DOI: 10.24205/03276716.2020.824

Examining the Effectiveness of Machine Learning Algorithms as Classifiers for Predicting Disease Severity in Data Warehouse Environments

Sakshi Hooda*, Suman Mann**

Abstract

Recent trends in the incorporation of computational science into medical and biological field of research have led to the accumulation of huge amounts of data regarding medical and experimental information. The application of data mining in healthcare sectors enables early prediction of patient conditions and their behaviors by performing data analysis and discovering relations from seemingly unrelated large volume of collected data. There is also increasing popularity of data mining in healthcare operations due to its ability to benefit all parties. For instance, data mining application in this sector aids in ensuring that patients receive more affordable and better healthcare services, physicians identify the best practices and effective treatments, healthcare firms make informed decisions about customer relationship management, and healthcare insurers detect abuse and fraud. Despite these promising trends, however, the resultant and huge data amounts that healthcare transactions generate prove voluminous and too complex to process and analyze using traditional approaches. Also, the conventional mechanism involved in the extraction of information from the data warehouse does not identify the hidden patterns involved and hence a new approach is adopted in this research to classify the data to predict the medical conditions of a patient. Moreover, in this research, prediction of severity of brain-related diseases is described based on the medical attributes using advantages of machine learning algorithms as classifier. This was implemented by utilizing the data obtained from the medical data warehouse (DWE). The extract, transform, load (ETL) process and the Online Analytical Processing (OLAP) approach were used for feature extraction, training and testing of data. Machine learning algorithms such as the Artificial Neural Network (ANN) and Support Vector Machine (SVM) are applied to generate optimized input arguments (weights and bias) for the selection of best kernel to classify the data for further diagnosis. The proposed model is found to provide quick response time and least error rate in the identification of diseases. Thus, the proposed framework can be used to predict the conditions of the patients and to provide optimal decisions in the treatment of diseases in healthcare institutions or organizations.

Keywords: Support Vector Machine (SVM), Artificial Neural Network (ANN), ETL (extract, transform and load) process, machine learning, disease severity, data warehouse

1. Introduction

Healthcare sectors are found to produce large volume of data on a regular basis based on the diagnosis made. The information technologies are utilized for the extraction of data and for eliminating the task of manual extraction of data to achieve information regarding regularities directly from electronic records. This is observed to reduce the overall cost incurred in the healthcare sector and early detection of

*(Research Scholar, IPU, New Delhi. Email id-sakshi.hooda@gmail.com)
** (Associate Professor, MSIT, New Delhi. Email-ID-sumanmann2007@gmail.com)

contagious diseases based on the advanced data collections (Dewan & Sharma, 2015). Diagnosing

diseases require an efficient evaluation of the statistical health records of a patient that is stored in the data base housed in the data warehouses (DW) (Pavithra & Parvathi, 2016). From a massive data warehouse system, the medical records of patient, techniques of data mining are utilised in the discovery of the unknown facts and patterns. It can greatly reduce the effect caused by

the medicinal drugs, and it also provides a methodical strategy treatment at lower costs as the disease is predicted in its early stages (Srinivas, Rani, & Govrdhan, 2010). The maintenance of the complex, dynamic and heterogeneous database, also known as big data, in the healthcare sector is of high importance in order to analyse and determine a systematic decisive treatment for the investigation of the prevailing disease. To effectively handle the big unstructured data and store it in an organized manner, many novel techniques have been modelled using machine learning and data mining concepts (Zia & Khan, 2017).

Data mining technology determines the hidden patterns in the data which are used in the healthcare sector for improvising the detection of diseases at an early stage. The obtained data is generally utilised by the practitioners for reducing the effect of drugs such that the therapeutic expenses are reduced (Patidar, Pachori, & Acharya, 2015). This approach is used in health care management for determining the behaviour of the patients based on the available data (Qureshi & Mir, 2017). Besides, data mining uses the clustering approach based on the similarity of the data provided by the health centres (Pavithra & Parvathi, 2016). Data mining technique are used in hospitals with data ware to educate the nurses or medical graduates in order to diagnose the patient more precisely (Lakshmi, Krishna, & Kumar, 2013). The increasing number of health records has led to the introduction of data mining technique for assisting the doctors in extracting the required data of a particular patient. Hence, the machine learning and data mining have been introduced lately in order to systematically organise the clinical reports which will prove to be very efficient in the diagnosis and treatment of the patient (Kavakiotis et al., 2017).

Machine learning in the heath domain helps in the detection of diseases and anticipates the best approach for the treatment based on the big data collected by specialists in the health organisations. There are number of machines

learning algorithms depicted in Fig.1 which can be used by the experts. Subsequently, several algorithms were used in order to detect the accurate predictability of the disease in multiple scenarios such as different diseases, age group, the image patterns or the other morphological features (Elangovan & Sethukarasi, 2016). However, there is a deficiency of an effective interpretation to discover the similarities in the patterns with the huge data which is diverse in nature. Hence, there is a need for more innovative analytical tools to measure the potential of the health risk factors. The quality of prediction in the medical diagnosis needs to be improved for better healthcare management (Moudani, Hussein, & Mora-Camino, 2014). The specialist such as doctors and the data programmers were found to collaborate with the big data and the machine learning concepts for efficient prediction of health issues (Luo et al., 2017). Recently, diverse data mining approaches such as Neural network, Naive Bayes, Decision Tree etc has been employed by health services for the prediction of heart disease (Srinivas et al., 2010). The use of data mining techniques was found to eliminate the manual tasks involved in the extraction of data by using questionnaires. Moreover, the extraction of data by an atomized process is found to improve the decision making and avoiding human errors in the diagnosis ofdiseases.

Further, a conventional data warehouse (DW) does not have the ability to find similarity from the dissimilar image database as its mechanism is restricted to ETL process (extract, transform and load). Therefore, the primary agenda of the present study is to implement a novel technique using the concept of machine learning for predicting the severity of heart diseases as discussed in the following sections.

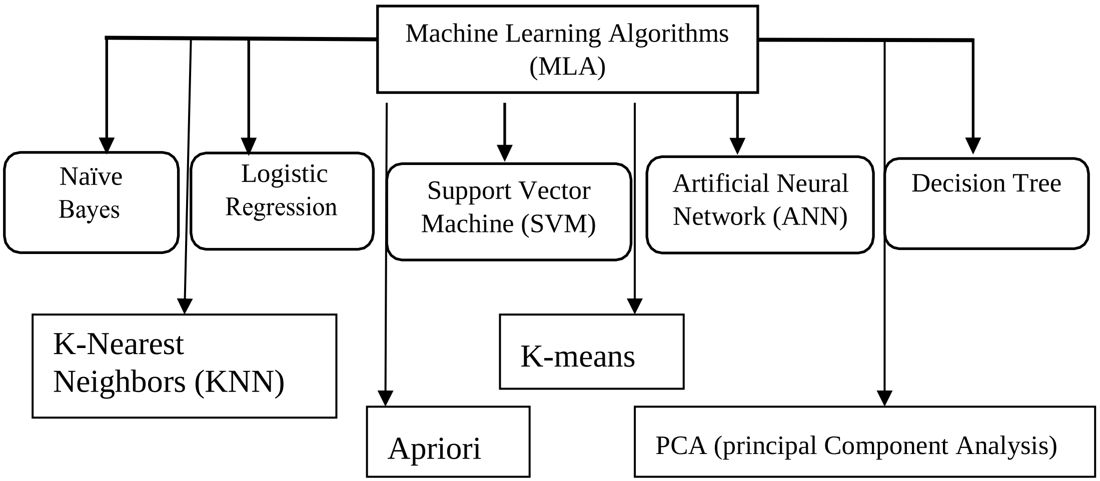

Machine Learning Algorithms (MLA)
Naïve Bayes
Logistic Regression
Support Vector Machine (SVM)
Artificial Neural Network (ANN)
Decision Tree
K-Nearest Neighbors (KNN)
K-means
Apriori
PCA (principal Component Analysis)

Figure 1. **Different Machine Learning Algorithms Based on Different Mathematical Formulations**

2. Related Work

Patidar et al. (2015) proposed Tunable-Q Wavelet Transform (TQWT) method utilizing the heart signals for diagnosing heart disease. This research is found to employ least squares support vector machine (LS- SVM) for improving the accuracy to 96.8%, sensitivity of 100%, and specificity by 93.7%. In a similar study, Lyu et al. (2015) used data mining algorithm to evaluate coronary artery disease. This research is found to utilize linear time-invariant approach to assess coronary heart diseases. Rajeswari et al. (2012) ascertained the use of Network in identifying heart diseases. This study is found to analyse feature selection approach used in neural network to diagnose Ischemic heart disease. The results obtained from the study indicate a precision rate of 89.4% while considering diverse feature attributes.

