Journal of Engineering Sciences

Vol11,Issue2,FEB/2020
ISSN NO:0377-9254

INCREASING ACCURACY OF PREDICTING HEART STROKE USING EFFECTIVE MACHINE LEARNING TECHNIQUES BY APPLYING PRINCIPAL COMPONENT ANALYSIS (PCA)

P. JAGADISH KUMAR1, DR.P.MEENA KUMARI[2], PUPPALA PRIYANKA[3]

1. Research Scholar, BIHER, Chennai, Tamil Nadu 600073

2. Asso. Professor, Samskruti College of Engg & Tech., Hyd, TS

3. JNTUH, Hyderabad, India.

Abstract:-In medical science, heart stroke is one of the main challenges; This is because many parameters and technical factors are required to accurately predict the disease. The main idea behind this discussion is to highlight how useful machine learning approaches are for assessing heart stroke using medical data. (Detrano, R., et.al.1989). This research emphasizes overcoming the vulnerability and creating a computer system so that misinterpretations and misinterpretations of data by heart consultants cannot be avoided. Machine learning is a good choice for high accuracy assessment of not only heart stroke but also other diseases. This is because this different tool uses a functional vector and its different data types under different conditions for predicting heart stroke, algorithms such as Naive Bayes, Decision Tree, and Neural Network. Classified reports of heart stroke The Decision Tree is used to deliver, but the neural network provides opportunities to reduce the risk of heart stroke. (M. A. Jabbar, 2013). The main goal of this paper is to reduce the dimensions of data by applying Principal component analysis of projecting the (k) dimensional subspace to increase computational efficiency while retaining more information. An important question is "what is the size of k, which represents" well "data?". This research is dedicated to extensive research in the field of machine learning techniques for heart stroke. It also shows the future prospectus of the machine learning heart stroke algorithm. This paper provides an in-depth analysis of the use of deep learning in the area of heart stroke assessment.

Keywords: Heart stroke, Data Preprocessing, Train & Test datasets, Machine learning algorithms, Classification, Naive Bayes, Decision Tree, and Neural Network algorithms, Dimensionality Reduction, Principle component analysis

INTRODUCTION

Mining data for timely analysis of heart infection appears to be severe. Different people's body may show different symptoms of heart stroke, which may change. However, they often include back pain, jaw pain, neck pain, abdominal pain and shortness of breath, chest pain, arms and shoulders. There are many types of heart stroke including heart failure and stroke and ischemic heart stroke. Although heart stroke is considered one of the most chronic diseases in the world, it is highly preventable at the same time. A healthy lifestyle (lead prevention) and early diagnosis (worse prevention) are the two main sources of heart stroke director. Conducting permanent screening (inferior prevention) shows an excellent role in predicting and preventing heart stroke problems. A few tests, including angiography, chest x-ray, echocardiography, and exercise tolerance tests, support this important issue. However, these tests are expensive and have limited availability of accurate medical equipment.

The content of this paper focuses on different approaches to obtaining valuable data in the diagnosis of heart stroke, using the various data collection tools available. If the heart is not functioning properly, it can damage other parts of the human body, such as the brain, kidneys, and so on. Heart stroke is a type of disease that affects heart function. Nowadays, heart stroke is the leading cause of death. (Avinash Golande, 2019). WHO - The World Health Organization (WHO) estimates that 12 million people die of heart stroke each year? Some heart stroke is including cardiovascular, heart attack, coronary and knocking. Knock is a type of heart stroke that occurs as a result of strengthening, blocking, or narrowing of blood vessels that pass through the brain or can be initiated by high blood pressure.

