iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.49209

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 IV Apr 2023- Available at www.ijraset.com

Diagnosis of Cardiovascular Disease using Deep Learning

Vasamsetti Ramya Sri Sushma[1], K Nikhil[2], K Ashish Reddy[3], L Sunitha[4]
[1, 2, 3]*Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad*
[4]*Assistant Professor, Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad*

Abstract: ***Cardiovascular disease is one of the most horrendous illnesses, particularly the silent heart attack that strikes a person so abruptly that there is no time for treatment. It's difficult to diagnose a disease of this nature. One of the scariest diseases that c an kill a person at any time without warning is heart disease, and most doctors are unable to predict silent heart attacks. The lac k of specialists and an increase in cases of wrong diagnoses have fueled the demand for the creation of an efficient cardiovascul ar disease prediction system. This resulted in the exploration and development of original machine learning and medical data mi ning methodologies. The principal goal of this research is to identify the most crucial qualities for silent heart attack identificatio n by using classification algorithms to extract significant patterns and features from medical data. Although it is not innovative t o build such a system, the current ones have flaws and are not designed to detect the likelihood of silent heart attacks. Another is sue with the present heart attack prediction method is the use of characteristics. Choosing the typical features for the heart attac k prediction algorithm frequently yields unreliable results. To increase prediction accuracy, the suggested method aims to extract suitable attributes from the datasets. We developed a framework in this exploration that can understand the principles of predicti ng the risk profile of patients with the clinical data parameters. This research suggests an effective neural network with convolut ional layers to classify clinical data that is noticeably class-imbalanced. In order to forecast the development of Coronary Heart Disease, data from the National Health and Nutritional Examination Survey (NHANES) is collected (CHD). This research aime d to design a robust deep-learning algorithm to predict heart disease. Heart disease prediction is performed using SMOTE and MLP Classifier algorithms and Deep Neural Network Algorithms. The effectiveness of the model that accurately predicts the pre sence or absence of heart disease was examined using DNN and ANN. In this research article, we'll look at a machine-learnin g model that can clearly assess cardiac issues and be utilized by analysts and medical professionals.***
Index Terms: cardiovascular disease, machine learning, MLP Classifier, Deep Neural Network, SMOTE

I. INTRODUCTION

The terminology "cardiovascular diseases" (CVDs) encompasses a variety of heart and blood vessel abnormalities. This would include cerebrovascular disease, a disorder of the blood arteries feeding the brain, and coronary heart disease, a disorder of the blood vessels supplying the heart muscle. a condition called peripheral arterial disease that affects the blood arteries supplying the arms and legs; Streptococcal bacteria, which further cause rheumatic fever, which harms the heart muscle and heart valves, are the cause of rheumatic heart disease; Blood clots in the leg veins known as deep vein thrombosis and pulmonary embolism have the potential to break free and go to the heart and lungs. Congenital heart disease is a congenital defect that interferes with the heart's normal growth and operation. Worldwide, cardiovascular diseases (CVDs) constitute the leading cause of death.

In 2019, 17.9 million people are estimated to die from CVDs, accounting for 32% of all deaths globally. Heart attacks and strokes were responsible for 85% of these fatalities. Around 75% of CVD fatalities occur in developing and middle-income countries. In 2019, noncommunicable illnesses caused 17 million unexpected deaths in those under 70, with cardiovascular diseases (CVDs) accounting for 38% of those deaths. By addressing behavioral risk factors like cigarette use, poor eating habits, obesity, physical inactivity, and reckless alcohol consumption, the majority of cardiovascular illnesses can be cured. The cardiovascular disease must be identified as soon as feasible in order to begin treatment with counseling and medication. Machine learning is a branch of computer science and artificial intelligence (AI) that focuses on simulating human learning by using data and algorithms to improve the system's accuracy over time. A more prominent machine learning technique is deep learning. It is used for more than only picture classification jobs; it also uses regular tabular data. We develop a deep learning neural network model for this model. With the Keras library, a deep learning neural network toolkit, we can apply the Talos optimizer in this model. A high-level neural network model is produced by Keras. It was created to make experimentation simple and quick. Convolutional neural networks (CNN) and recurrent neural networks (RNN), as well as hybrids of the two, are also supported by Keras. Both the CPU and GPU

operate well. Many different methods and strategies have been used in this sector to conduct a significant quantity of studies. This study attempts to improve the system's efficiency and accuracy in predicting the likelihood of a heart attack.