Teixeira, Annibal, Felipe, Ciferri, and de Aguiar Ciferri (2015) proposed a novel technique to store the enormous data by implementing the image DWE (data warehouse environment) from the health care data warehouse systems. From the computational results obtained, it is observed that the proposed algorithm has a greater adaptability and reliability in terms of organising and routing the clinically big data for disease assessment. Jerez et al. (2010) proposed a novel technique method based on machine learning using the ANN algorithm to predict the reoccurrence and survival chances of the patients related with breast cancer. From the computation results it was seen that the proposed method based on machine learning was more appropriate in predicting the cancer in a patient when compared with the other statistical based methods. Lakshmi et al. (2013) carried out various experimental simulations on ten different data mining concepts in predicting the symptoms of heart attack by varying the different parameters. The results were compared and it was seen that the best data mining technique was PLS-LDA (Partial Least Squares-Linear Discriminant Analysis Target) in terms of accuracy and time efficiency with negligible error. Zou et al. (2018) proposed a novel technique in predicting diabetes mellitus using the tool of machine learning and bioinformatics. This innovative method using the big data has been proved highly accurate and very efficient in monitoring and controlling of diabetes disease. Anita and Priya (2017) proposed a model to accurately predict the early stages of the Parkinson's disease (PD). The results of about three years were collected and analysed and it was found that the proposed regression model showed 99.3% accuracy in comparison with the threshold values. Hence the proposed model is very useful in predicting the PD by measuring the amount of GABA. Julie and Kannan (2012) proposed a novel algorithm based on artificial neural network (ANN) to analyses the learning disability (LD), which is a neurological disease among kids so as to find the best approach in creating a learning experience for children. The researchers carried out a comparative study on the results in determining the efficiency and simplicity of the algorithms. It was concluded from the experimental data that when the ANN was implemented with the data mining techniques it showed better results than other classifiers like Naïve Bayes, SVM and J48. The proposed algorithm can be extended with the implementation of fuzzy neuro logic in determining the LD in the child. Luo et al. (2017) proposed a novel software program which automates the machine learning process (Auto-ML). It automatically selects the big data such as the images from the data warehouse environment. The model was implemented in diagnosing the disease with the real-time patients in a hospital and it was found that the results had least predicting error.

Kavakiotis et al. (2017) proposed a novel method based on data mining using the fuzzy association rule. The experts utilise the weight value which is very useful in predicting the symptoms of disease. The selection algorithm reduces the increasing number of fuzzy weighted association rules by prioritising the important parameters first which shortlists based on the most important attributes. The results are compared with the other fuzzy association algorithms and it was found that the proposed model has greater accuracy and lesser computational time in detecting the diseases. This research work can be extended in curing prognosis and also in detecting the malicious attacks in the wireless communication. Nithya and Duraiswamy (2014) proposed a novel method of data mining based on the fuzzy association rule algorithm and verified with the other data mining classifiers. The proposed model aims in reducing the association rules and membership functions of the collected big data from the clinical centres. Aydilek and Arslan (2013) proposed a novel method called hybrid fuzzy c-means logic by integrating the Support vector regression (SVR) and Genetic algorithm (GA) for estimating the missing data efficiently. From the simulation results it was observed that the lower value of RMSE specified better performance, a higher value of accuracy and lesser run time specified better computation and a higher value of Wilcoxon rank implied that the proposed method is superior to other statistical methods.

The significant challenge observed in data mining in the application of medicine is presence of enormous and heterogeneous raw medical data (Zhang, Qiu, Tsai, Hassan, & Alamri, 2015). These data are generally extracted from patients, and laboratory result. These factors were observed to have critical influence on the diagnosis and treatment of the patient. Besides, the complexity of extraction and collection of data is found to be the barriers in the implementation of data mining approaches in the medical sector.

Another crucial area that has received scholarly attention entails how the severity of brain-related diseases occurs through machine learning (ML) approaches. To achieve this objective, there is the extraction of imaging data on brain functioning. In most cases, functional magnetic resonance imaging (fMRI) data is obtained and aids in reflecting the brain's functional integration. Should observations depict alterations in the functional connectivity (FC) of the brain, potential biomarkers through which brain disorders could be classified or predicted are provided (Zou et al., 2018). Hence, typical classification strategies and brain FC measures aid in depicting the severity of brain-related diseases. Through fMRI, a non-invasive procedure that investigates the functioning of the brain via high spatial resolution, brain connectivity or networks could be detected and characterized in relation to regions that are functionally interconnected. To discern disease severity, especially in conditions such as bipolar disorder (BP) and schizophrenia (SZ), changes in connectivity measures are monitored because the alterations represent critical biomarkers through which individual patients are classified via machine learning techniques (Kavakiotis et al., 2017).

3. Methodology

In this research, a prediction model is developed by selecting the perceptual layer from the image data warehouse which comprises details regarding severity in brain diseases. The input data is captured from the online website and stored in data a warehouse. This data is processed through Extraction Transformation Loading (ETL) stage where it is trained by extracting the specific feature set. This study employed the Kohonen Self Organizing Neural Network variant of ANNs. This variant was chosen because the Euclidian distance algorithm played the part of mathematical categorization and the objective was to discern how machine learning could aid in disease severity prediction by basing on healthy and diseased tissues or organs. A Euclidean distance similarity measurement technique is used to optimize the extracted feature set, and the trained data obtained is stored using image data warehouse. OLAP query process is considered for testing purpose. Both ETL and OLAP are processed through a classifier to predict the severity of the disease.

Indeed, OLAP, the online analytical processing approach, is applied in computing and aims at swiftly answering multi-dimensional analytical (MDA) queries. It is also notable that OLAP comes with three basic analytical operations. These operations include slicing and dicing, drill-down, and roll-up or consolidation. Regarding the consolidation process as one of the basic operations of OLAP, the process ensures that data is accumulated in a manner that would allow for its computation in at least one dimension (Aydilek & Arslan, 2013). On the other hand, drill-down as a basic operation of OLAP reflects a technique that paves the way for system users to navigate through various details of their interest. Lastly, slicing and dicing allows users to slice or take out specific data sets of OLAP cubes before dicing or viewing the slices from various viewpoints – or dimensions. When a database is configured for OLAP, the implication is that multidimensional data models could be used and paves the way for complex ad hoc and analytical queries. Also, the configuration of databases for OLAP has been associated with rapid execution time. In this case, therefore, OLAP was preferred rather than an approach such as online transaction processing (OTP) because the latter approach is associated with less complex queries. Similarly, OLAP was chosen because it is optimized for reads in the majority of cases while an approach such as OTP is expected to process all forms of queries, such as delete, update, insert, and read (Zou et al., 2018).

For the case of the Extraction Transformation Loading (ETL) technique that was used during data processing, it is important to note that the process entails a general procedure through which data is copied from at least one source to a given or targeted destination system. Indeed, the destination system represents data differently relative to its source. Thus, the representation reflects a different context compared to the source of the data. From the ETL concept, three database functions are expected (Zhang et al., 2015). The initial function of extraction implies that from a given database, data is read. The implication is that the stage involves different and multiple types of data sources. In relation to the second function of transforming, it involves the conversion of data that has been extracted from a previous form to establish a new form into which it is expected to be, allowing for its placement in other databases. Notably, data transformation can be seen to take place based on lookup tables or rules, as well as the combination of the given data with other data. Lastly, the loading function of the ETL function constitutes data writing into a given databases that is targeted (Patidar et al., 2015).

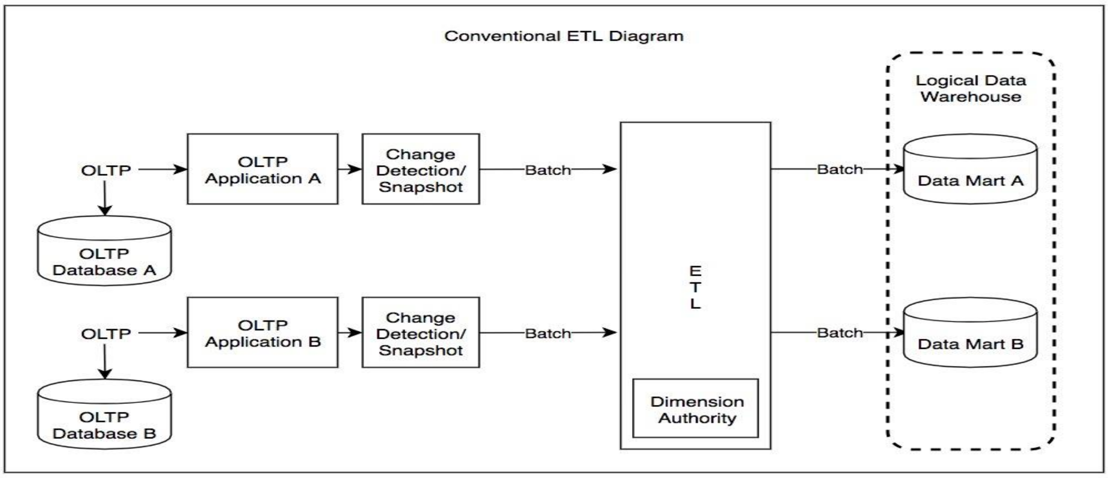


Figure 2(a). **ETL Flow Chart**

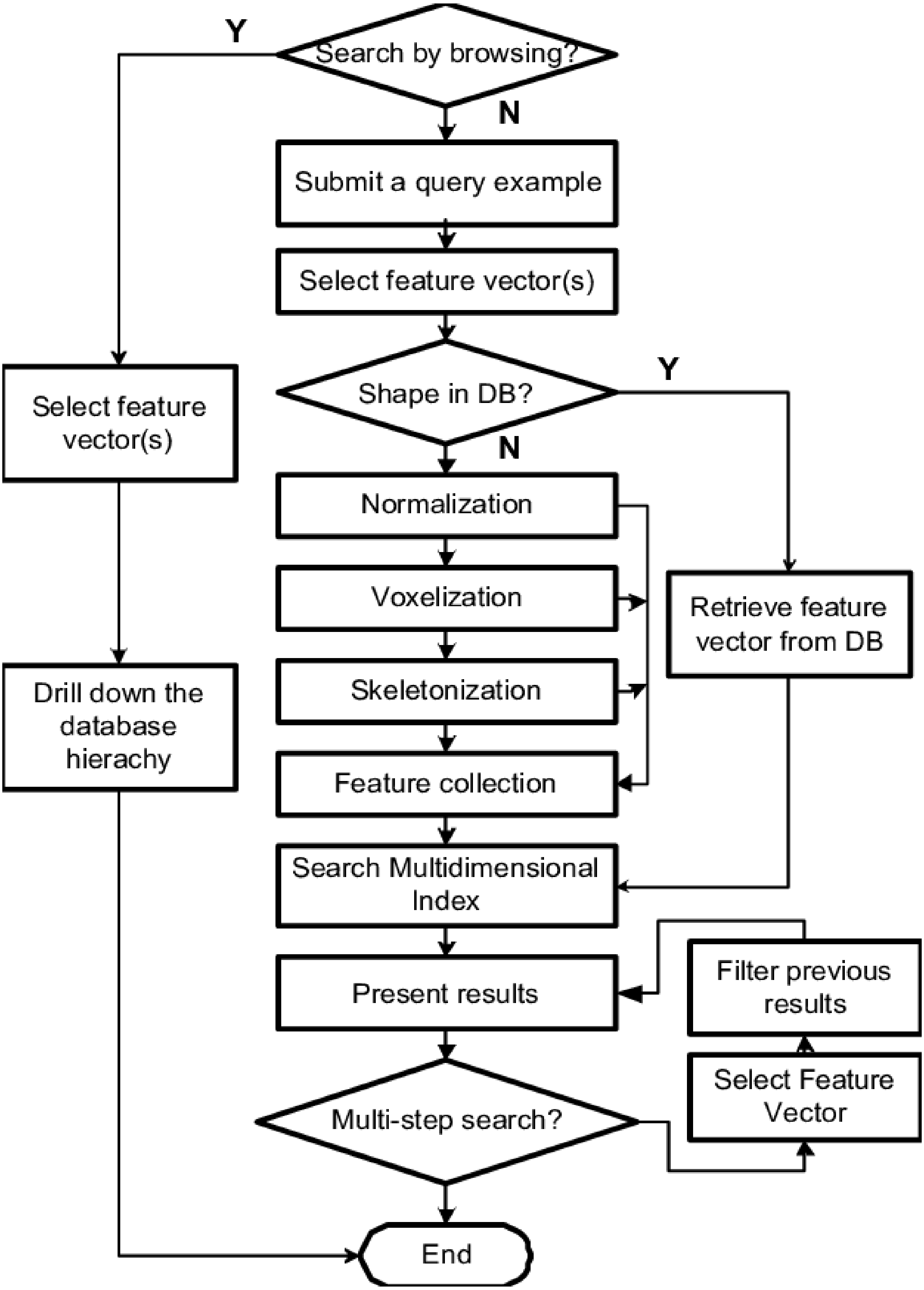


Figure 2(b). **A Query Process Flow Chart**

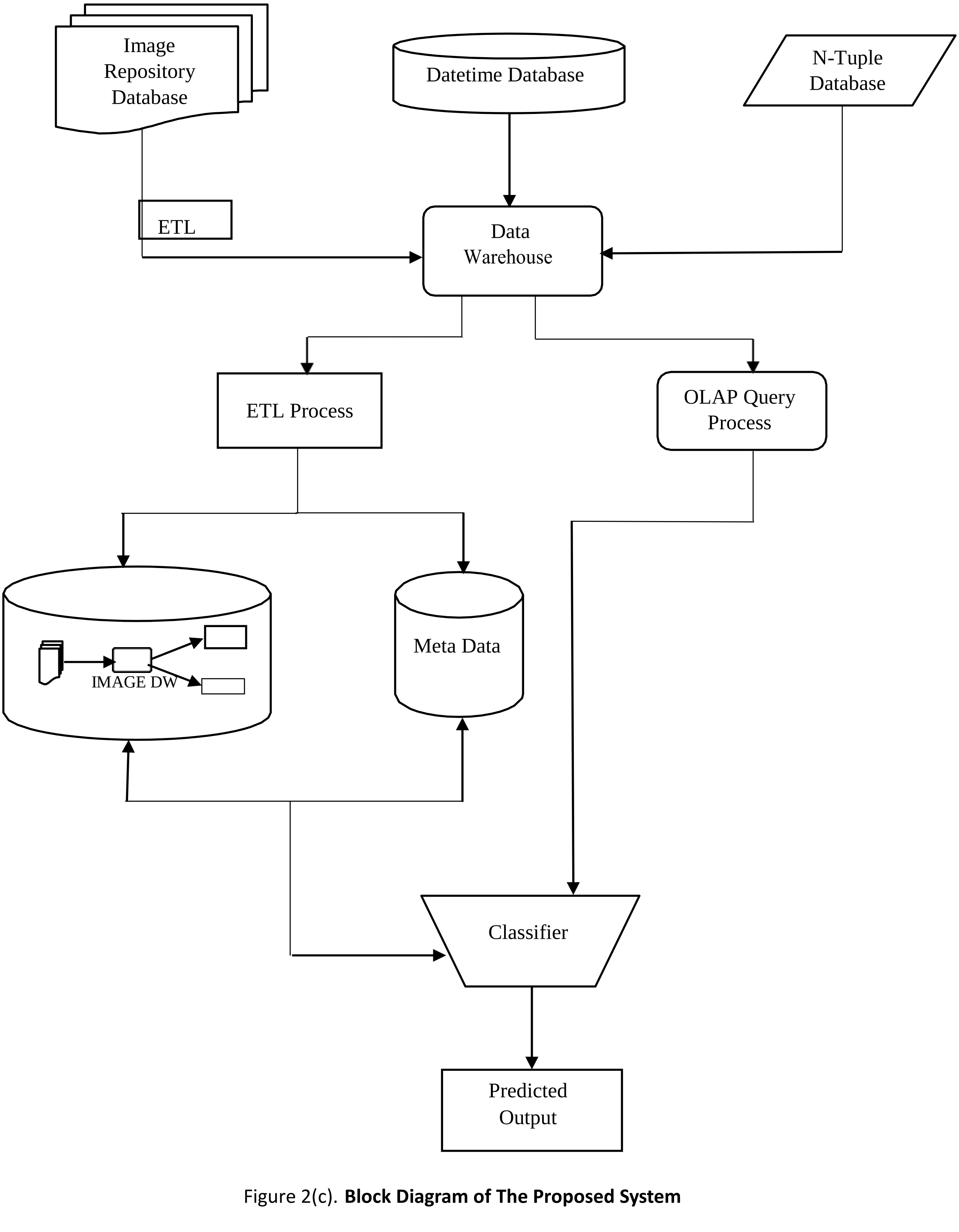


Figure 2(c). **Block Diagram of The Proposed System**

The block diagram of the proposed system is shown in Fig. 2©. In this research, the flowchart of the proposed framework is divided into four stages, first stage explains about the input data and extraction of the input data from data warehouse. Second stage explains about the feature extraction in ETL process while third stage deals with OLAP process, and the fourth stages entails the classifier algorithms and output results. The detailed explanation of the proposed research will be explained as follows:

Input Data Warehouse

The input database comprises with details regarding numerous brain diseases such as brain tumour, dementia, brain injury, multiple sclerosis (S. Hooda & Mann, 2019). These data are obtained from the online website “www.datasus.gov.br/datasus/index.php.” In this research, we have considered data warehouse to store the information obtained from image repository, Date, time, and Tuple data. The image repository is divided into number of perceptual layers for representing the specific feature descriptor in metric space. Date time data provides the details regarding image captured date and time at that particular instant. Further, the key age, age of the person, gender and category is extracted and stored in’ N’ tuple database. ETL step is designed to extract the effective features from input and processed through classifier along with OLAP query process. For evaluation purpose, we have considered 100 images for ETL process and 15 images for OLAP process which comprises details regarding two types of heart diseases (Hooda, 2020). These data are processed through classifier for predicting the type of heart disease.

Extraction Transformation Loading (ETL) process

In this research, we have developed an efficient ETL process for image feature extraction, storage of key features, task execution, conversion and loading of conventional data. The ETL process is divided into three stages on the basis of task execution namely feature extraction, representation of images and similarity calculation of images. These tasks are explained as follows:

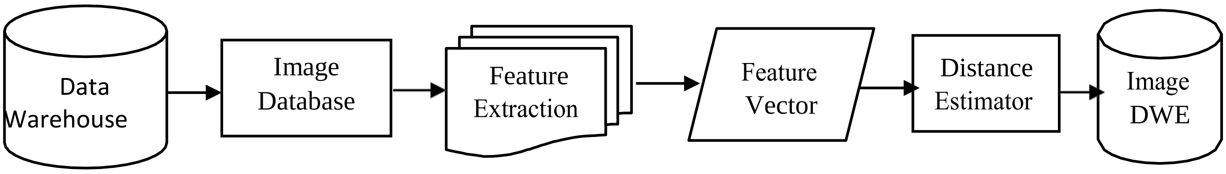


Figure 3. **Flowchart of ETL process**

The flowchart of the proposed ETL process is shown in Fig.3. In this research, 100 input images comprising different heart disease is used in ETL framework for training. The detailed description of various steps involved in ETL process such as processing, feature extraction will be explained in the following stages.

Feature Extraction

Feature extraction is defined as the extraction of most relevant and unique information from raw data and further processed to provide discrimination amongst the set of data (Hooda, 2020). The input image extracted from the data warehouse is stored in image database and processed through a number of feature extractors for extracting the effective features. In this research, we have considered SURF descriptor, haralick descriptor, Zernike movement descriptor, and GLCM as effective features. Indeed, it is important to note that the Zernike moment descriptor play a crucial role in beam optics and their use lies in the extraction of features from the given images. The extracted features aid in the description of an object’s shape characteristics. In practical applications, an illustration of the Zernike moment descriptor in use is that which entails malignant and benign breast mass classification, as well as the classification of surfaces of vibrating disks. Another illustration is that in which Zernike moment descriptors could be used towards the quantification of the shapes of osteosarcoma cancer cell lines, especially on a single-cell level basis. The suitability of the Zernike moment descriptor arose from several reasons or desirable properties with which the descriptor is associated. Some of these properties include multi-level representation towards pattern shape description, fast computation, expression efficiency, robustness to noise, and rotation invariance.

It is further explained as follows:

Speedup Robust Feature (SURF) descriptor

SURF detector comprises with square shaped filters for Gaussian smoothing approximation. It has been observed that square filtering image is faster than integral image. The equation for the same is givenby,

$$S(x,y)=\sum_{i=0}^{x}\sum_{j=0}^{y} I(i,j) \qquad (1)$$

Furthermore, six features are considered for capturing the uniqueness in the image and they are as follows:

A. **Mean:** The mean ‘m’ is defined by capturing the pixel values from the defined window.

$$\bar{x}=A+\frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i} \qquad (2)$$

The mean is calculated by using the formula, Where, A and B defines the pixel image size.

B. **Standard Deviation:** The standard deviation ⍰is used to calculate the mean square deviation of the grey scale pixel P (i, j). The mathematical formula for calculating SD is given by:

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2} \qquad (3)$$

C. **Variance**

Variance is defined as the root means square of standard deviation. The formula for calculating VAR is given by,

$$\text{VAR}=\sqrt{SD} \qquad (4)$$

SD = Standard deviation

D. **Smoothness**

The relative smoothness R is defined as a measure of contrast used to establish descriptors of the relative smoothness. The equation for calculating smoothness is given by,

$$\frac{\text{S⍰1⍰1}}{\text{(1⍰⍰2)}} \qquad (5)$$

E. **Kurtosis**

The flatness or peak value observed amongst distributive relative to normal distribution is defined as kurtosis. The equation for calculating kurtosis is given by,

$$\text{Kurtosis}=\frac{\sum_{i=1}^{N}\frac{(X_i-\bar{X})}{N}}{s^4} \qquad (6)$$

where,
$\bar{X}$ is the mean,
s is the standard deviation
and N is the sample size

F. **Skewness**

Skewness S is used to characterize the asymmetrical degree of pixel distribution in the specified window around its mean. The formula for calculating skewness is given by,

$$G_1=\frac{k_3}{k_2^{3/2}}=\frac{n^2}{(n-1)(n-2)}\frac{m_3}{s^3}$$
$$=\frac{\sqrt{n(n-1)}}{n-2}\frac{m_3}{m_2^{3/2}}=\frac{\sqrt{n(n-1)}}{n-2}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}\right] \qquad (7)$$

2. Haralick Descriptor

The grey level metrics developed by haralick is a most popular and well-known text measure technique. The haralick feature is captured through the correlation and intensity of the pixel in feature space. Fourteen developed of texture features is developed by haralick and from that we are considering six features, which is same as that in SURFdescriptor.

3. Zernike Moments Descriptors

Zernike movement is defined as the projection of image function on the basis of orthogonal basis functions. The parameters such as moments (z), amplitude (A), and phase of the angle (phi) is extracted to calculate the Zernike movement descriptors.

4. Grey Level Co-occurrence Matrix (GLCM)

$$Contrast\square \sum_{a,b\square 0}^{n\square 1} P_{ab}\,(a\square b)^{2} \qquad (8)$$

$$Correlation\square P_{ab} \sum_{a,b\square 0}^{n\square 1} \frac{(a\square\square)(b\square\square)}{\square^{2}} \qquad (9)$$

$$Energy\square \sum_{a,b\square 0}^{n\square 1} (P_{ab})^{2} \qquad (10)$$

$$Homogeneity\,.\square \sum^{n\square 1} P_{ab} \qquad (11)$$

$$a,b\square 0\ 1\square\ (a\ \square b)^2$$

Where, Pab= Element a, b of normalized symmetric GLCM, n. = set of grey level pixels in the image, μ = GLCM mean, σ 2 = Variance of the intensities.

Representation of Images

After the extraction of key features from the input image, it is further processed through individual perceptual layer for image representation. It is observed that tasks are executed only after the entire image set is stored in image DW. The feature vector comprising text data after feature extraction is stored in metadata repository. The dimensionality parameter and diameter of the individual feature dataset is extracted from individual perceptual layer and stored in metadata base and further processed to calculate the distance between the images.

Distance Measurement Calculation

Euclidean distance is the frequently used and effective distance measurement technique for feature set optimization. In this process, the distance between the co-ordinates of the pair of objects is examined using root square distances. The mathematical equation for calculating the Euclidean distance between the two points is given by,

$$X \square (a1, a2, a3, ..., an)$$

$$Y = (b1, b2, b3, ..., bn)\ E(X, Y) = \sqrt{(a1-b1)^2+(a2-b2)^2+(a3-b3)^2+......+(an-bn)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (a_i - b)_i^{\ 2}} \qquad (12)$$

OLAP Query Process

OLAP similarity queries comprise with conventional as well as similarity prediction techniques for predicting the query conditions (Hooda, 2020). This stage is similar as in ETL process except the input data comprises with 15 test images, which is obtained from the database.

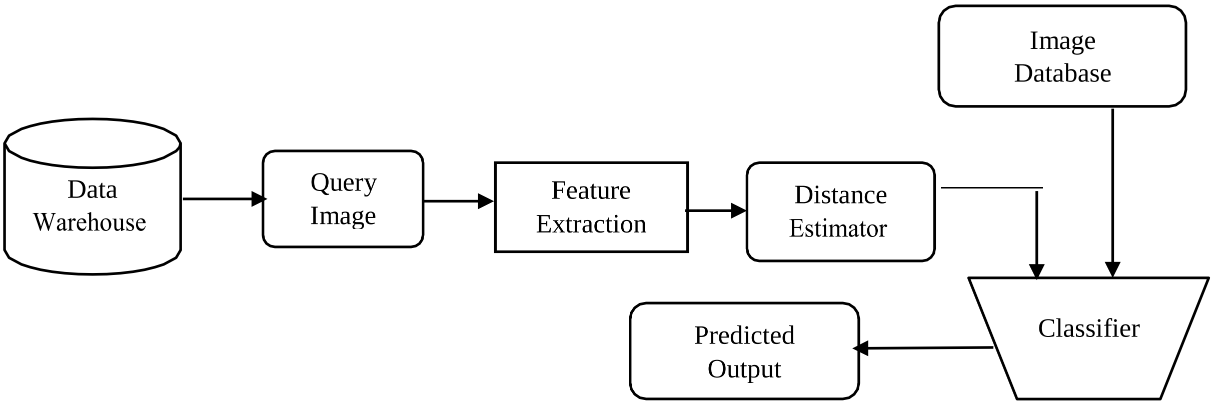


Figure 4. **Flow Chart of Query Process**

Fig. 4 shows the flow chart of the query process. The query image is obtained from the data warehouse and processed through feature extraction stage. These features obtained are processed through similarity measurement Euclidean distance technique for optimization. The data obtained from the ETL process is stored in the image database and the feature set obtained from OLAP query process is compared by using the classification algorithm to predict the state of the disease. In this research, both Artificial Neural Network (ANN) and Support Vector Machine (SVM) are used to validate the effectiveness and accuracy of the system in predicting the heart diseases.

Classifier

In this research, an advantage of machine learning techniques is that it is used for classification and prediction purpose. From the aforementioned literature review, it is observed that artificial neural network and support vector machines are advantageous and effective in prediction of brain disease. The classification model using ANN and SVM is implemented to predict the diseases of the brain. The output obtained from individual model is tabulated and compared with each other to predict better classifier for future studies.

Support Vector Machines (SVM)

In this research, SVM classifier is considered as one of the predicting algorithms to determine the brain disease. The SVM uses three different kernels such as linear kernel (SVM-L), polynomial kernel (SVM-P) and radial basis function (SVM-RBF) in order to predict the severity of the heart diseased data. A kernel function k (x,y) represents a dot product in a new dimensional space as it is used in the support vector machine, and there values are mathematically calculated using the following equations:

Linear Kernel: Training a support vector machine based on linear kernel can enhance the computational time of the data classification. A SVM using the linear kernels has to optimize the ‘C’ parameter, but whereas while training the classifiers with different kernels even the ‘γ’ parameter has to be optimized. Hence a grid search using more kernels will consume lot of time. Linear kernel is given by $K (x, y) = x T y + c$. It is best suitable when the features are more in the data. Gaussian Radial Basis Function has the default value of one. It is given by equation (13)

$$K (xi, xj) = exp(-\gamma ll\ xi - xj \|2) \ldots(13)$$

It considers two parameters ‘C’ and ‘γ’ in generating the input arguments. Unlike the linear kernel, RBF maps the non-linear data samples into the multi-dimensional space with lesser complexity in design.

$$K (xi, yj) = (xi.\ xj)\ d \ldots (14)$$

Polynomial kernel takes into account more number of hyper parameters which require higher mathematical formulation. Generally, a polynomial kernel is considered up to the order three and it is given by equation (14).

Artificial Neural Networks (ANN)

Artificial neural network is defined as an adaptive model for analysing and comparing the predictive data. The functional process of ANN is inspired by the process of neurons in human brain. These systems have the capability to modify the internal structure in accordance with the objective function. From the review, it is observed that ANN is effective in solving the nonlinear type of issues. The advantages of the supervised ANN along with processing is already predefined and is used to determine the error function obtained by calculating the distance amongst the desired and fixed output. The general form of ANN is given by

$$z \square f (x, y^{*}) \qquad (15)$$

Where, w* comprises with the set of parameters to select the best optimization function.

In general, ANN model comprises with input layer, hidden layer and output neuron layer. The input layer comprises with details regarding features extracted from the SURF and OLAP model. These features are processed through the hidden layer. The number of hidden layers is selected on the basis of system complexity and the network trains the hidden layer by adjusting the weights. After processing, the hidden layer output is forwarded to the output layer for prediction.

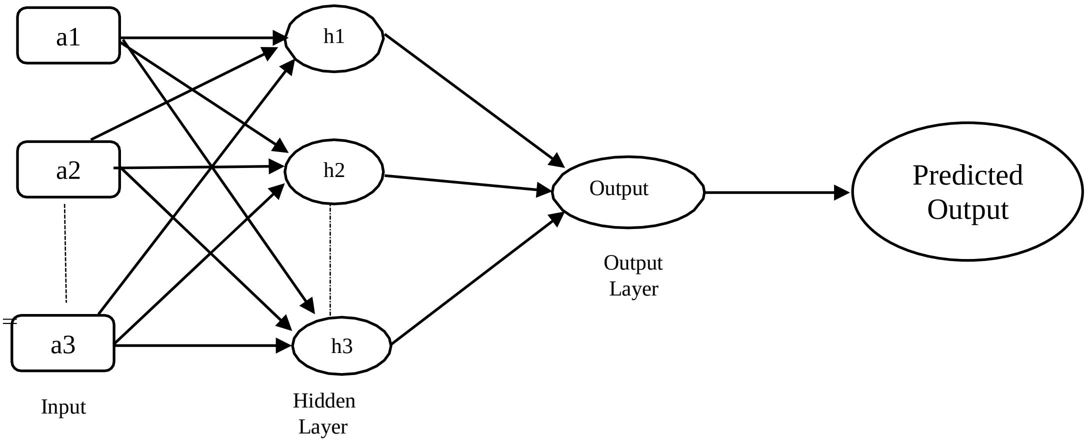


Figure 5. **Block Diagram of ANN**

The block diagram of the proposed system is shown in Fig. 5. The input signals are defined in terms of individual node. Each node manipulates the input signal to provide single output signal. The weight factors are defined in terms of w1, w2 …wn. Weights are defined as adaptive co-efficient within the network and used to calculate the intensity of the input. The output signal is calculated through the product of the individual input (a1, a2 …an) to its corresponding weight factor. The summation of the weighted inputs (a1*w1, a2*w2...an*wn) along with its output signal is used to estimate the predicted output.

4. Results and Discussion

The proposed research study comprises with 115 inputs captured from the online website www.datasus.gov.br/datasus/index.php and simulated using MATLAB R2014a software. The results obtained from the simulation are compared on the basis of SVM and ANN algorithms to identify the most effective algorithm. Thus, by the aforementioned process the data consisting of brain related disease are classified appropriately in predicting the severity. Parameters such as specificity, precision, accuracy, recall, elapsed time and the error rate with respect to algorithms are evaluated using MATLAB tool. Both the algorithms provide the input arguments to the SVM and ANN and the results are evaluated by comparing the performance of the two algorithms. The following are the performance parameters considered for selecting the best algorithm:

Accuracy (ACU): It is given by the ratio of the truly classified samples to the total samples (true positive (TP), true negative (TN), false positive (FP) and false negative (FN)) as given by equation (5)

ACU = (TP+TN)/ (TP+TN+FP+FN) … (5)

Specificity (SP): It is formulated as the ratio of true negative (TN) to the sum of true negative and false positive (FP) samples.

S.P = TN/ (TN+FP) … (6)

Precision (P): It is defined as the ratio of true positive (TP) to the sum of positive samples (true positive and falsepositive).

P= TP/ (TP+FP) … (7)

False Positive Rate (FPR): It is given by the ratio of false positive to the sum of true negative and false positive.

FPR= FP/ (TN+FP)… (8)

False Negative Rate (FNR): It is given by the ratio of false negative to the sum of true positive and false negative.

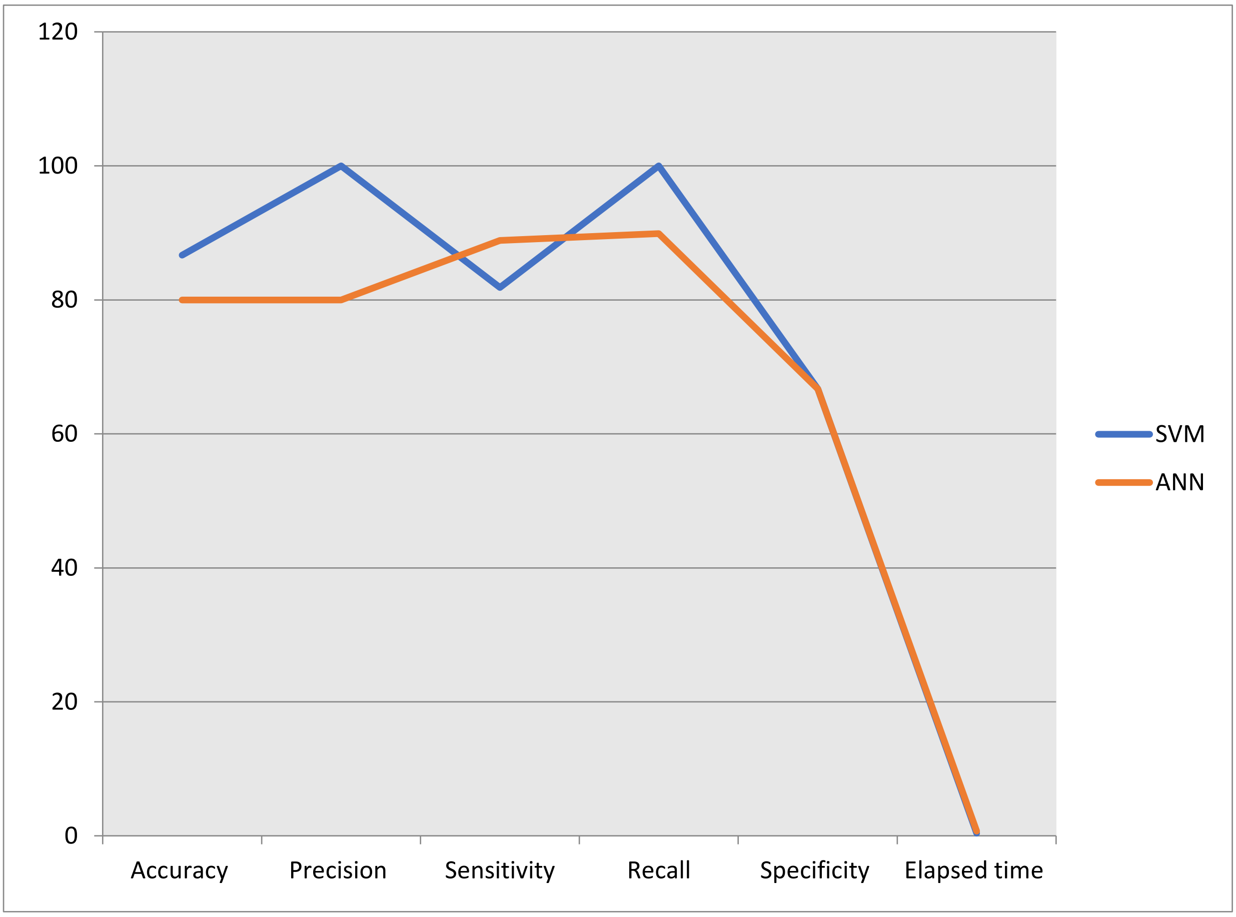
FNR= (FN)/ (TP+FN) … (9)

The input arguments generated by ETL and OLAP Query process are provided to the classifier and is compared with the sample data to identify the best kernel in the SVM classifier. Further, the input data is provided to the SVM classifier and ANN for the prediction of severity of diseases and is compared to determine the performance evaluation of the proposed model.

Notably, the statistical measures of specificity and sensitivity were used to discern the performance of the classification tests that were examined or employed. The aim was to discern how the classification functions employed in this investigation were likely to perform. Of importance to indicate is that sensitivity involves a probability of detection, the recall, or the true positive rate. In this study, therefore, true positive rate or the sensitivity parameter aided in measuring the proportion of the real positives that were identified correctly, a specific example being the percentage of sick persons that were likely to be identified correctly as actually worth diagnosing for a given condition. On the other hand, the case of the parameter of the true negative rate or the specificity aided in measuring the proportion of the real negatives that would be identified correctly, including situations such as those that involved the percentage of healthy people who would have been identified correctly as not having a given condition. Overall, therefore, the true negative rate reflected the specificity parameter concerning healthy persons identified correctly as not to have had a given condition while the true positive rate constituted the sensitivity parameter involving the correct identification of sick persons who were likely to have had a given condition.

Table 1. **Performance Evaluation in Predicting Disease Severity**

Parameters	SVM	ANN
Accuracy	86.67	80
Precision (P) (%)	100	80
Sensitivity (%)	81.82	88.89
Recall	100	88.89
Specificity (%)	66.67	66.67
Elapsed time	0.389280 seconds	0.700700 seconds



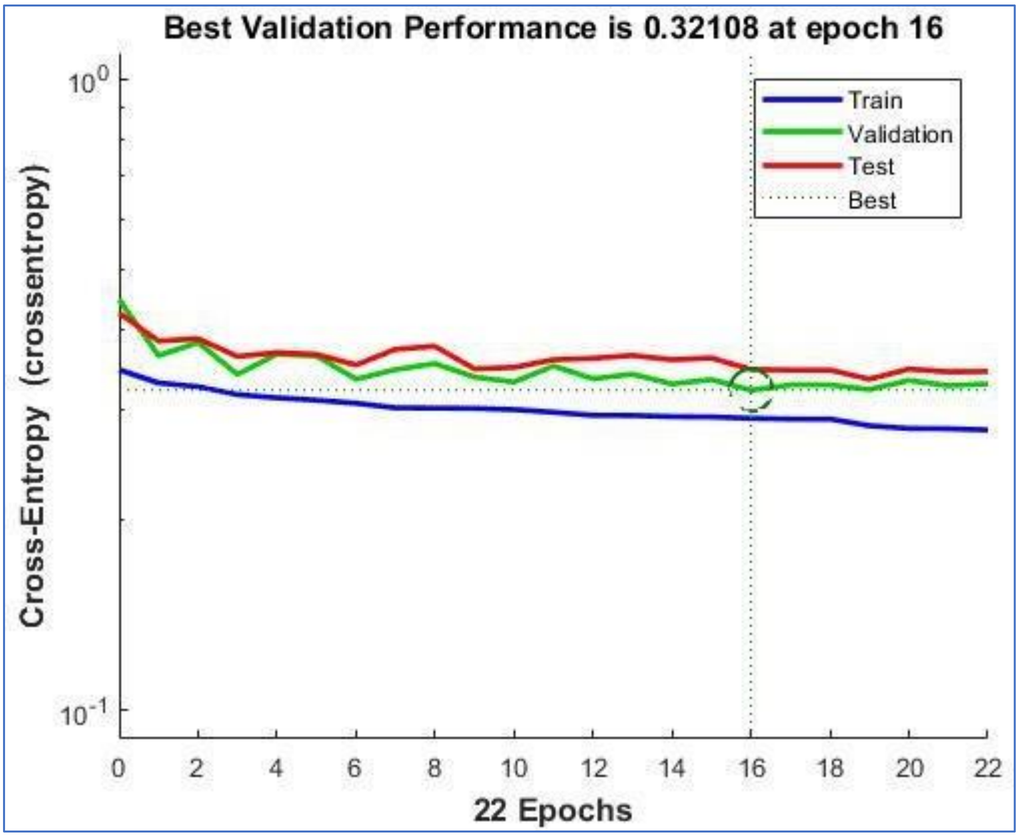


Figure 6. **Performance Model of Proposed ANN Best Validation**

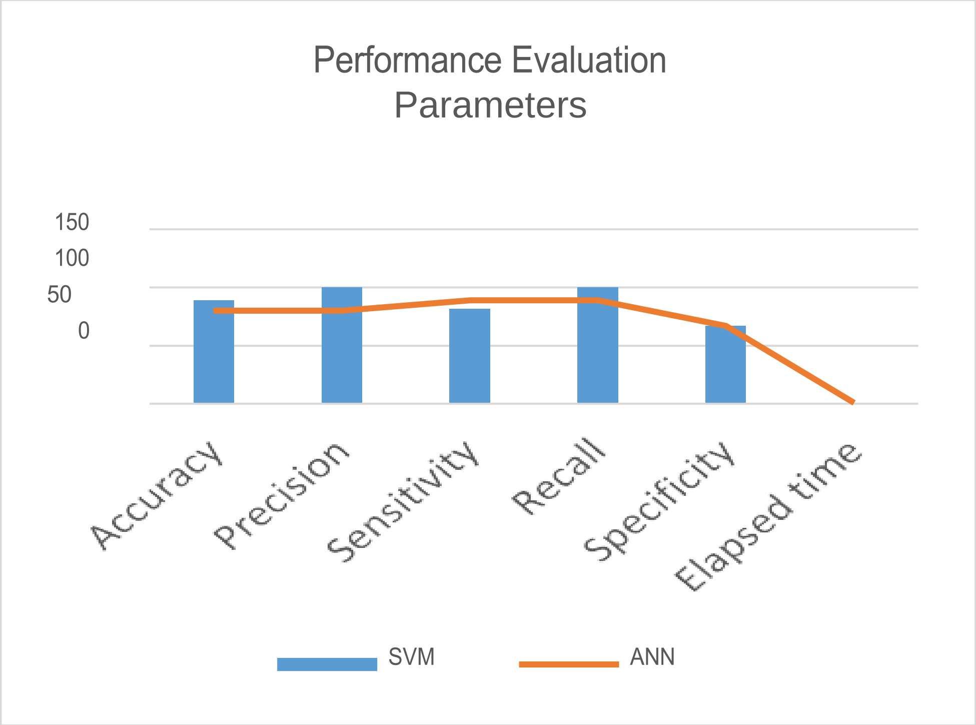


Figure 7. **Comparative Analysis (Performance Evaluation) of SVM and ANN**

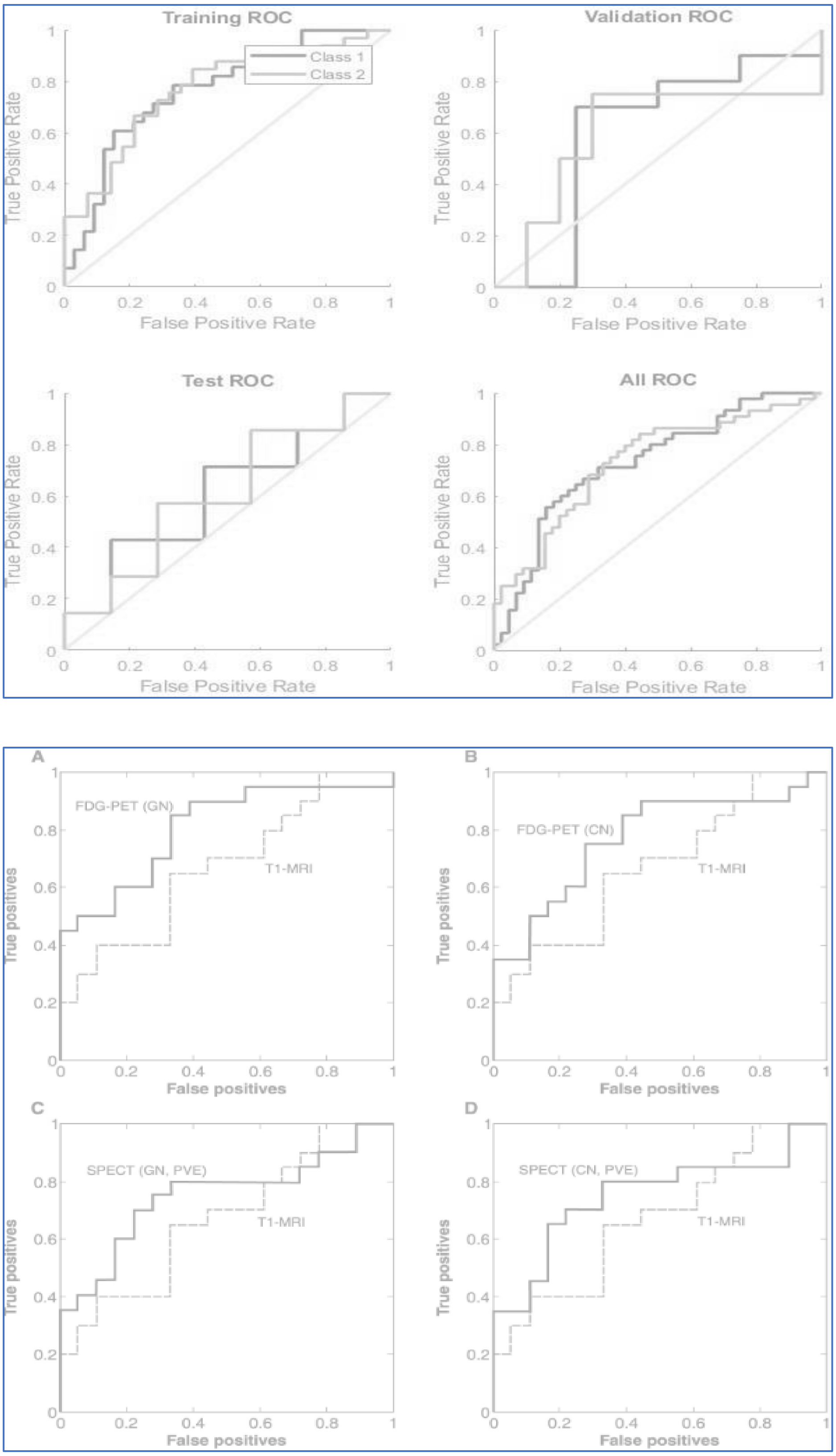


Figure. 8 **Performance Validation Using ROC Curves for ANN and SVM Respectively**

The proposed model is validated in terms of performance parameters and the result obtained for 15 test images and individual classifier is tabulated in Table. 1. Fig. 7. Shows the comparative graph for the performance evaluation parameters and it is observed that SVM outperformed ANN in terms of accuracy, precision, recall rate and elapsed time. The precision rate defines the correctness in defining the output of the predictive model and it has been observed that SVM classifier has 100% precision, which is efficient compared to other existing techniques. The training, testing and validating errors observed during the experimental analysis is used to evaluate the reliability of the classifier. Fig. 6 shows the performance graph of the proposed ANN classifier. It is observed that the best performance validation for ANN classifier is observed at an error rate of 0.32108. Furthermore, it is demonstrated that the significant over fitting for test set and validation set error and best validation performance is obtained at epoch16.

The ROC curve is plotted by varying the threshold voltage to show the performance of the true positive vs. false positive rates. Fig. 8 shows the ROC plot for the proposed ANN method. It is observed that ANN based approach covers the larger area amongst diagonal and upper left corner of the curves. Through the comparative model as shown in Fig. 7, in terms of specificity and sensitivity, it is observed that SVM based model is efficient in predicting the disease severity in big data.

This study also established that given big data, the discovery of pattern trend similarities (hence the ability to predict disease severity) arises from four main forms of health analytics, with machine learning techniques playing a crucial role. One of these analytics is the case of descriptive analytics, whereby machine learning aids in the calculation, interpretation, development, and implementation of data via graphical interpretations. Specific details that machine learning yields via descriptive analytics involves how diseases are treated and managed, the diseases diagnosed, patient symptoms, the revenue generated, and the number of patients treated. On the other hand, machine learning establishes trend similarities via predictive analytics in which future trends and probabilities of similar magnitudes of severity are projected, having learned patterns from big data sets. It is also, notable that machine learning aids in establishing pattern similarities relative to disease severity by focusing on how practitioners could or might have responded, upon which, relative to the level of severity of a disease, the most accurate solutions are established. Lastly, machine learning can be seen to discern similarities in the patterns of disease severity based on discovery analytics in such a way that the discovered knowledge is used as a foundation for developing new innovations and inventions in the data mining healthcare field. Indeed, the proposed model strived to demonstrate the possibility of employing these four forms of health analytics towards predicting disease severity and informing optimal interventions, including situations where brain-related diseases are reported.

5. Conclusion

Data mining is found to have significant application in the field of medical sector and proposed technique is used to improve the health care management. However, the voluminous raw medical data that is available from the health sector develops the complexity in implementing this technique. Besides, these data are required to be stored in data warehouses. Therefore, this research has conducted a comparative analysis of the algorithms such as Artificial Neural Network (ANN) and Support Vector Machines (SVM) for the selection of best machine learning technique such that the brain diseases are diagnosed at an early stage. Furthermore, new technique comprising ETL and OLAP is implemented for extracting the effective features, training and testing of data. The performance of the SVM system is evaluated by considering five major parameters such as specificity, precision, accuracy, recall and error rate. The weighed values are generated by these algorithms for each of the individual classifiers and the highest weight value of the classifier is selected as the best classifiers through performance evaluation. From the proposed research, SVM classifier is found to be the effective technique which showed an accurate prediction of the brain disease data. The algorithms were simulated using the MATLAB 2014 tool and it was found that the data classification was better using the optimisation of the SVM than ANN. Moreover, the results obtained from the research indicate that SVM showed an improved accuracy of 86.67% with high precision rate. Thus, the proposed research will be used to predict the conditions of the patients and to provide optimal decisions in the treatment in Healthcare institutions or organizations. Overall, the proposed model is found to provide quick response time and least error rate in the identification of diseases. It is also notable that the proposed SVM classifier can be simulated using different algorithms and with more numbers of performance parameters to yield more accurate results in the future work.

References

Anita, S., & Priya, P. A. (2017). ESTIMATION OF PARKINSON'S DISEASE RISK BY STATISTICAL MODEL. IIOAB JOURNAL, 8(3), 42-48.

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences, 233, 25-35.

Dewan, A., & Sharma, M. (2015). Prediction of heart disease using a hybrid technique in data mining classification. Paper presented at the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).

Elangovan, K., & Sethukarasi, T. (2016). Knowledge enrichment of prediction using machine learning algorithms for data mining and big data: a survey. Advances in Natural and Applied Sciences, 10(15), 23-31.

Hooda. (2020). A Focus on the ICU's Mortality Prediction Using a CNN-LSTM Model. International Journal of Psychosocial Rehabilitation, Vol. 24(6), 8045-8050.

Hooda, S., & Mann, S. (2019). Distributed Synthetic Minority Oversampling Technique. International Journal of Computational Intelligence Systems, 12(2), 929-936.

Hooda, S. and Mann, S., 2020. A Focus on the ICU's Mortality Prediction Using a CNN-LSTM Model. International Journal of Psychosocial Rehabilitation, Vol. 24, Issue. 6, pp. 8045-8050.

Hooda, S. and Mann, S., 2020. Imbalanced Data Learning with a Novel Ensemble Technique: Extrapolation-SMOTE SVM Bagging. International Journal of Grid and Distributed Computingx, 13(1), pp.1202-1207.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2), 105-115.

Julie, M. D., & Kannan, B. (2012). Attribute reduction and missing value imputing with ANN: prediction of learning disabilities. Neural Computing and Applications, 21(7), 1757-1763.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

Lakshmi, K., Krishna, M. V., & Kumar, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. International Journal of Scientific and Research Publications, 3(6), 1-10.

Luo, G., Stone, B. L., Johnson, M. D., Tarczy-Hornoch, P., Wilcox, A. B., Mooney, S. D., . . . Nkoy, F. L. (2017). Automating construction of machine learning models with clinical big data: proposal rationale and methods. JMIR research protocols, 6(8), e175.

Lyu, Y., Hong, J., Wei, Y., Yang, J., Tang, Y., Wang, W., & Agoulmine, N. (2015). Dynamic evaluation model of coronary heart disease for ubiquitous healthcare. Computers in Industry, 69, 35-44.

Moudani, W., Hussein, M., & Mora-Camino, F. (2014). Heart disease diagnosis using fuzzy supervised learning based on dynamic reduced features. International Journal of E-Health and Medical Communications (IJEHMC), 5(3), 78-101.

Nithya, N., & Duraiswamy, K. (2014). Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface. Sadhana, 39(1), 39-52.

Patidar, S., Pachori, R. B., & Acharya, U. R. (2015). Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. Knowledge-Based Systems, 82, 1-10.

Pavithra, M., & Parvathi, R. (2016). A SURVEY ON DATA MINING APPROACHES FOR HEALTHCARE DOMAIN. Journal of Recent

Research in Engineering and Technology, 3(9), 01-11.

Qureshi, M. A., & Mir, I. A. (2017). Comparative Study of Existing Techniques for Heart Diseases Prediction Using Data Mining Approach. Asian Journal of Computer Science and Information Technology, 50-56.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

Teixeira, J. W., Annibal, L. P., Felipe, J. C., Ciferri, R. R., & de Aguiar Ciferri, C. D. (2015). A similarity-based data warehousing environment for medical images. Computers in Biology and Medicine, 66, 190-208.

Zhang, Y., Qiu, M., Tsai, C.-W., Hassan, M. M., & Alamri, A. (2015). Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. IEEE Systems Journal, 11(1), 88-95.

Zia, U. A., & Khan, N. (2017). An Analysis of Big Data Approaches in Healthcare Sector. International Journal of Technical Research & Science, 2(4), 254-264.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.

List of Abbreviations

ANN – Artificial Neural Network
DW – Data warehouse
ETL – Extraction Transformation Loading
FN – False negative
FP – False positive
GA – Genetic algorithm
GLCM – Grey-Level Co-occurrence
LD – Learning disability
LS-SVM – Least squares support vector machine
OLAP – Online analytical processing
PD – Parkinson’s disease
PLS-LDA – Partial Least Squares-Linear Discriminant Analysis Target
SURF – Speed Up Robust Feature
SVM – Support Vector Machine
SVM-L – Linear kernel
SVM-P – Polynomial kernel
SVM-RBF – Radial basis function
SVR – Support vector regression
TN – True negative
TP – True positive
TQWT – Tunable-Q Wavelet Transform

Conflict of Interest Disclosure

“The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.”

Funding Statement

‘‘The research and publication of this article is funded by the corresponding author. The research did not receive specific funding.”

Data Availability Statement

“ All data used to support the findings of this study are included within the article.The proposed research study comprises with 115 inputs captured from the online website www.datasus.gov.br/datasus/sindex.php”