The major challenge facing the healthcare industry today is the superiority of equipment. The quality of service is defined by proper diagnosis of disease and providing effective treatment to patients. A bad diagnosis can lead to disastrous consequences. History records or data are very large, but from very different bases. The interpretations that doctors make are an important part of this data. (E. O. Olaniyi,. 2015). Real-world data may be noisy, incomplete, and inconsistent, so pre-processing of directive data is necessary to fill in the discarded values in the database. Although cardiovascular diseases are an important source of death in the world in ancient

times, they have been reported to be very preventable and manageable. All and accurate treatment of the disease depends on the prediction of the disease. This is still an open domain waiting to be implemented in the assessment of heart stroke. Guidelines are discussed along with machine learning algorithms and some deep learning techniques for assessing heart stroke. Analytical comparison was made to find the best algorithm available for the medical data set. In the future, our goal is to work in a temporary medical dataset, where the dataset changes over time and the dataset is re-trained.

The next section of this research deals with different algorithms of machine learning of heart stroke and their relative comparison with different parameters. All these methods use old patient records to obtain a new patient assessment. This cardiovascular disease approach helps doctors assess heart stroke at an early stage of the disease, saving millions of lives. In this research study, the Hybrid Intelligent Predictive System is based on machine learning to diagnose heart stroke. The system was tested in the *Cleveland Heart stroke* Dataset. The seven known classifications, such as logistic regression, Naive Bayes, Decision Tree, and Neural Network algorithms to select the most important elements. (Amin Ul Haq, et.al. 2018). Various evaluation metrics have also been adopted to check the performance of classifiers.

The main advantages of the proposed research work are as follows:

(A) The performance of all classifiers in terms of classification accuracy and execution time is checked for full characteristics. (Amin Ul Haq, et.al. 2018).

(B) Classification performances were checked on selected items selected by Feature selection Algorithms (FS) with cross-validation.

(C) After implementing the machine learning algorithms, by applying Principal component analysis for reducing dimensionality.

(D) The study indicates which functional algorithm is possible with which classification to create a highly intelligent heart stroke system that accurately characterizes heart stroke and healthy people.

The remainder of the paper is divided as follows: In Part 2, the basic information for the heart stroke data set is briefly reviewed by theoretical and mathematical background for Feature selection and machine learning classification algorithms. (Amin Ul Haq, et.al. 2018). It also discusses cross validation techniques and performance evaluation metrics. The experimental results are discussed in detail. The main cause of stroke is arterial obstruction. It has many other names such as cardiovascular diseases and arterial hypertension. About 26 million people worldwide suffer from heart stroke. It is alarming that this ratio will increase rapidly in the coming years if preventive measures are not taken effectively. (Alotaibi, F. S. 2018).

Heart stroke has raised serious concerns among researchers; One of the major challenges in heart stroke is to properly detect and detect its presence in humans. Early methods were not effective in diagnosing, and the medical professor was not as effective in predicting heart stroke. There are various medical tools on the market for assessing heart stroke, two of which are major problems, the first being that they are very expensive and the second is the inability to assess the incidence of heart stroke in humans. (Himanshu Sharma,. 2017). According to the latest WHO survey, only 67% of heart stroke s are accurately assessed by a medical professional, so there is extensive research on the prediction of heart stroke in humans.

With the advancement of computer science, there are huge opportunities in various fields, and medical science is one of the areas where it is possible to use the computer science tool. In the field of IT applications, it shifts from metrology to ocean engineering. Medical science has also used some of the major tools available in the field of computer science. (O. Olaniyi. 2015). Over the past decade, artificial intelligence has grown thanks to the progression of computational power. Machine learning is a tool that is widely used in domains because it does not require a different algorithm for different data sets. Programmable machine learning capabilities bring a lot of strength and open new doors to medical science opportunities.

In a variety of life-threatening diseases, medical research is increasingly seeking heart stroke. Diagnosis of heart stroke is a challenging task that provides an automated assessment of the patient's heart condition, making further treatment more effective. Diagnosis of heart stroke usually depends on the patient's symptoms, symptoms and physical examination. (Singh, P., Singh, S., & Pandi-Jain, G. S. 2018). There are many factors that increase the risk of heart stroke, such as smoking habits, body cholesterol levels, family history of heart stroke , esophagus, high blood pressure and lack of physical activity.

The main challenge facing healthcare organizations, such as hospitals and health centers, is to provide quality services at an affordable cost. Quality service requires proper patient diagnosis and effective treatment. The available heart stroke database

contains numerical and taxonomic data. Before further processing, these records can be cleaned and filtered to filter out irrelevant data from the database. (Singh, P., Singh, S., & Pandi-Jain, G. S. 2018). The proposed system can determine accurate encryption, i.e. patterns and relationships associated with heart stroke from the historical heart stroke database. It can also answer complex questions about the diagnosis of heart stroke; Therefore, it is useful for healthcare professionals to make smart clinical decisions. The results show that the proposed system has unique potential in achieving the objectives set by the extraction target.

Literature study

The "Cleveland Heart stroke 2016 Data Set" is used by various scientists and can be accessed from the online data repository at the University of California, Irvine. This dataset was used in this research study to design a machine learning system to diagnose heart stroke. (Amin Ul Haq, et.al. 2018). The Cleveland Heart stroke Dataset contains a sample size of 303 patients, 76 functions, and some missing values. At the time of analysis, 6 samples were omitted because there were no values in the column columns and the remaining sample size was 297 13 with more independent input features and the target output tag was extracted and used to diagnose heart stroke. The target exit label contains two classes to refer to the heart patient or the general patient. Thus, the data set contains a matrix of 29713 elements. A complete description of 13 data and 297 examples of data set functions. Along with a healthy lifestyle and diet, proper time analysis and comprehensive analysis are other important factors that can ultimately save lives. Most of the time, patients undergo numerous tests that can be overloaded with unnecessary physical activity, over time, and certainly with additional financial charges. As previous studies have suggested, the most common causes of heart stroke are unhealthy diet, tobacco, high sugar, overweight or body fat. (Alotaibi, F. S. 2018).

Frequent symptoms may be hand and chest pain. Obviously, these causes are independent of each other; Proper analysis of this type of data set improves the diagnostic process and helps cardiologists. This research therefore seeks to improve the performance of classifiers by experimenting with multiple machine learning models to make better use of data sets derived from various medical databases. (Alotaibi, F. S. 2018). This work is divided into the following sections: The next section describes a comprehensive overview of the use of machine learning models for predicting heart stroke. In coming sections describes the data overview, the number of attributes, and the description of each attribute. The pre-processing steps used in this study. In addition, after that refer to the design, implementation, and classification performance of the experiment. Finally, the study is completed in conclusion.

Implementation

The proportion of heart failure patients increases daily. To overcome this dangerous situation and reduce the risk of heart failure, a system that can make rules or classify data through machine learning is needed. Therefore, this research discusses, designed, and implemented the machine learning model by combining five different algorithms. Rapid Minor is the tool used in this research, which calculates the high accuracy of Matlab and Weka. (Ghwanmeh, A. 2013). Compared with previous research, this study showed significant improvement and higher accuracy than previous work. The main limitation of this work is the small size of the data set. The data set includes a limited number of patient records; For this reason, the data set is expanded using appropriate methods. The results suggest that the system can be useful and useful for cardiologists and cardiologists to diagnose heart attacks in the patient.

Methodology of the Proposed System

A proposed system for classifying heart stroke and healthy people was developed. The performance of various predictive machine learning models has been tested for the diagnosis of heart stroke in both complete and selected tasks. The Cleveland Heart stroke data set has been implemented in many studies and is used in our study. The popular logistic regression of machine learning classifiers is used in the Naive Bayes, Decision Tree, and Neural Network algorithms. Validation and performance evaluation metrics of the model were calculated.

Steps to predictions

The methodology of the proposed system structured into five stages including:

(1) Selection of dataset/Data Preprocessing
(2) feature selection,
(3) Model Implementation
(4) Dimensionality Reduction Result Comparison

Data Preprocessing

Data preprocessing is essential for efficient representation of data and for the classification of machine learning that must be effectively trained and tested. Pre-processing techniques such as missing values, standard scalar, and MinMax Scalar are used in the dataset for effective use in classifiers. (Amin Ul Haq, et.al. 2018). The standard property ensures that each property has a difference of 0 and 1, so that all properties match the same coefficient. Similarly, in the MinMax scalar, it scrolls the data so that all functions are between 0 and 1. The missing value

function is removed from the row data set. All these preprocessing techniques were used in this research.

Feature Selection Algorithms

Feature selection is the machine learning process is essential because sometimes inconsistent characteristics affect the performance of the machine learning classification. Selecting features improves the accuracy of a classification and reduces model execution time. (Amin Ul Haq, et.al. 2018). We use three known FS algorithms to select elements in our system, and these algorithms select important features. The operator of the least absolute contraction and choice. The minimum absolute reduction and selection of the operator selection functions depend on updating the absolute value of the element coefficient. Some element multiplier values are zero and these zero multiplier properties are excluded from a subset of elements. Lasso has excellent performance with low coefficient values. Selected subsets of elements contain elements that have high coefficient values. Some irrelevant attributes can be selected and include a subset of the selected function.

Data preprocessing is an indispensable step to clean up data and use it in any experiment involving machine learning or data acquisition. In this study, several pre-processing steps were applied to the selected dataset. First, the size of the data set was found to be insufficient to implement machine learning approaches. As stated in, machine learning data set size can cause bias and affect the results generated by machine learning models. (Amin Ul Haq, et.al. 2018). Therefore, a random number generation technique is used to generate random values for each column for each attribute using the minimum and maximum values . As seen in the results section, this has helped us to increase data efficiency, which has a positive impact on classification performance. In the end, the data increased the volume by three times.

In addition, outlier detection techniques have been used to measure noise in the data. The data does not detect noisy values , and no outliers are found in the dataset. The outlier values are found using a fast miner operator with distance. Data discernment, transformation, and binning methods were also used to check for further differences in the data set. The next step is to convert the data values into the appropriate data type. In this study, several models were used to check the performance of the assessment. Therefore, it is necessary to convert the data type of some attributes to the required format based on the model specification. In particular, the design of the experiment is based on binary classification, a process of classifying a data set according to predefined classes, which is widely used in the application of machine learning algorithms. (Amin Ul Haq, et.al. 2018). Therefore, the same binary classification was used in the data set, where the binary classification presented in this study is a good way to show the performance accuracy of the selected classification.

Mainly proposed method has separated into two different phases, training and testing.

Training:

- Collect information from the Internet, equivalent to artificial information, as well as time-consuming audit information.
- Pre-processing, processing of information, retrieval of information, search for external information and conversion of information.
- Time ago, this step information is stored in a record called Background, which is used during the test of time.

Testing

- The Scheme Basic Scheme creates an IoT-based health care environment where we can use a small number of sensors as wearable devices.
- Then we associate each sensor with the Raspberry Pi and we also collect access information along with the sensor. (Amin Ul Haq, et.al. 2018).
- Everything Harvest cultivated is integrated into the world record using an association-centered design.
- Ing When testing, we study all manufacturing information at the same time.
- Using different classifications also evaluates the efficiency of the selection system.

Algorithm: Predicting Heart Failure Disease

Step 1: Selection of dataset/Data Preprocessing

{

Data overview
Detect and remove outliers
Detect and impute missing data
Data enhancement using random number generators
Applying suitable normalization techniques

}

Step 2: Feature Selection

{

Understanding data value (classes)
Machine learning model selection

}

Step 3: Model Implementation

{

Import Data Implementing all models together using Machine Learning algorithms
}

Step 4: Dimensionality Reduction
{
Calculate Accuracy using applied ML algorithms
Analyzing the result through applying Principle component analysis (PCA)
}

Step 5: Result Comparison
{
Comparing the accuracy among all models Comparing the result with previous ML algorithms and with the final output after applying PCA
}

Most of the features in the selected dataset are nominal. For example, the attribute describes the thallium test value based on four predefined values (0, 1, 2, 3). Similarly, another independent feature of the CP dataset was that the patient underwent chest pain conditions (0, 1, 2, 3) upon hospitalization, where "0" was normal and ―3‖ was worse. The goal column in the data set is also called the class attribute, and the two types of predefined classes are called ―0‖ and ―1‖. This symptom refers to the overall condition of patients using other independent variables. (Amin Ul Haq, et.al. 2018). The value of ―0‖ means that the patient may not have heart failure and heart means the patient may have heart failure. For example, using the value of all independent variables, if the patient has high blood pressure, sugar, and high thallium values, he or she is likely to have heart failure and vice versa.

Description of 15 used parameters

Sno	Parameters	Parameter description	Values
1	age	Age in years	Continuous
2	sex	Male or female	1= male 0= female
3	thestbps	Resting blood pressure	Continuous value in mmHg
4	cp	Chest pain type	1= typical type 1 2= typical type angina 3= non-angina pain 4= asymptomatic
5	chol	Serum cholesterol	Continuous value in mm/dL
6	fbs	Fasting blood sugar	1≥120 mg/dL 0≤120 mg/dL
7	restecg	Resting electrographic results	0= normal 1= having ST-T wave abnormal 2= left ventricular hypertrophy
8	thalach	Maximum heart rate achieved	Continuous value
9	old peak	ST depression induced by exercise relative to rest	Continuous value
10	exang	Exercise induced angina	0= no 1= yes
11	ca	Number of major vessels colored by fluoroscopy	0–3 value
12	slope	Slope of the peak exercise ST segment	1= unsloping 2= flat 3= downsloping

13	thal	Defect type	3= normal 6= fixed 7= reversible defect
14	obes	Obesity	1= yes 0= no
15	num	Diagnosis of heart stroke	0%≤50% 1%.50%

Importing three algorithms

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
form sklearn.naive-bayes import GaussianNB
```

Naïve Bayes.

NB is a classification-based learning algorithm. It is based on conditional probability theory to determine the class of a new character vector. Note. NB uses a training data set to determine the conditional probability value of a vector for a given class. (Amin Ul Haq,. 2018). After calculating the probable conditional value of each vector, the new vector class is calculated based on the conditional probability. Note NB is used to classify text-related problems.

The NB is a classification supervised learning algorithm.

It is based on conditional probability theory to determine the class of a new character vector. NB uses training data to determine the conditional probability value of vectors for a given class. After calculating the probable conditional value of each vector, the new vector class is calculated based on the conditional probability. Note NB is used to classify text-related problems.

Another classification used in this study is the Nav Bayes. It is a supervised learner classification model that classifies the data by calculating the likelihood of independent variables. After calculating the probability of each class, a high probability class is assigned to the entire transaction. Naïve Bayes is a common method used to estimate classes for a variety of data sets, such as obtaining educational data and obtaining medical data. (Alotaibi, F. S. 2018). The model can also be used to classify different types of data sets, such as sentiment analysis and virus detection. It works by using values for independent variables and predicting a predefined class for each record. They measure the probability of A shown in B in the following equation. Work to find a different class for each attribute; in this scenario, and all other variables are not interdependent. Naïve Bayes uses the following equation to measure probability:

```
model = GaussianNB()
model.fit(X_train, y_train)
predict = model.predict(X_test)
predict
test_score = model.score(X_test, y_test)
Print("NBtest_score:",test_score)
```

NBtest_score: 0.9255641

Decision Tree

The decision tree is a graphical representation of the specific decision situation used in the model attendance model. The root of the decision tree is the root, nodes, and branching decisions. There are several approaches to building a tree, such as ID3, CART, CYT, C5.0, and J48, using methods to classify the dataset using J48, and compare the decision tree with the classification product of another algorithm. Decision tree is used in medical science fields with many parameters when classifying data set.

Decision tree is the most compressed approach of all machine learning algorithms. These clearly reflect important features of the data set. In heart stroke, a number of parameters, such as blood pressure, blood sugar, age, sex, genetic and other factors, affect the patient. Based on the decision tree, the doctor can clearly identify the most effective functions of all parameters. They can also create the most effective characteristics of a population. (Avinash Golande, 2019). Decision tree is based on entropy and obtaining information clearly indicates the importance of data set. The disadvantage of the decision tree is that it suffers from two major assembly problems and is based on the greedy method. Over-adaptation is caused by the expansion of data sets of alphabetically aligned trees, which means that many nodes are needed to spill the data, a problem described by J48.

A decision tree is a supervised machine learning algorithm. The decision tree shape is only one tree, where each node is a leaf node or a decision node. Decision tree methods are simple and easy to understand how to decide. The internal and external nodes in the decision tree are interconnected. Internal nodes are part of decision making, and child nodes visit subsequent nodes. On the other hand, the wing node has no child nodes and is attached to the label.

It is a tree-like classification model that builds on branches and nodes based on the evidence gathered for each attribute in the model learning stage [30]. Depending on the number of entities described in the data set, tree branches, and decision branches are linked. The upload process uses the number of values reserved for each attribute. Further, as per the terms

described for each branch and node, it has reached a decision for each transaction. Finally, according to the Decision Node, the class code is assigned to the record. (Alotaibi, F. S. 2018). This process is repeated and repeated until each transaction class is received. Therefore, this algorithm converts properties to branches and nodes and selects one of the properties as a Decision Node, also known as Class Status. When importing a dataset, you can select a class in the rapid miner.

```
Dt_mod = DecisionTreeClassifier(criterion = 'entropy', max_depth=8)
Dt_mod.fit(X_train, y_train)
y_pred = dt_mod.predict(X_test)
y_pred
ts_dt_score = dt.mod.scpre(X_test, y_test)
print("DTest_score:", ts_dt_score)
```

DTest_score: 0.978811

Artificial Neural Network:

The artificial neural network is a supervised machine learning algorithm, and it is a mathematical model that integrates the neurons that send messages. ANN has three components, including inputs, outputs, and transmission functions. Input units have extraordinary values and weights that can be adjusted during the network training process. Artificial neural network output is calculated for the known class; The mass is converted using the margin of error between the mass and the output of the actual class. ANN is created by integrating neurons. (Alotaibi, F. S. 2018). This diverse combination of neurons from different structures is like multilevel perception.

```
mlp_model = MLPClasssifier()
mlp_model.fit(X_train, y_train)
mlp_predict = mlp_model.predict(X_test)
mlp_predict
ts_mlp_score = mlp_model.score(X_test, y_test)
print("NNtest_score", ts_mlp_score )
```

NNtest_score = 0.979911942

Final Scores:

```
scores = pd.DataFrame({"scores:" {" nb_score":
test_score,"dt_score": ts_ dt_score, "nn_score":
ts_mlp_score}})
```

Algorithm	Ranking	Score
nb_score	III	0.925564
dt_score	II	0.978811
nn_score	I	0.979912

Dimensionality Reduction- Principle component analysis (PCA)

Principal Component Analysis (PCA) is a simple but popular and useful linear transformation technique that is used in many applications such as stock market estimation, gene expression data analysis, and more. In this tutorial, we will see that PCA is not just a "black box" and we understand our internal aspects in three basic steps. Next, we compute the eigenvectors (main components) of the data set and collect them in the projection matrix. These eigenvectors are associated with each eigenvalue, which can be understood as the "length" or "size" of the eigenvector. If some eigenvalues are much larger than others, it is reasonable to reduce the dataset to a smaller subset by PCA by omitting "less informative" eigenvalues.

Major Component Analysis (PCA) is a mathematical process that converts many (probably) interrelated variables into (small) number of unrelated variables called principal components. The first major component is responsible for the greatest variability of the data possible, and each subsequent part represents the greatest variability of the remainder. The principal components analysis is like another multivariate approach called factor analysis. They are often confused, and most scientists do not understand the differences between these two methods or types of analysis, each of which is more appropriate.

PCA reduces the characteristic space from many variables to a small number of factors, and this approach is "independent" (i.e., not considered a dependent variable). PCA is a method of size reduction or data compression.

PCA trends and models simplifies the complexity of high-dimensional data. It does this by turning the data into smaller dimensions that serve as function summaries. High-dimensional data are very common in biology and arise when measuring multiple features, such as the expression of many genes for each sample. This type of data presents several challenges that reduce PCA: increased error rates due to computational costs and multiple test corrections when testing each function to match the result. PCA is like unrestricted teaching methodology and clustering 1 - it finds principles without prior knowledge of whether patterns come from different treatment groups or have phenotypic differences.

A Summary of the PCA Approach

- Standard data authentication.

- Obtain the Eigenvectors and Eigenvalues from a covariance matrix or correlation matrix, or perform Singular Value Decomposition.
- Sort eigenvalues in descending order and choose kk's eigenvectors, which corresponds to the largest eigenvalues of kk, where kk is the number of dimensions ($k \leq dk \leq d$) of the new subspace of the elements.
- kk Create a Ww projection matrix from selected kk owners.
- Change the original Xx dataset by Ww to obtain the Yy subspace of the kk-dimensions.

Applying PCA

Comprehension of results:

Analysis of results is the final stage of research, which involves the completion of experiments, obtained results, and their analysis and discussion. Research is carried out by performing various experiments to verify the effectiveness of the proposed algorithm in terms of different parameters such as data set size, and data set type and different algorithm inputs. Element selection algorithms select classification accuracy, specificity, and sensitivity, important features that improve classification performance in terms of MCC and reduce algorithm calculation time. The logistic regression of classifiers with 10-fold cross-validation showed 89% better accuracy when selected using the FS relief algorithm. If we look at the good performance of logistic regression with mitigation, it is a good estimation system in terms of accuracy. FS algorithms select important features related to separating heart stroke from healthy individuals. According to FS algorithms, thallium scanning, chest pain and exercise-induced angina type are the most important and appropriate features; The results of the three FS algorithms show that fasting blood sugar performance is not appropriate for classifying heart stroke and healthy people.

The novelty of this research is the development of a diagnostic system for heart stroke. The system uses three FS algorithms, seven classifications, a cross-validation method, and a performance evaluation metric for heart stroke diagnostics. The system was tested in the Cleveland Heart stroke Data Set to classify heart stroke and healthy people. For the diagnosis of heart stroke, it is advisable to create a decision support system based on machine learning. In addition, some irrelevant features reduce the performance of the analysis system and increase the computational time. (Amin Ul Haq, et.al. 2018). Another innovative aspect of this study is the use of element selection algorithms to select the best features that improve classification accuracy and reduce the analysis system execution time as well as by reducing dimensionality using Principle component analysis will get the accurate results than regular machine learning algorithms predictions. In the future, we will conduct further experiments to improve the performance of these predictive classifiers in the detection of heart stroke using other function selection algorithms and optimization techniques.

Reducing the dimensionality using Principle component analysis:

```
pca = PCA(c_components=3)
pca = pca.fit(X_train) pca
PCtrain = pca.transform(X_train) PCtrain
PCtest = pca.transform(X_test) PCtest
```

Final Scores using PCA=

```
pca_score = pd.DataFrame({“pc_score:” {“pc_nb_score”:pc_nb_ts_score, ”pc_dt_score”: pc_dt_ts_score,
“pc_nn_score”: pc_nn_ts_score}})
```

Algorithm	Ranking	Score
pc_nb_score	II	0.979499
pc_dt_score	III	0.977160
pc_nn_score	I	0.979912

Considerations:

- *For both normal and after applying PCA the Neural Networks gives best and accurate results with the highest accuracy of 0.979912.*
- *The considerable change occurred in Naive Bayes algorithm which is from 0.925564 to 0.979499.*
- *But for decision tree there is no improvement even though it has decreased its value.*

Conclusion

In conclusion, the obvious research gap in previous surveys is that the measured accuracy does not reduce recall. As many researchers discussed general machine learning approaches are nowhere used. This section, therefore, provides a comprehensive overview of previous work with the assessment of heart stroke in a patient using ML procedures. The aim of this study is to refine previous work using selected data and ML models, as described in the next section. The performance of each model is discussed in the Results section. (Alotaibi, F. S. 2018). The models and datasets selected in this research are based on previous work. The most common ML procedures found and used in this study are; Decision Tree, Naïve-Bayes, Neural network algorithms. This study used a set of data collected from the database, which was originally published in the Machine Learning UCI Data Repository. Previously, experiments that attempted to assess heart stroke, detail and measurement accuracy there we are comparing three algorithms. Finally, and the results section presents a comparative study of understanding the performance of classifiers in this study by applying Principle component analysis with the previous work.

Heart stroke is a major health problem in human society. This article summarizes the latest methods and available methods for predicting the disease. Deeper practice in the emerging field of artificial intelligence has shown good results in another area of medical analytics with high accuracy. To select a subset of variables from a large set that has the highest correlation with the original set of variables. Principal Component Analysis (PCA) is a dimension reduction tool used to reduce a small set of variables in a large set to a smaller set, which still contains more information in the larger set. The goal is to reduce the dimensions and there is no guarantee that the measurements will be understandable (a fact that most frequent statisticians do not realize).

REFERENCES

Alotaibi, F. S. (2018). *Implementation of Machine Learning Model to Predict Heart Failure Disease.* . Retrieved from https://thesai.org/Downloads/Volume10No6/Paper_37-Implementation_of_Machine_Learning_Model.pdf .

Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018. https://doi.org/10.1155/2018/3860146.

Avinash Golande, Pavan Kumar, 2019, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019. Retrieval Number: A11740681S419/19©BEIESP

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.

E. O. Olaniyi and O. K. Oyedotun, "Heart diseases diagnosis using neural networks arbitration," International Journal of Intelligent Systems and Applications, vol. 7, no. 12, pp. 75–82, 2015.

Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," Journal of Intelligent Learning Systems and Applications, vol. 5, no. 3, pp. 176–183, 2013.

Himanshu Sharma, M A Rizvi, Prediction of Heart Disease using Machine Learning Algorithms: A Survey,2017, International Journal on Recent and Innovation Trends in Computing and Communication. ISSN: 2321-8169 Volume: 5 Issue: 8, Available @ http://www.ijritcc.org

M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using artificial neural network and feature subset selection," Global Journal of Computer Science and Technology Neural & Artificial Intelligence, vol. 13, no. 11, 2013.

M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," International Journal of Control Theory and Applications, vol. 9, pp. 256–260, 2016.

Mourão-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data," NeuroImage, vol. 28, no. 4, pp. 980–995, 2005.

Predicting the presence of Heart Diseases using Machine Learning . (2018). Retrieved from https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c.

Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal of Computer Science Issues, vol. 8, no. 2, pp. 150–154, 2011.

Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. International journal of nanomedicine, 13(T-NANO 2014 Abstracts), 121–124. doi:10.2147/IJN.S124998