II. LITERATURE REVIEW

1) In 2016, BayuAdhi Tama et al. recommended a study on the chronic condition known as diabetes. This illness was thought to be incredibly prevalent and to be the primary cause. According to a poll by the International Diabetes Federation (IDF) [8], there are around 285 million diabetics worldwide. Due to the lack of an effective method for the absolute minimization and prevention of this disease, these numbers may rise in the near future. Type 2 diabetes is the most prevalent form of the disease. The discovery of TTD posed the biggest challenge because it was difficult to foresee all of its effects. Data mining was used as a result because it produced the best results and assisted in the finding of information from readily available data. Throughout the data mining process, the Support Vector Machine (SVM) classifier was utilized to extract useful data about all patients from the previous records. The ability to recognize TTD in a timely manner aided in the making of effective decisions. U-Net was employed by Hajar Cherguif et al. to semantically segment medical images. A decent convoluted 2D segmentation network was created using U-Net architecture. The suggested model was tested and assessed using the BRATS 2017 dataset. 27 convolutional layers, 4 deconvolutional layers, and a Dice coef of 0.81 were incorporated into the proposed U-Net design.
2) According to Shahriar Satu et al., cardiovascular disease is a widespread illness that causes enormous loss of life throughout the world. The creators have thought of a few outlandish strategies to deal with learning crucial components of heart disease. They used data on Cleveland and Hungarian heart disease, which were divided into three categories: 33%, 65%, and 100%. Estimates of the numerous individual credit scopes in this data are used to identify key components of this illness. The knowledge on heart illness is then examined using a variety of semi-directed learning algorithms, including Collective Wrapper, Filtered Collective, and Another Semi-Supervised Idea. In order to support singular classifiers and identify the optimum semi-managed learning computation, measurements of these classifiers such as precision, f-measure, and region under the ROC have been determined to be useful. By removing credits in a steady progression consecutively and observing the outcomes of the order, this calculation has examined significant and minor causes of heart disease. Results from exploratory analyses on two reliable pieces of data demonstrate the validity and competence of the assessment.
3) Dinesh Kumar conducted research on early methods for predicting cardiovascular diseases and assisted in making decisions about the progressions that should have occurred in high-chance patients, which reduced their risks. This forecast's data pre-processing makes use of methods like removing noise from the data, eliminating missing data, changing default values when appropriate, and combining attributes for prediction at different levels. In order to demonstrate an accurate model for predicting cardiovascular diseases, these progressions are concluded by contrasting the correctness of applying rules with individual outcomes of classifiers such as gradient boosting, random forest, naive Bayes, support vector machines, and logistic regression onto the dataset taken from a district.
4) In their research proposal for 2020, Haviluddin et al. compared the performance of Euclidean, Manhattan, and Minkowski distances in K-Means clustering. The Euclidean Distance, Manhattan Distance, and Minkowski Distance methods have all been used to analyze the data distance calculation to each centroid in the K-Means approach. In the Bontang region of East Kalimantan, Indonesia, crime rates have been categorized as high, medium, and low using the K-Means approach. Based on the experiment, a value of 16.3418 was achieved for the Minkowski Distance approach, 18.3532 for the Euclidean Distance, and 29.4712 for the Manhattan Distance. Group data has been processed using the best distance approach in the interim. Results from tests using two and three clusters have been performed. where two clusters and three clusters with a value of 12.6850 are the results of the Euclidean Distance. The Manhattan Distance yielded 2 clusters and 3 clusters, both with a value of 25.2956. Two clusters and three clusters with a value of 11.3562 are the findings of the Minkowski distance. Because the results of the three clusters have a lower value than the SSE value of the two clusters, they are more ideal in this study than the findings of the two clusters. The addition of factors including the incident's location, the amount or proportion of crime in previous years, and it is crucial to provide weight to each new criminal case in order to improve the accuracy of clustering results in the next study.
5) SahayaArthyet.al analyses the prior research on data mining-based heart disease prediction. Data mining methods are frequently employed in the prediction of heart disease. Also, they talk about the databases and tools that were used, including the heart disease data set from the UCI repository and Weka, Rapid Miner, Data Melt, Apache Mahout, Rattle, KEEL, R data mining, and soon. They come to the conclusion that using a single algorithm improves prediction accuracy. Yet, using the hybridization of two or more algorithms can strengthen and improve the accuracy of heart disease prediction.

III. PROPOSED WORK

The term "deep learning" refers to a subset of machine learning and is essentially a neural network with three or more layers. Artificial neural networks try to replicate how the human brain works, but they can't come close to matching it when it comes to learning from huge amounts of data. Even though a neural network with only one layer can still approximate the predictions, more hidden layers can help tune and optimize for accuracy. The goal of this study is to create a reliable deep-learning algorithm to forecast cardiac disease.

SMOTE and MLP Classifier algorithms and a Deep Neural Network model are used to forecast heart disease. According to this study, a neural network with convolutional layers can successfully categorize clinical data that is distinctly class-imbalanced. Data from the National Health and Nutritional Examination Survey (NHANES) is collected to predict the onset of coronary heart disease (CHD).

Contrary to the majority of the currently used machine learning models that have been applied to this class of data, our neural network model exhibits resilience to the imbalance with reasonable harmony in class-specific performance and is not affected by class imbalance even after the adjustment of class-specific weights. With a severely unbalanced dataset, it gets more challenging to simultaneously obtain a high class 0 accuracy and a high class 1 accuracy as the test sample size increases. If we input high imbalance data, as the test sample size grows, it becomes increasingly difficult to simultaneously obtain a high class 1 accuracy (actual CHD prediction rate) and a high class 0 accuracy.

We adopt a two-step approach: first, we built a Multi-Layer Perceptron Classifier model and measured the metrics like accuracy, precision, recall, and f1-score. Next, we built a Deep Neural Network model with three layers. The activation functions used in this network are Relu and sigmoid along with the optimizer Adam.

A. Method

A system for predicting cardiac disease has been created, allowing users to enter patient information and have the system utilize the generated model to forecast the outcome for that specific patient. The result will be classified by the model as either normal or dangerous. We have retrained our model in this project using Keras, TensorFlow, and the applied relu function. The generated model is then used to find specific patients. The result will be classified by the model as either normal or dangerous. In this project, we retrained our model and obtained an accuracy of 88% with the aid of Keras, TensorFlow, and the relu function.

1) Dataset

Data from the National Health and Nutritional Examination Survey (NHANES) is gathered in order to forecast the occurrence of coronary heart disease (CHD). The dataset yielded 37079 records with 50 medical attributes (factors), the majority of which showed the absence of CHD. This demonstrates the dataset's asymmetry. Our study makes use of NHANES data from 1999–2000 through 2015–2016.

The dataset is produced by combining information from 37,079 (CHD - 1300, Non-CHD - 35,779) individuals' demographic, examination, laboratory, and questionnaire records. Demographic information includes the age and gender of survey participants at the time of screening. As a group of risk factor factors, body mass index (BMI), height, blood pressure, and participant weight are also considered in order to study their effect on cardiovascular disease. The study's dichotomous dependent variable is coronary heart disease (CHD). You are deemed to be aware of CHD if you responded "yes" to the question about receiving a diagnosis of coronary heart disease.

The list of attributes is gender, age, annual family income, ratio-family-income-poverty, 60 s pulse rate, systolic, diastolic, weight, height, body mass index, white blood cells, lymphocyte, monocyte, eosinophils, basophils, red blood cells, hemoglobin, mean cell volume, the mean concentration of hemoglobin, platelet count, the mean volume of platelets, neutrophils, hematocrit, red blood cell width, albumin, alkaline phosphatase (ALP), aspartate aminotransferase (AST), alanine aminotransferase (ALT), cholesterol, creatinine, glucose, gamma-glutamyl transferase (GGT), cholesterol, creatinine, glucose, iron, iron, lactate dehydrogenase (LDH), phosphorus, bilirubin, protein, uric acid, triglycerides, total cholesterol, high-density lipoprotein (HDL), glycohemoglobin, vigorous-work, moderate-work, health-Insurance, diabetes, blood-related diabetes, and blood-related stroke. However, a few of these variables are linearly dependent in terms of how they were obtained or quantified, and a few of them are uncorrelated (annual family income, height, the ratio of family income-poverty, 60 s pulse rate, health insurance, lymphocyte, monocyte, eosinophils, total cholesterol, mean cell volume, the mean concentration of hemoglobin, hematocrit, segmented neutrophils). These variables are not taken into account for further processing and analysis.

Table. 1: Description of the dependent variable and the Independent variables for the risk factor

Variable Name	Description	Code	Meaning
Gender	Gender of the participant	1	Male
		2	Female
Vigorous Activity	Vigorous activity in last one	1	Yes
12 years and above	week or 30 days	2	No
		3	Unable to
			do activity
Moderate Activity	Moderate activity in last	1	Yes
12 years and above		2	No
		3	Unable to
			do activity
Diabetes	Doctor told that the	1	Yes
1 yr and above	participant has diabetes	2	No
		3	Borderline
Blood Relative	Biological blood relatives	1	Yes
Diabetes	ever told that they have	2	No
20 yrs and above	diabetes		
Blood Relative Stroke	Biological blood relatives	1	Yes
20 yrs and above	ever told that they have	2	No
	hypertension or stroke		
	before the age of 50		
Coronary Heart Disease	Ever told that the participant	1	Yes
Gender	Gender of the participant	1.5995	Male

2) Data Pre-Processing

To make the data more useful for creating the machine learning model, we must process it (data pre-processing). An essential phase in the data mining process is data pre-processing. The Pandas library is better at handling data manipulation. We can view the dataset's statistical data using the describe the () method. Some of the tasks involved in data preprocessing are dropping insignificant columns, selecting important features, and converting categorical values into discrete values.

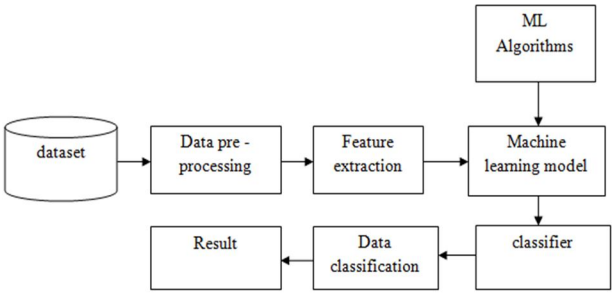


Fig. 1. Process model

3) Dealing With Imbalanced Data

Resampling the data is one of the most popular solutions for a dataset that is unbalanced. For this, under-sampling and over-sampling are the two main types of approaches. Oversampling techniques are favored over undersampling ones. The cause is that we frequently omit data items that could contain important information when we undersample. In this study, we thoroughly investigated SMOTE.

B. Synthetic Minority Oversampling Method: SMOTE

Synthetic samples are created for the minority class as part of the oversampling technique known as SMOTE. By using this method, the overfitting problem caused by random oversampling is reduced. It focuses on the feature space to create new instances by interpolating between positive instances that are close to one another.

C. Working

First, N, the total number of oversampling observations, is determined. The binary class distribution of 1:1 is frequently adopted. But that can be turned down if necessary. The first member of a positive class is chosen at random to start the iteration. The KNNs for that instance is then obtained, with a default of 5. In order to interpolate additional synthetic instances, N of these K instances is ultimately chosen. To do that, any distance metric is used to calculate the distance between the feature vector and its neighbors The preceding feature vector is now multiplied by this difference plus any random value within the range of [0, 1]. Below is a visual representation of this:

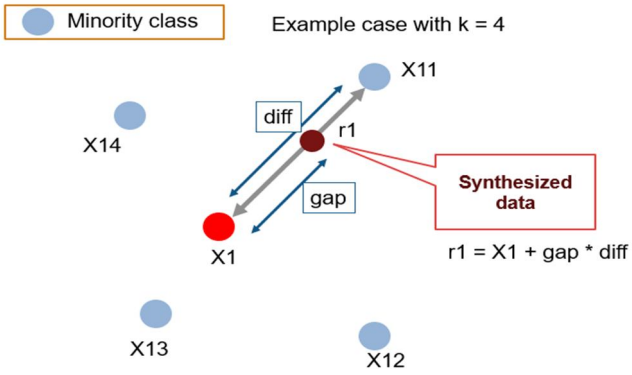


Fig. 2. SMOTE working

In our research, we are using SMOTE because most of the samples are of negative output which might affect the accuracy of the prediction of positive samples.

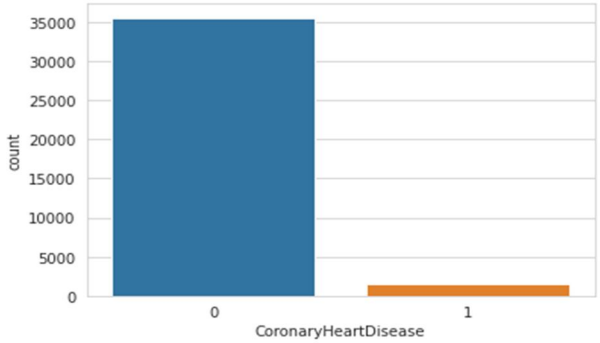


Fig. 3. Before applying SMOTE

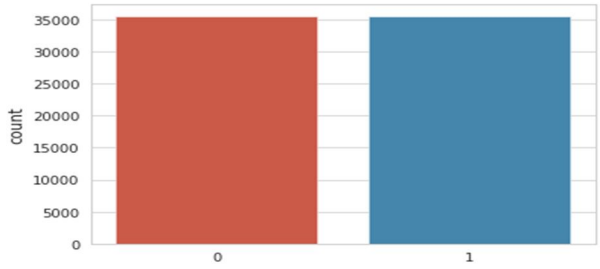


Fig. 4. After applying SMOTE

1) MLP Classifier

A typical neural network for classification applications in machine learning is the MLP Classifier or multilayer perceptron classifier. It is a feedforward artificial neural network that creates a prediction or classification using a set of inputs and one or more hidden layers of nodes. The weights and biases of the network are modified during the training of the MLP Classifier using an optimization approach, such as gradient descent, to reduce the error between the predicted class labels and the actual class labels.

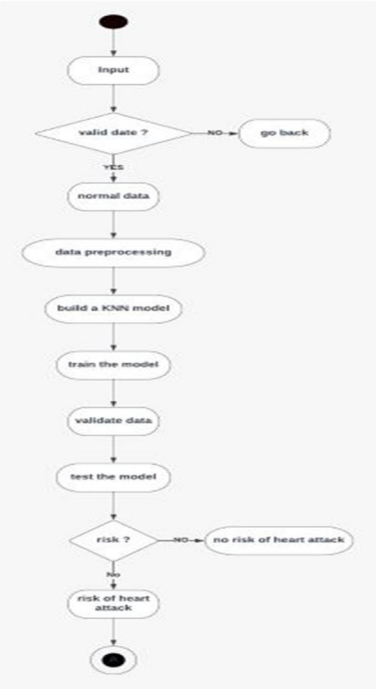


Fig. 5. Model working

In this research, we are using the inbuilt MLP Classifier model from the sci-kit learn library with maximum iterations of 300 and the random state as 1. Here, we are calculating different metrics like accuracy, f1-score, precision, recall, etc. Our model is working fine for the taken dataset with an accuracy of 90%(approx.).

2) Deep Neural Network

A type of machine learning method called deep neural networks can be used to predict a variety of outcomes, including the presence of heart disease. Deep neural networks often need a lot of data in order to predict the presence of heart disease. A deep neural network, a kind of artificial neural network that can learn complicated patterns and correlations between several variables, is then trained using the data.

In research, we have built a neural network with 3 layers using the Keras library. Here, we used different activation functions like Relu and Sigmoid. We used the Adam optimizer along with binary cross entropy to calculate the loss. We trained the model with 100 epochs and a batch size of 10. This network is giving an accuracy of 89%(approx.) and a loss of 27%(approx.).We then plotted a graph between epoch and loss or accuracy for training data.

```
model.summary()

Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_3 (Dense)             (None, 12)                600

 dense_4 (Dense)             (None, 8)                 104

 dense_5 (Dense)             (None, 1)                 9

=================================================================
Total params: 713
Trainable params: 713
Non-trainable params: 0
```

Fig. 6. Neural Network model summary

IV. RESULTS

By assigning 80% of data for training and 20% of data for testing against an MLP Classifier model with maximum iterations of 300, we found that our model is working with an accuracy of 90.01% (approx.). We made a classification report on this model where we calculated distinct metrics like precision, recall, accuracy, etc.

Table. 2: Classification Report of MLP Classifier

Risk of Heart disease	precision	recall	f1-score
No	0.89	0.91	0.9
Yes	0.91	0.89	0.9

For the Deep Neural Network model that we have built, we found that our model is giving an accuracy of 88.66% (approx.) and a loss of 27.34%(approx.).

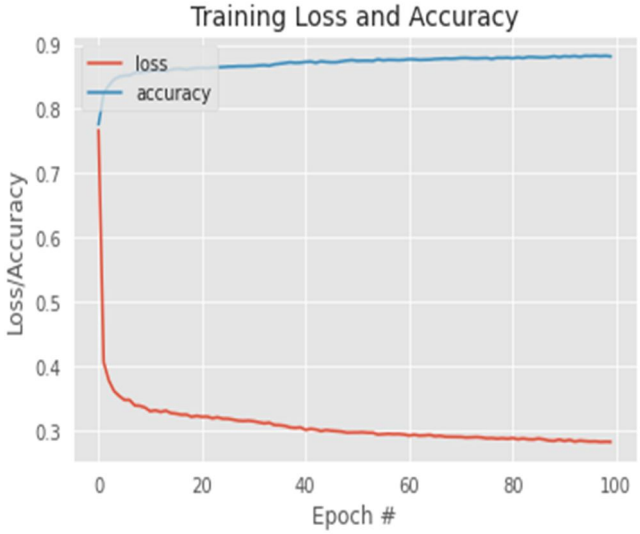


Fig. 7. Epoch Vs Loss/Accuracy for Deep Neural Network Model

Model	Accuarcy(in %)
MLP Classifier	90.01
Deep Neural Network	88.66

Table. 3. Accuracy

V. CONCLUSION

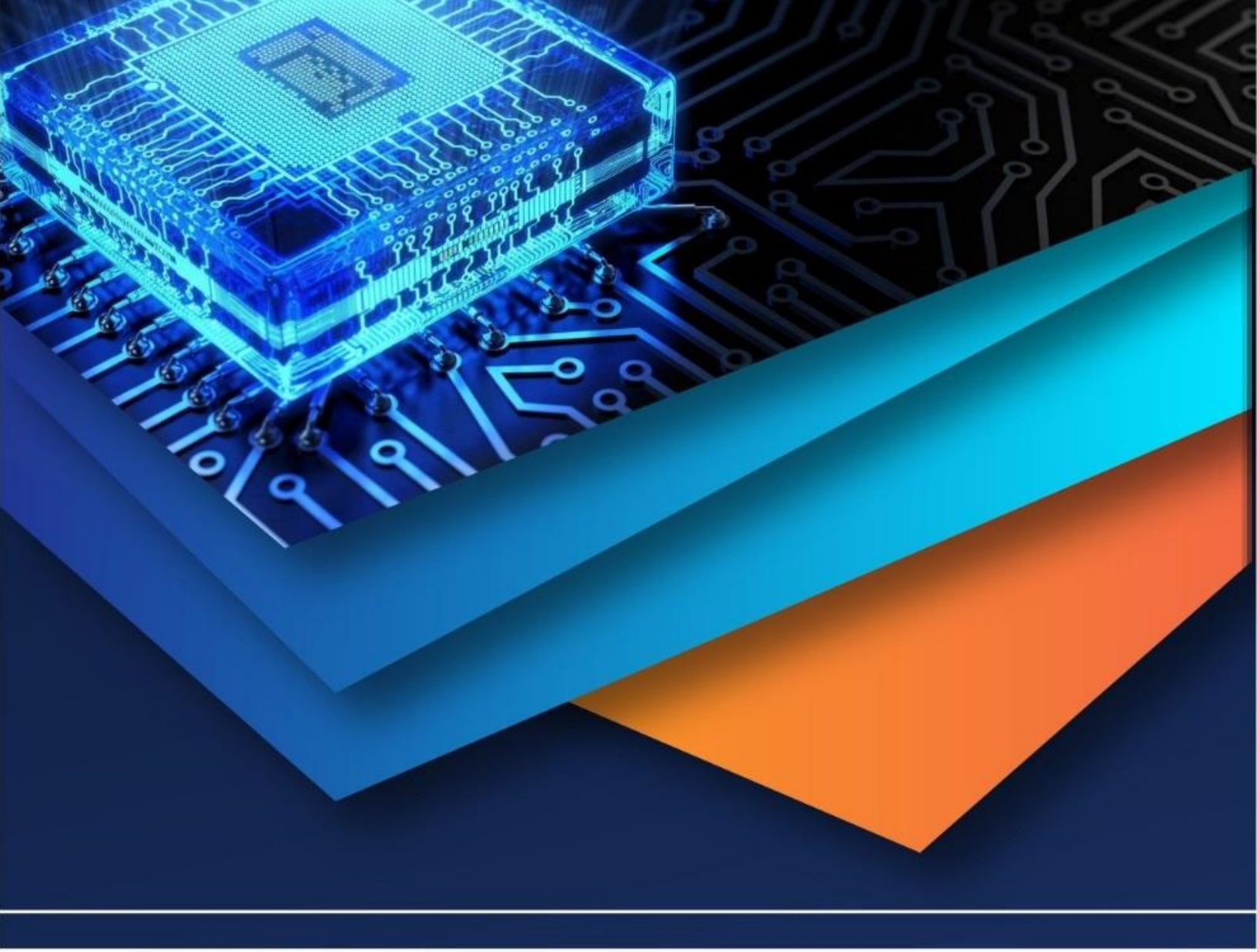
In this research, we worked with the prediction of cardiovascular disease using the NHANENS dataset which is completely imbalanced. Here, we worked this data by handling it with an over-sampling algorithm like SMOTE. The data is split into training data and testing data. We found that using algorithms like MLP Classifier and building a deep neural network model it is possible to predict the risk of heart disease even for the imbalanced data with an accuracy of nearly 90%. Our project is easy to implement and predicts the heart attack risk of a patient very accurately.

VI. FUTURE SCOPE

Since Deep Learning is one of the fastest-growing filed, we want to extend our research by experimenting with different RNN architectures based on different activation functions, and optimizers in order to improve accuracy. To improve accuracy and better prediction of the risk associated with heart disease, we will continue to train our model by using real-time data rather than NHANENS dataset by handling missing values, detecting outliers, etc.

REFERENCES

[1] S. Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning", Second International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018), pp. 1275–1278, 2018.
[2] R. Abhay Kishore, Ajay Kumar, Karan Singh, Yogita Hambir and Maninder Punia, "Heart Attack Prediction Using Deep Learning," in 2018 International Research Journal of Engineering and Technology(IRJET), vol. 5,pp. 4420–4423,2018.
[3] Z. Praneetha M, Sri Varsha M, Jesudoss A, and Albert Mayan , "Cardiovascular Disorder Prediction using Machine Learning" e Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021), pp. 1665–1670, 2021.
[4] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded, and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-5.
[5] A. Mohini Chakarverti, Saumya Yadav, Rajiv Rajan, "Classification Technique for Heart Disease Prediction in Data Mining," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), pp. 1578–1582, 2019.
[6] Mr. ChalaBeyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.
[7] N. Komal Kumar, G.Sarika Sindhu, D.Krishna Prashanthi , A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 15–21,2020.
[8] Ms. Tejaswini U. Mane, "Smart heart disease prediction system using Improved K-Means and ID3 on Big Data", 2017 International Conference on Data Management, Analytics, and Innovation (ICDMAI), vol. 8, issue 11, pp. 123-148, 2017.
[9] SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", vol. 5, issue 1, pp. 13-28, 2008.
[10] KanikaPahwa, Ravinder Kumar, "Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer, and Electronics (UPCON), vol. 4, issue 5, pp. 23-48, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY