iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53377

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 11 Issue V May 2023- Available at www.ijraset.com

Prediction of Heart Disease using Machine Learning Techniques

Pranshul Kaushik[1], Richa Tiwari[2], Prateek Mohan[3], Nishika Singh[4], Rajshree Singh[5], Amit Kumar[6]
[1, 2, 3, 4, 5, 6]*IMS Engineering College, Uttar Pradesh, India*

Abstract: ***Cardiovascular diseases account for approximately 30% of global deaths, with a significant number resulting from delayed diagnoses. The challenge arises when addressing this issue through data analysis. To address this, the adoption of Machine Learning Techniques has become prevalent, and it was also the chosen approach in our study. Our research paper presents a novel classification model developed after extensive stages of pre-processing.***
The prediction model incorporates diverse feature combinations and utilizes well-established classification techniques. The selection of these techniques was based on a meticulous comparison of their respective accuracy levels. Our ultimate objective is to achieve the highest level of accuracy when compared to existing models in this domain.
By leveraging the power of data analysis and employing advanced machine learning methodologies, we aim to enhance the early detection and prediction of cardiovascular diseases. This can potentially contribute to a significant reduction in mortality rates and pave the way for more effective healthcare interventions in the field of cardiovascular health.
Keywords: Heart Disease; Machine Learning; K Nearest Neighbor (K-NN), Random Forest, Multi-Layer Perceptron Neural Network(MLPNN)

I. INTRODUCTION

Heart disease is one of the leading causes of mortality worldwide, accounting for a significant number of deaths each year. Timely detection and accurate prediction of heart disease can play a crucial role in improving patient outcomes and reducing the burden on healthcare systems. In recent years, machine learning techniques have shown great promise in the field of healthcare, particularly in predicting and diagnosing cardiovascular diseases.

This research paper aims to explore the application of machine learning techniques in the prediction of heart disease. The primary objective is to develop a robust and accurate prediction model that can aid in the early identification of individuals at risk of developing heart disease. By leveraging the power of machine learning algorithms and utilizing a comprehensive dataset, we seek to enhance the accuracy and efficiency of heart disease prediction.

The paper will present an in-depth analysis of various machine-learning algorithms employed for heart disease prediction. It will explore the strengths and weaknesses of these algorithms, highlighting their performance in terms of accuracy, sensitivity, specificity, and computational efficiency. Furthermore, different feature selection and preprocessing techniques will be investigated to optimize the predictive model's performance.

The findings of this research will provide valuable insights into the potential of machine learning in predicting heart disease. By developing a reliable and accurate prediction model, healthcare professionals can make informed decisions, initiate preventive measures, and provide timely interventions to high-risk individuals. Ultimately, the aim is to improve patient outcomes, reduce healthcare costs, and contribute to the advancement of cardiovascular health research.

II. LITERATURE REVIEW

The application of artificial intelligence and machine learning algorithms has gained much popularity in recent years due to the improved accuracy and efficiency of making predictions. The importance of research in this area lies in the possibility to develop and select models with the highest accuracy and efficiency. Hybrid models which integrate different machine learning models with information systems (major factors) are a promising approach for disease prediction[1].

Senthil Kumar Mohan and Gautam Srivastava in their prediction model proposed a hybrid method HRFLM approach by combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The prediction models are developed using 13 features and the accuracy is calculated for modeling techniques. The results show that RF and LM are the best.

The RF error rate for dataset 4 is high (20.9%) compared to the other datasets. The LM method for the dataset is the best (9.1%) compared to DT and RF methods. They combined the RF method with LM and proposed HRFLM method to improve the results [1].

Sami Azam in his research demonstrates that the Relief feature selection algorithm can provide a tightly correlated feature set which then can be used with several machine learning algorithms. The study has also identified that RFBM works particularly well with the high impact features (obtained by feature selection algorithms or medical literature) and produces accuracy substantially higher than related work. RFBM achieved accuracy of 99.05% with 10 features. Considering 13 features, the most accurate prediction 89.07% was obtained from the AB Classifier, whereas the accuracy of KNN was 83.61% [2].

Muhammad Syafrudin and Jongtae Rhee proposed an effective heart disease prediction model (HDPM) for heart disease diagnosis by integrating DBSCAN, SMOTE-ENN, and XGBoost-based MLA to improve prediction accuracy. The DBSCAN was applied to detect and remove the outlier data, SMOTE-ENN was used to balance the unbalanced training dataset and XGBoost MLA was adopted to learn and generate the prediction model. The experimental results confirmed that the proposed model achieved better performance than that of state-of-the-art models and previous study results, by achieving accuracy up to 95.90% and 98.40% for Stat log and Cleveland datasets, respectively[3].(look into Table 3 and Table 4 for accuracies)

The proposed HDPM was then loaded to diagnose the patients' current heart disease status, which was later sent back to the HDCDSS's diagnosis result interface. Thus, the developed HDCDSS helped clinicians to diagnose patients and improve heart disease clinical decision making effectively and efficiently[3].

In the proposed work, a machine intelligence framework MIFH is presented for heart disease diagnosis. The proposed framework MIFH can be used to predict the instances either as normal subjects or heart patients. MIFH utilizes the characteristics of FAMD to extract as well as derive features from the UCI heart disease Cleveland dataset and train the machine learning predictive models for classification of instances as well as prediction of heart disease and normal subjects. MIFH returns the best classifier based upon the weight matrix corresponding to performance metrics[4]

The proposed framework, i.e., MIFH, inputs the UCI Cleveland CHD dataset D, imputed the dataset for missing values ca and thal using majority labels as presented in Section V-A. The imputed Cleveland dataset is partitioned into training and validation datasets, i.e., DT and DV , respectively using the hold-out validation scheme with validation ratio 0.2. Stratification is performed to keep the partitioning balanced for heart patient and normal subject instances in both datasets, DT and DV[4].

III. PROPOSED METHODS

A. Dimensionality Reduction Techniques

1) Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a powerful dimensionality reduction technique that allows for the identification of correlations and patterns within a dataset. By transforming the data into a lower-dimensional representation, PCA retains crucial information while simplifying its structure. This process facilitates visualization of the data in 2D or 3D plots and helps in identifying the most significant features. PCA contributes to improved performance at a minimal cost to model accuracy, offering benefits such as noise reduction, feature selection, and the generation of independent and uncorrelated features. As an unsupervised learning algorithm, PCA enables efficient data exploration and analysis.

2) Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a commonly employed technique in machine learning models for predicting heart disease. It aims to find a linear combination of features that maximizes the separation between different classes of heart disease. By reducing the dimensionality of the data while preserving class separability, LDA can effectively identify the most discriminative features for accurate prediction. LDA helps uncover underlying patterns and relationships within the data, contributing to improved classification performance and aiding in the early detection and management of heart disease.

3) Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a statistical method widely used in machine learning models to predict heart disease. By estimating the parameters of a probability distribution that best fit the observed data, MLE allows for the identification of the most likely values for the model's parameters. In the context of heart disease prediction, MLE enables the model to learn and make predictions based on the probability distribution that maximizes the likelihood of the observed data, enhancing the accuracy and reliability of the predictions.

B. Classification

For this project, we employed five different classifiers to determine whether a patient is afflicted with heart disease or not. These classifiers are machine learning algorithms that analyze and learn from the provided data to make accurate predictions. By utilizing multiple classifiers, we aimed to explore their individual strengths and weaknesses in identifying heart disease cases. The use of diverse classifiers enables a comprehensive evaluation of their performance, ultimately enhancing the accuracy and robustness of our predictions.

1) Support Vector Classifier

The Support Vector Classifier (SVC) using a decision boundary, SVC separates different classes of heart disease cases based on features extracted from the dataset. SVC is particularly effective in handling complex, non-linear relationships within the data. It maximizes the margin between different classes, reducing the risk of misclassification. With its ability to handle high-dimensional data, SVC serves as a valuable tool in accurately predicting heart disease and aiding in timely interventions.

2) Random Forest Classifier

It combines multiple decision trees to create a robust ensemble model. By aggregating the predictions of individual trees, it improves accuracy and handles overfitting. The Random Forest Classifier is effective in capturing complex relationships and identifying important features, making it a valuable tool in heart disease prediction.

3) Multi-Layer Perceptron Neural Network (MLPNN)

MLPNN consists of multiple layers of interconnected neurons, allowing for complex pattern recognition and nonlinear relationships in the data. By training the MLPNN on a large dataset of heart disease cases, it can learn to accurately classify and predict the presence or absence of the disease. The MLPNN's ability to handle high-dimensional input and its flexibility in capturing intricate relationships make it a valuable tool in developing robust heart disease prediction models.

4) Decision Tree

It constructs a tree-like model where each internal node represents a feature, and each leaf node represents a class label. By recursively partitioning the data based on feature values, decision trees can capture complex relationships between variables. They offer interpretability, as the resulting tree structure can be easily understood and visualized. Decision Trees are effective in handling both categorical and numerical features, making them suitable for heart disease prediction tasks where various risk factors need to be considered.

5) k-nearest neighbors algorithm(KNN)

KNN classifies an unknown sample by identifying its k nearest neighbors based on a similarity metric. In the context of heart disease prediction, KNN analyzes patient data and compares it to the nearest neighbors to determine the likelihood of heart disease. This algorithm provides a straightforward and effective approach to classify and predict heart disease outcomes.

IV. TOOLS AND TECHNOLOGIES

A. Programming Language Used- Python

Python is a widely adopted programming language for machine learning projects. Its extensive libraries such as scikit-learn, TensorFlow, and PyTorch provide powerful tools for developing and deploying machine learning models. Python's simplicity and readability make it accessible for both beginners and experienced developers.

Its vast ecosystem offers a wide range of resources, documentation, and community support, making Python an ideal choice for machine learning projects.

B. Integrated Development Environment (IDE) Used- Spyder

Spyder is a popular integrated development environment (IDE) for machine learning projects. It provides a user-friendly interface, efficient code editing features, and integrated tools for data exploration and analysis.

Spyder's interactive console and debugging capabilities make it suitable for iterative development and testing of machine learning models.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

C. Libraries Used

1) Pandas

The pandas library is a powerful tool for data manipulation and analysis in Python. It provides data structures like DataFrames and Series, enabling efficient handling of structured data. Pandas offers a wide range of functions for data cleaning, transformation, and exploration, making it indispensable for data-intensive tasks in machine learning.

2) Matplotlib

Matplotlib is a widely-used plotting library in Python for creating visualizations. It offers a range of plot types, customization options, and high-quality output. With Matplotlib, developers can generate various charts, plots, and graphs to analyze and present data in machine learning projects.

3) Seaborn

Seaborn is a Python library that enhances data visualization in statistical graphics. It provides a high-level interface for creating visually appealing and informative plots. Seaborn simplifies the process of creating complex visualizations, making it a valuable tool for exploratory data analysis and communicating results in machine learning projects.

4) Numpy

NumPy is a fundamental library for scientific computing in Python. It provides powerful tools for working with large, multi-dimensional arrays and matrices. NumPy offers efficient mathematical functions, linear algebra operations, and tools for data manipulation, making it essential for machine learning tasks.

5) Sklearn

Scikit-learn is a comprehensive machine-learning library in Python. It offers a wide range of algorithms and tools for classification, regression, clustering, dimensionality reduction, model selection, and evaluation, making it a go-to choice for building and deploying machine learning models.

6) Scipy

The scipy library in Python is a powerful tool for scientific and numerical computing. It offers a wide range of functions and modules for optimization, interpolation, linear algebra, signal processing, statistics, and more. Scipy is extensively used in various fields, including machine learning and data analysis.

V. EXPERIMENTAL SETUP

The initial stage of the setup involves obtaining the dataset containing the features of an individual afflicted with heart disease and an individual without it, along with the corresponding disease status. The dataset utilized in this experiment is sourced from Kaggle, a website (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset). Python serves as the programming language for conducting the experiment. Thirteen attributes, accessible within the dataset, are utilized, and their descriptions can be found on Kaggle. Subsequently, the data is subjected to analysis. To obtain a concise summary of the DataFrame, the info() function from the Pandas library is employed on the dataset.

```
In [3]: runcell(0, 'C:/Users/prans/OneDrive/aaa VIII sem/Project/code/final_project.py')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 302 entries, 0 to 301
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       302 non-null    int64
 1   sex       302 non-null    int64
 2   cp        302 non-null    int64
 3   trestbps  302 non-null    int64
 4   chol      302 non-null    int64
 5   fbs       302 non-null    int64
 6   restecg   302 non-null    int64
 7   thalach   302 non-null    int64
 8   exang     302 non-null    int64
 9   oldpeak   302 non-null    float64
 10  slope     302 non-null    int64
 11  ca        302 non-null    int64
 12  thal      302 non-null    int64
 13  target    302 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.2 KB
```

Figure 1. Concise summary of the DataFrame

The describe() function provided by the Pandas library is used to retrieve some statistical information of the data set like mean of the values of the attributes used.

```
In [8]: df.describe()
Out[8]:
              age         sex          cp  ...          ca        thal      target
count  302.00000  302.000000  302.000000  ...  302.000000  302.000000  302.000000
mean    54.42053    0.682119    0.963576  ...    0.718543    2.314570    0.543046
std      9.04797    0.466426    1.032044  ...    1.006748    0.613026    0.498970
min     29.00000    0.000000    0.000000  ...    0.000000    0.000000    0.000000
25%     48.00000    0.000000    0.000000  ...    0.000000    2.000000    0.000000
50%     55.50000    1.000000    1.000000  ...    0.000000    2.000000    1.000000
75%     61.00000    1.000000    2.000000  ...    1.000000    3.000000    1.000000
max     77.00000    1.000000    3.000000  ...    4.000000    3.000000    1.000000

[8 rows x 14 columns]
```

Figure 2. describe() function in pandas

Here, we consider people having age from 29 to 35 to be young, 36 to 50 to be middle and more than 50 are elderly.

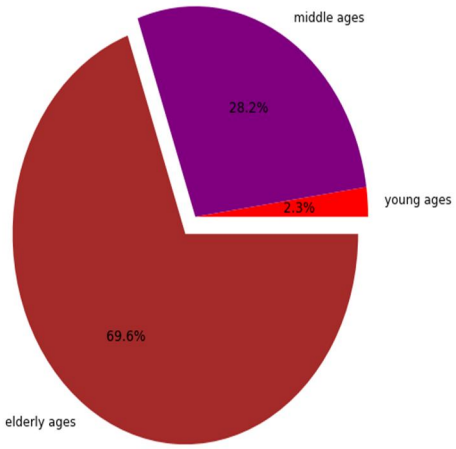


Figure 3. Pie Plot of Age analysis

The target attribute has 2 values 0 and 1. Here 0 means no disease and 1 means disease.

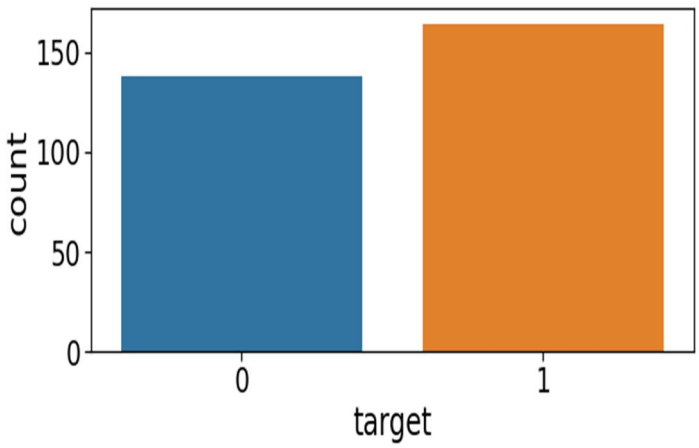


Figure 4. Count plot of target attribute

After checking that the data is balanced, the correlation between the data is found out and is plotted as a heat map using the Seaborn library.

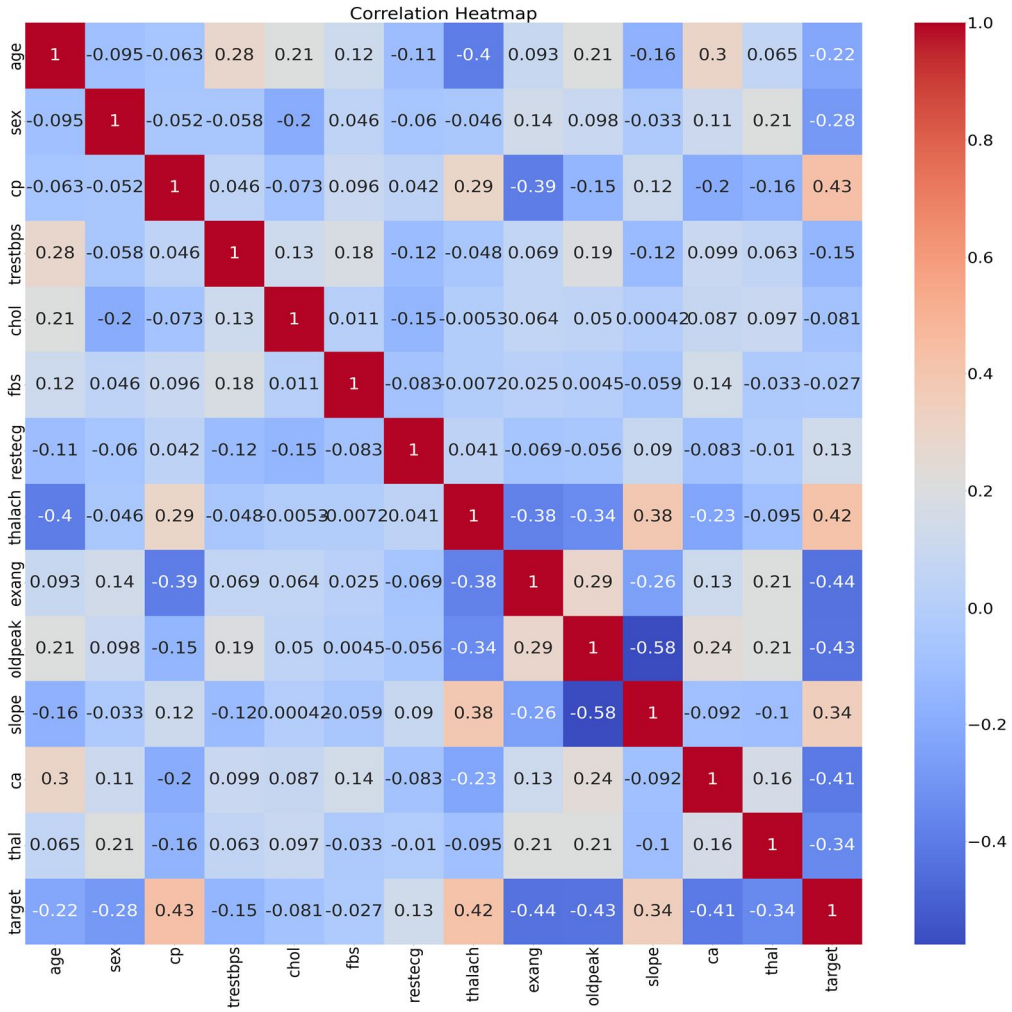


Figure 5. Correlation between variables

The heat map unmistakably demonstrates that attributes such as cp (chest pain) and thalack (maximum heart rate achieved) exhibit a positive correlation with the target attribute. Since we have assessed the correlation, the next step involves transforming categorical variables like sex, cp, fbs, restecg, exang, slope, ca, and thal into dummy variables. This can be accomplished using the get_dummies method from the Pandas library. After dummy variable creation, standard scaling is necessary for columns like age, trestbps, chol, thalach, and oldpeak, as they possess dissimilar quantities and units. To achieve this, the Scikit-learn library in Python can be employed.

The data set has been divided into two parts, training data which is 75% of the whole data set and testing data which is 25% of the whole data set. After preparing the data, the algorithms are applied and the confusion matrix has been found out. The results have been found in terms of accuracy of the algorithm. The accuracy has been found with the use of a confusion matrix.

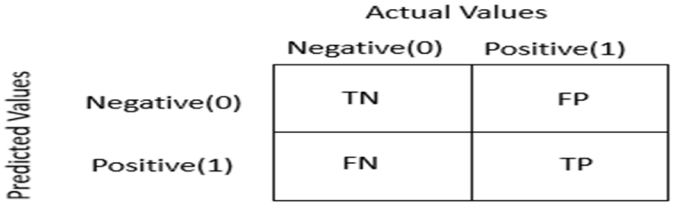


Figure 6. Confusion matrix layout

Confusion matrix can also be shown as a matrix in the following way:

[[TN FP
FN TP]]

The accuracy of the algorithm can be calculated using the formula:

Accuracy = {(TP + TN) / TP + FP + TN + FN)} * 100

VI. RESULTS

After applying the algorithms, the results obtained are as follows:

Table 1 Confusion matrices

Algorithms	Confusion Matrices			
	Without reduction technique	With PCA	With LDA	With MLE
Support Vector Classifier	[[23 11] [4 38]]	[[0 34] [0 42]]	[[24 10] [4 38]]	[[24 10] [12 30]]
Random Forest	[[23 11] [4 38]]	[[14 20] [11 31]]	[[24 10] [5 37]]	[[27 7] [12 30]]
MLPNN	[[27 7] [10 32]]	[[15 19] [11 31]]	[[27 7] [4 38]]	[[27 7] [11 31]]
Decision Tree	[[23 11] [18 24]]	[[15 19] [18 24]]	[[22 12] [5 37]]	[[20 14] [11 31]]
KNN	[[16 18] [13 29]]	[[21 13] [19 23]]	[[25 9] [7 35]]	[[26 8] [14 28]]

Table 2. Accuracies after applying each algorithm

Algorithms	Accuracy			
	Without reduction technique	With PCA	With LDA	With MLE
Support Vector Classifier	80.26%	55.26%	81.58%	71.05%
Random Forest	80.26%	80.26%	80.26%	75.00%
MLPNN	77.63%	60.52%	85.53%	76.31%
Decision Tree	61.84%	51.32%	77.63%	67.11%
KNN	59.21%	57.89%	78.95%	71.05%

Table 1 shows the confusion matrix for each combination of algoritm and Table 2 shows the accuracy of each algorithm after testing the machine learning model on the dataset.

VII. CONCLUSION

In this project, we have used total 20 combinations of three dimensionality reduction techniques and five classificaition techniques to predict the heart disease.We proposed an effective heart disease prediction model for heart disease diagnosis using machine learning techniques with accuracy 85.53 using MLPNN with LDA.

VIII. FUTURE SCOPE

In the future, we will consider the comparison of other data sampling with the model hyper-parameters and broader medical datasets.We will work on more algorithms for increasing our accuracy.

IX. CODE

The code for the entire project is hosted on github and linked below: https://github.com/pranshul2199/ML-InsightsHub

X. ACKNOWLEDGMENT

We are really thankful to Assistant Professor Mr. Amit Kumar from the IMS Engineering College in Ghaziabad's Computer Science and Engineering department for his assistance in assisting us with the application of our research to the real world. Its our privilege to express our sincere regards to our project guide, Mr. Amit Kumar for his valuable inputs, able guidance, encouragement, cooperation and constructive criticism throughout the duration of our project. We sincerely thank the Project Assessment Committee members for their support and for enabling us to present the project on the topic “PREDICTION OF HEART DISEASE USING MACHINE LEARNING TECHNIQUES”.

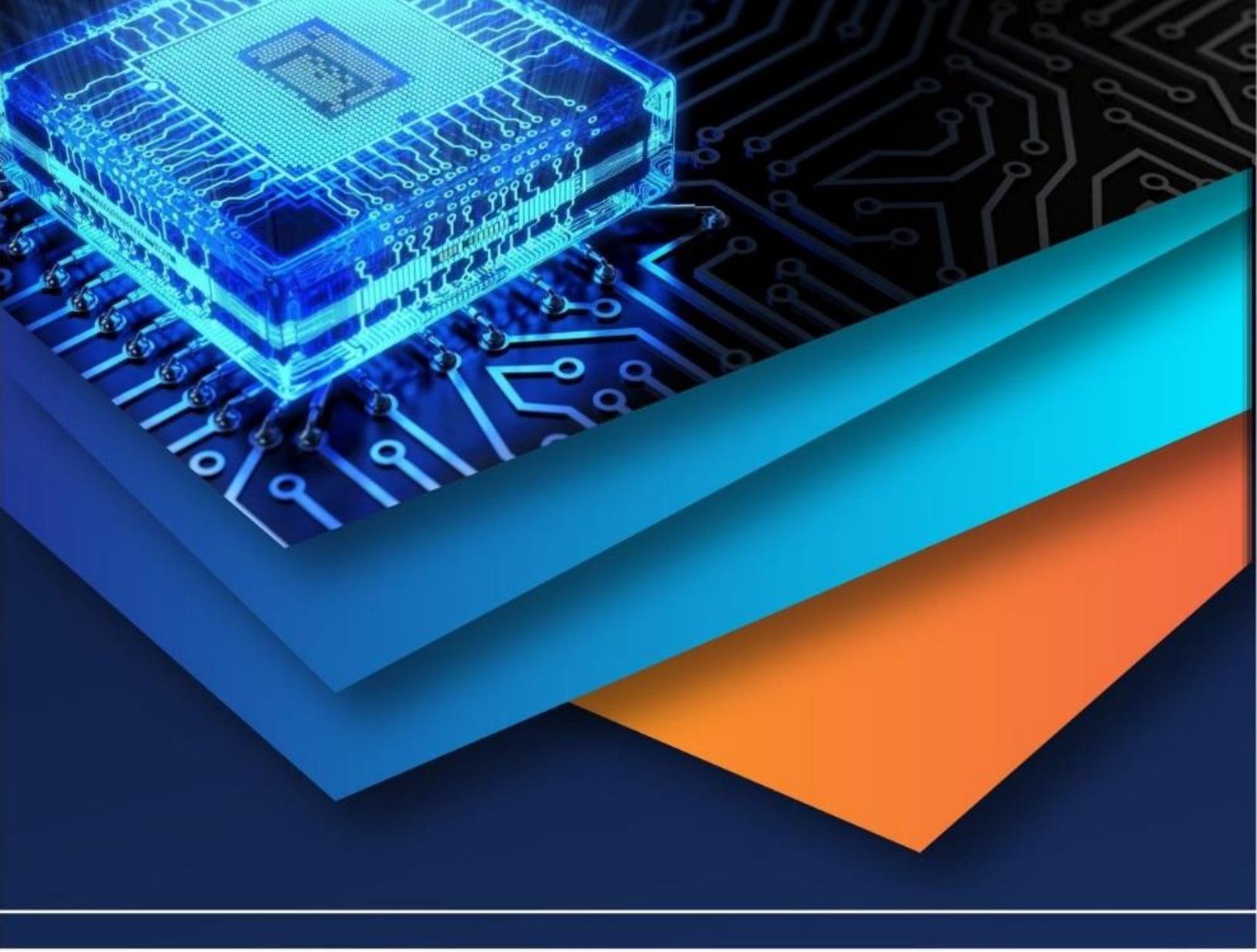
REFERENCES

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.

[2] P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in IEEE Access, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[3] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[4] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, ‘‘MIFH: A machine intelligence framework for heart disease diagnosis,’’ IEEE Access, vol. 8, pp. 14659–14674, 2020Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Intelligent and effective heart disease prediction system using weighted associative classifiers. International Journal on Computer Science and Engineering, 3(6), 2385-2392.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY