and 13 patient features, the data used in the tests were drawn from the Cleveland heart disease archive. The best performance in accuracy was 94.60%. Among the most widely used methods for data mining in grouping issues is KNearest Neighbor . It is a popular choice due to its simplicity and relatively high level of convergence. Nonetheless, the huge memory space required to store the entire sample is a major disadvantage of KNN classifiers. If the sample is large there is also a large response time on a sequential computer

Python and machine learning have been used in our research to support libraries such as sci-kit learn, numpy, pandas and matplotlib.

II. MATERIALS AND METHODS

The goals of the research are to perform a" comprehensive analysis of different algorithms in machine learning and develop a more accurate algorithm." People these days' work for hours and hours on machines, they have no time to take care of themselves. The poor health of people is severely affected because of the unhealthy lifestyle, hexagonal habits, and intake of junk food. It can determine how much a person would die from heart disease with the medical parameters, which are adequate patients as well as normal people.

The task is to determine which studies will be positive and which will be inaccurate in the cardiovascular identification process. The following subsections address extensively the paper's analysis materials and procedures.

A. Data set Description

Various studies use" Cleveland heart disease data set 2016" and can be downloaded from the Universityof California, Irvine's online data mining depot. Thisdata set has been used in this study to develop a machine learning algorithm based system for thediagnosis of heart disease. The data set for Clevelandcardiac disease is 303 cases with 76 attributes and certain incomplete attributes. During the study, six samples were rejected due to a missed value of the task columns and a resulting sample size of 297 weretaken with 13 further separate input features. There are two groups on the goal performance label to describe a heart patient or a regular topic

B. Classification Problem

Throughout Machine learning the question is to classify, on the basis of a training data set of observers whose group membership is identified, which of a number of categories belong to a new observation. Classification is a supervised learning problem, i.e. problems with poorly labeled tuples andalgorithms. A classification algorithm is classified asa classifier in particular in an actual implementation .

C. Binary and Multiclass Classification

Classification can be viewed as two distinct binaryand multiclass classification problems. A better understood function of binary classification needs only two types, while multi type class classification includes the assigning of an entity to one of various classes.

TABLE I HEART DISEASE DATASET

Sr. NO.	Attribute	Description	Values
1	Age	Age in Years	Continuous
2	Gender	Male or Female	0=male, 1=female
3	Cp	Chest Pain Variety	4 = typical type 3 = typical type angina 2 = nonangina pain 1 = asymptotic
4	Thest bps	Sleeping Blood Force	Continuous
5	Chloe	Serum Cholesterol	Continuous
6	Rest Ecg	Resting ECG	5 = normal 4 = having ST T waveabnormal 3 = left ventricular hypertrophy
7	Fbs	Fasting Blood Sugar	1 >= 120 mg/dl 0 <= 120 mg/dl
8	Thalach	Greatest heart speed reached	Continuous
9	Ex ang	Exercise induced angina	1 = no 2 = yes
10	Old peak	ST despair induced by apply virtual to relax	Continuous
11	Slope	Slope of the height effect ST section	3 = un sloping 2 = flat 1 = downsloping
12	Ca	Number of key vessels painted by fluoroscopy	1-4 value
13	Thal	Desert Category	2 = normal 4 = fixed 5 = reversible defect

D. Comparison of methods of classification

The methods of classification can be analyzed andcompared with the following criteria:

1) Accuracy: A labeling device's accuracy relatesto its ability to properly interpret a tuple or unlabeled data object. The stronger the accuratethe algorithm will be.
2) Speed: Classifier speed refers to the amount ofclassification time required by the classificationalgorithm.
3) Robust: a classifier's sturdiness is its capacity to distinguish accurately when the data are chaotic and outliers.
4) Scalability: Scalability of a classifier is its ability to classify efficiently when large amounts of data is given.
5) Interpretability: It applies to the degree that the classifier offers for comprehension and perspective.

E. Machine Learning Classifiers

Machine learning recognition systems are used to identify heart patients and healthy people. In the present paper, we address briefly some popular classification algorithms and their theoreticalcontext.

1) *Logistic Regression:* In terms of its name, logistic regression is not a regression model but a linear classification. Logistic regression is alsorecognized in literature as logistical regression, maximum-entropy classification . In this model, a logistic function is used to predict the probabilities that characterize the future consequence of a single test. The logistic regression in scientific learning can be approached from the Logistic Regression framework. For optional L2 or L1 regularizationthis design will work in a multi-class (one vs rest)logistic regression. In order to optimize the logistic regression penalized to binary class L2, the following risk function minimizes:

$$\min_{\omega,c} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\log(exp(-y_i(X_i\omega + c)) + 1)$$

Likewise, the L1 regularized logistic regression addresses the following problem ofoptimization: n minωk1 + C Xlog($exp(-y_i(X_i\omega + c)) + 1$) ω,c i=1

A logistical regression description may begin with a logistics feature explanation. The logistical function is valuable because its valueis always between zero and one and can therefore only be viewed as a probability. It cantake any amount from negative to positive infinities. This describes the logistic function asfollows:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{e^t + 1}$$

2) *K-means Clustering:* In order to separatesamples in n classes that have the same variance,the K-Means algorithm cluster results, minimizing the inertia criterion or using a squarenumber. The number of clusters to be defined is needed for this algorithm. This applies well to a huge number of samples and is used in several different areas of use. The k-means algorithm partitions a set of observations N through K disjoint clusters C. Each is the average sample ofμj in the cluster. The mean square often referred to as" centroids;" usually they are not X points, although they reside in the same space. The K- means algorithm attempts to identify centers within a cluster sum that reduces inertia or squared criteria.

$$\sum_{i=1}^{n}\min_{\mu_j \epsilon C}(\|x_j - \mu_i\|^2)$$

Inertia can be recognized as a calculation of howinternal stability clusters is in the cluster total ofthe criteria for squares. There are several disadvantages:

- Inertia assumes clusters which is not alwaysthe case, are convex and isotropic. It does notreact adequately to extended clusters or multiples of uneven structures.
- Inertia is not a measure that has been normalized: we realize that lower andnegative values are better.
- Nevertheless, Euclidean distances appear to become overflowing in very high dimensional spaces (the" curse of depth" is an example). If a dimensional algorithm is used before K-means clustering, this issue could be solved and calculations accelerated.

3) *Naive Bayes:* It is a supervised algorithms collection focused on the theories of the implementation of Bayes with the" naive" presumption that each pair of features is independent. The Bayes theorem states that a class variable y and a dependent $x1$ through a xn feature have been defined:

$$P(y|x_1, ..., x_n) = \frac{p(y)P(x_1, ..., x_n|y)}{P(x_1, ..., x_n)}$$

With the naive assumption of independence
$P(y|x1,.\ x_i{-}1,x_i{+}1,.\ \ ,x_n) = P(x_i|y)$
This relationship has been simplified for all

$$P(y|x_1,....,x_n) = \frac{p(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,....,x_n)}$$

As the input is constantly given by $P(x1,..xn)$, thefollowing classifying rule can be used:

$$P(y|x_1,....,x_n) \propto p(y)\prod_{i=1}^{n} P(x_i|y)$$

$$\propto p(y)\prod_{i}^{n} P(x_i|y)$$

$$\hat{y} = \text{argmax}^{y}$$

=1

and we will use the approximation for $P(y)$ so $P(xi|y)$ for Maximum a Posteriori (MAP) to measure the relative class Y frequency in the training sets, then the former. The various Naive Bayes classification schemes vary mostly because they believe that $P(xi|y)$ will be spread. Though it seems to be oversimplified, simplistic Bayes classification schemes have performed very well in many real-world situations, such as text classification and spam filtering.

4) *K-Nearest Neighbors (kNN):* kNN is used for classification and inference by a non-parametric approach of pattern recognition. The input is the closest example of k in the feature space in both cases. kNN implies an instance-based learning or lazy learning that only approximates the function locally and delays all calculations until classification is completed. The KNN algorithm is one of the easiest algorithms to learn. Also for grouping as well as for estimation, weight can be assigned to the neighbors' inputs so that the near neighbors create a more reasonable contribution than the farther ones. A typical weighting scheme is to assign a weight of 1/*d* for each neighbor, where d is the distance from the neighbor. The underlying objects are drawn from a collection of items defined for the class (kNN classification) or the property value for the object (kNN regression).This can be seen as the algorithm training set but no specific training is required. The kNN algorithmhas a weakness because it is sensitive to the local data structure. Another common machine learningmethod, the algorithm must not be mistaken with K-means.

5) *Neural Network:* Machine Learning incorporates the most widely utilized form of the neural network, feedforwards artificial neural networks or, more simply, multilayer perceptron (MLPs). The MLP contains an input layer, output layer and one or several hidden layers. The growing MLP layer comprises of one or more neurons synchronized with neuron layers prior and next. Through MLP, each neuron is the same. Each has multiple input connections and multiple output connections. The values from previous levels, separately for each neuron, and the biological termare summed up in certain weights. The total is converted by the activation function f, which for multiple neurons may be specific. That is,provided xj in n line, yi in $n + 1$ layer outputs are determined as:

$$ui = \text{X}(wi,jn+1\ xj) + (wi,biasn+1\ j \quad *$$

$$yi = f(ui)$$

Various functions can be used for activation. Machine Learning has three standard features:

- Identity function :$f(x) = x$
- Symmetrical sigmoid : $f(x) = \beta$ $(1-e^{-\alpha x})/(1- e^{-\alpha x})$, is the default MLP choice. *
- Gaussian function :$f(x) = \beta e^{-\alpha x\, x}$, which is currently not completely supported.

The activation functions of both neurons in the machine learning process have the same free parameter with (α,β) specified by a user and not determined by the training algorithms. To measure the network, all wi,j^{n+1} weights are required. The training algorithm computes the weights. The algorithm receives a training set, includes multiple input vectors and iteratively adjusts the weights so that the network can respond to the supplied input vectors. The larger the network size, the more robust the system becomes. Arbitrarily, the loss in the training setcould be high. But the network learns the noisein the training set as well, so it usually begins to grow after the network size is limited. The errorsin the test set are usually increased. Furthermore, larger networks are much longer trained than smaller networks, so that only essential features can be prepared and a smallernetwork formed. Another MLP feature is that categorical data can -not be handled in the manner it is

6) *Overfitting and Regularization:* Over-fitting is amajor problem in neural networks. This is especially true in modern networks, which often have very large numbers of weights and biases. We need a way to detect how to train effectively when over-fitting is going on, so we don't overtrain. to over-train. And we want strategies toreduce the overfitting results. Regularization is one way of combating overfitting.

The regulation changes the target function by adding additional conditions to penalize large weights, which we minimize. In other terms, we are shifting the targetto the Error+ *lambdaf*(θ), where the power of regularization (a learning algorithm hyper-parameter), grows larger as the components of θ grown larger and λ becomes larger. A $\lambda = 0$ means that the possibility of overfitting is not taken. If λ is too high, our model will focus on keeping θ theparameter value as small as possible when trying to find the values which are good for our workout. Consequently, λ is a very important task and some trial and error may be required.

7) *Fuzzy K- Nearest Neighbors:* The method is based on assigning membership to the available groups depending upon the difference between the function and its closest neighbors and those neighbors. In that it still requires to scan the sample set labeled for the K-nearest neighbors, thefuzzy algorithm is identical to the crisp variant. The methods vary considerably beyond thecollection of the K samples. While the fuzzy K- nearest neighbor technique often reflects a rating algorithm, the results are different from the flat form. A class affiliation refers to a display function instead of a particular class the Fuzzy K-nearest neighbor algorithm. The drawbacks are that the program does not execute random tasks. Furthermore, the membership values of the variable should provide a confirmation standard that facilitates the subsequent classification. Large memory space is required to store the all thetraining data. It is also computationally expensiveand has a greater time complexity.
8) *K Means Clustering with Naive Bayes Classifier:* We initially implemented K Means individually. Then we worked on hybrid by combining K Means with naive bayes. K Means isused to group together similar data. Then nave bayes was implemented on each cluster and model was made. For each new test case, it was first determined to which cluster it belongs. Then the naive bayes model for that particular cluster was used to make prediction for the given test case. KMeans was used with the hope that grouping together similar data will help in increasing accuracy of naive bayes algorithm. Here we had atradeoff between time of computation and accuracy but additional gained accuracy was preferred. For this algorithm, we first discretized the data as naive bayes requires data which is in discrete form. We could not use gaussian Naive Bayes because the distribution of data was not gaussian and it would have contributed to poor accuracy if anything were used. We had two choices to discretize records, equal discretizing width and equal discretizing frequency. Equal variable widths contributed to improved performance.

F. Cross Validation

Cross-validation, also referred to as the estimation of rotation, is a testing technique to evaluate whether mathematical observational findings are translated toan individual set of data. It is primarily used in quantitative environments and one needs to approximate the actual output of a predictive model.A model generally contains a dataset of known data (training data set), a dataset of unknown data (or firstused data) against which the model is evaluated (testing dataset). In a prediction problem, a model canbe used. To order to limit issues like overfitting and provide an overview into how the model is to extendto a different dataset etc. For k-fold crossvalidation, the initial sample is divided randomly into k similar to k. The aim of the cross-validation is to create a dataset to" test" the model in the training phase. One sub-sample of k is held for validation of the model, and the remaining k 1 subsamples are used as rain results.The method of cross-validation then is repeated k times (folds), and the validation data is used preciselyonce for each of the k subsamples.

G. Performance Evaluation Metrics

Numerous performance evaluation tests have beenused in this work to validate the performance of classifiers. We used a confusion matrix, each result isprojected in precisely one chamber in the test array. There are two resting groups, 2 × 2 matrix. In comparison, there are two forms of accurate classifier estimation and two kinds of inaccurate prediction classifier. The following can be determined from theconfusion matrix: TP: expected performance as true positive (TP); we found an appropriate diagnosis of the HD topic, with cardiovascular disease.

We've noticed that a healthy person is properly labeled and the topic is safe. TN: The results predicted to be truly negative (TN).

FP: The conclusion is that it is wrong to classify a stable subject as having heart disease (Type 1 error): expected performance as false positive(FP).

FN: We consider that heart disease is incorrectly labeled as not having heart disease because the target is stable (type 2 error). Predicted to be false-negative(FN).

III. EXPERIMENTAL RESULTS ANDDISCUSSION

Two separate diagnosis classes were included in this study: normal and patient who may have heart diseases. Many researchers have suggested different methods for the detection of heart disease as stated inthe previous section. The accuracies recorded range from 50% to 87%. The collection includes 303 samples, 297 of which are full samples and six of which are incomplete samples. While the neural network ensemble model was trained in 70 percent ofthe cardiac disease database, the remaining 30% of the cardiac disease database was used for validating the proposed system. ROC curve is also one commonly used graph that summarizes the performance of a classifier over all possible threshold for assigning observations to a given class. Here, confusion matrix is used, which contains information__about actual and predicted classifications performed by a classification system. The performance is evaluated using the data in the matrix. The results ofthe various machine learning algorithms are shown infigure 1 From fig 1,

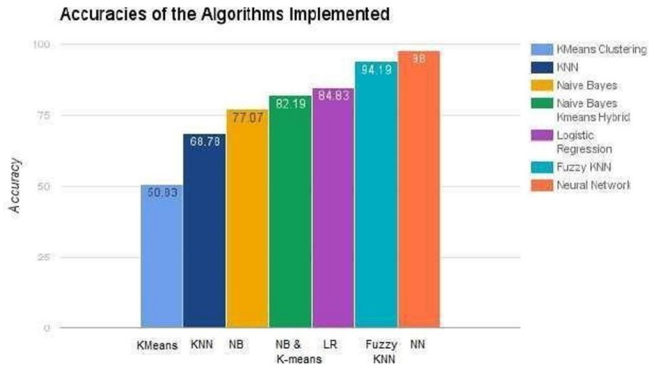


Fig. 1. Accuracies of various machine learning algorithms

it observed that the proposed system outperforms theexisting methods of machine learning. Using a confusion matrix, the system classification findings were shown. The cell comprises the total number of copies listed as an appropriate combination of expected and actual network outputs in a confusion matrix. The following Table II display the confusion matrix showing the results of this network's classification.

Our reports were contrasted with previous findingsrecorded using different approaches. Table III demonstrates our system

TABLE II CONFUSION MATRIX

ML Classifier	TP	TN	FP	FN
K Means Clustering	30.8	30.8	14.6	23.1
K Nearest Neighbors	31.2	39.5	14.1	15.2
Naive Bayes	33.1	43.5	11.1	12.3
K Means Clusteringwith Naive Bayes	35.1	47	7.7	10.2
Logistic Regression	37.7	47.9	6.9	7.5
Fuzzy K Nearest Neighbors	39.8	52.2	2.6	5.4
Neural Network	43.9	54.5	0.2	1.4

In order to predict heart disease and then analyze the best method of analysis for treatment for the disease the works conducted a study of different algorithms, such as the Naive Bayes, the K Nearest neighbors, and logistic regression. Alongside simple algorithms, the hybrid algorithms generated are better evaluated and implemented than several previously cited research papers. From test results and evaluates, it is inferred that, compared to the success of K Means Clustering, KNN, logistic regression, etc., Neural Network and Fuzzy KNN appear to be better performance

IV. ACKNOWLEDGMENT

We extend our heartfelt gratitude to all those who have contributed to the successful completion of our research paper titled "Heart Disease Detection using ML."

This project has been an endeavor of dedication, collaboration, and persistent effort, and we are indebted to numerous individuals and organizations for their invaluable support.

First and foremost, we express our deepest appreciation to our research supervisor, MR K.S Pathak, whose guidance and mentorship have been instrumental throughout the entire research process. Their insightful feedback, constructive criticism, and unwavering encouragement have significantly enhanced the quality and depth of our work.

We are also thankful to the members of our research team who have dedicated their time, expertise, and collaborative spirit to ensure the success of this project. Each member's unique contribution has played a crucial role in shaping the methodologies, conducting experiments,and analyzing results.

Our sincere thanks go to Babu Banarasi Das Northern India Institute of Technology, for providing the necessary resources, infrastructure, and financial support that enabled us to carry out this research. The conducive research environment and access to cutting-edge technologies have greatly facilitated our exploration into the field of machine learning for heart disease detection.

We would like to express our gratitude to the participants who willingly volunteered their health data for this study. Their contributions have been essential in building a robust and reliable machine learning model for heart disease detection, and their willingness to sharepersonal information for the advancement of medical research is truly commendable.

Finally, we thank our friends, families, and colleagues for their unwavering support, understanding, and encouragement throughout this research journey. Their patience and belief in our abilities have been a constant source of motivation.

V. CONCLUSION

This research paper represents a collective effort, and we are grateful to everyone who hascontributed in various capacities. The knowledge gained from this study has the potential to make a meaningful impact on the early detection and prevention of heart diseases, and we look forward to further advancements in this crucial area of healthcare.

Thank you all for being an integral part of thisresearch endeavor.

REFERENCES

[1] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, *et al.*, "Forecasting the future of cardiovascular disease in the united states: a policy statement from the american heart association," *Circulation*, vol. 123, no. 8, pp. 933–944, 2011.

[2] S. S. Yadav and S. M. Jadhav, "Machine learningalgorithms for disease prediction using iot environment," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6, pp. 4303–4307., 2019.

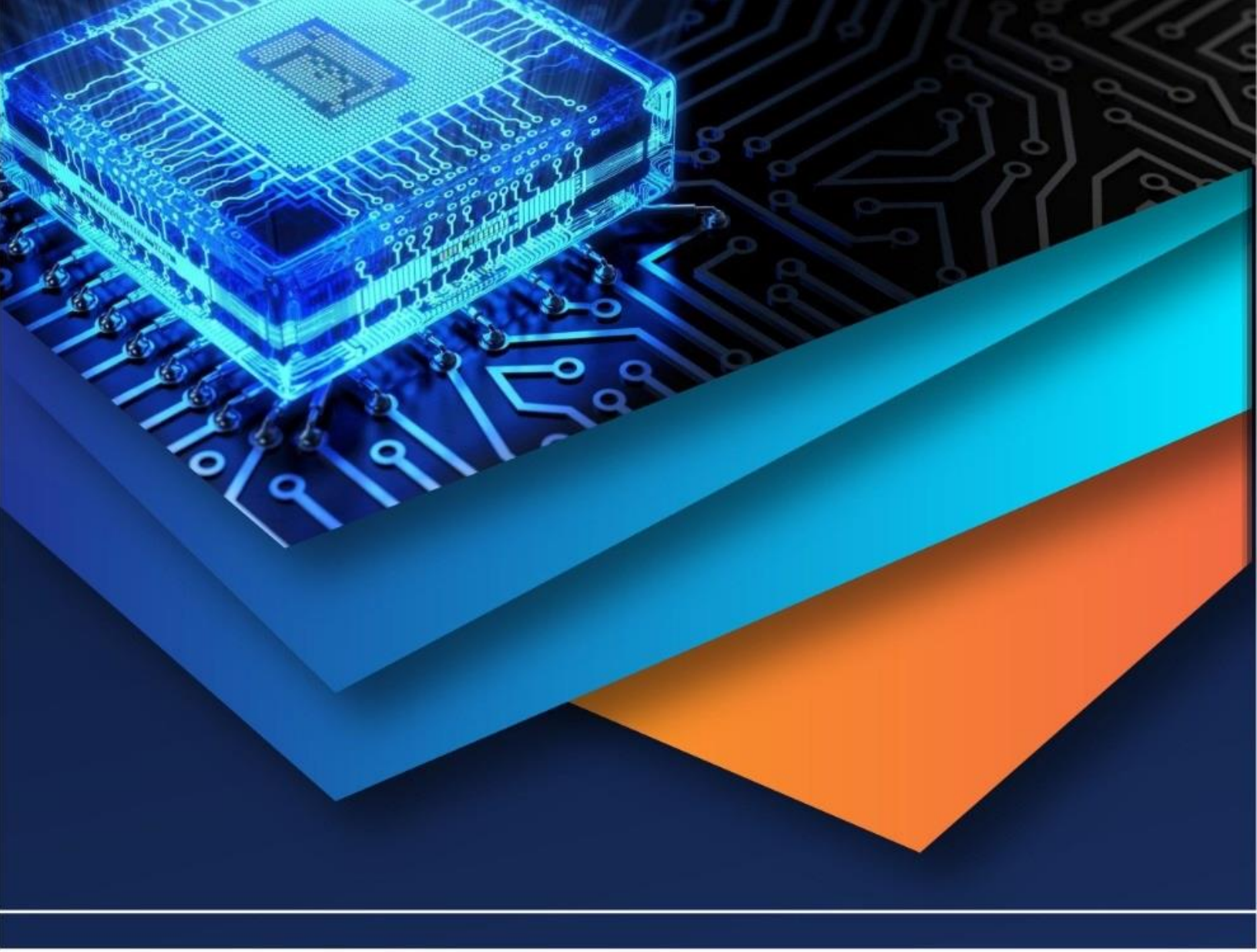
[3] S. Ghwanmeh, A. Mohammad, and A. Al- Ibrahim, "Innovative artificial neural networks- based decision support system for heart diseasesdiagnosis," *Journal of Intelligent LearningSystems and Applications*, vol. 5, no. 3, p. 176, 2013.

[4] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Systems with Applications*, vol. 40, no. 1, pp. 96– 104, 2013.

[5] 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.

[6] N. Cheung, *Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering*. PhD thesis, B. Sc. Thesis, University of Queenland, 2001.

[7] S. Palaniappan and R. Awang, "Web-based heartdisease decision support system using data mining classification modeling techniques.," in *iiWAS*, pp. 157–167, 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY