Migration Letters

Volume: 21, No: S4 (2024), pp. 2015-2027

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)

www.migrationletters.com

Stroke Disease Prediction Using K-Nearest Neighbor And Decision Tree Algorithms With Machine Learning Pre-Processing Techniques

Malak Roman*[1], Ifra Naz[2], Muhammad Ayass Luqman[3], Junaid Ali[4], Mian Sahib Jan[5], Habib Ullah Nawab[6]

*Corresponding Author: malak_5116@uoch.edu.pk

ABSTRACT

Medical professionals require a trustworthy prediction methodology to diagnose stroke patients' data. A vast amount of data regarding patients and their health issues exists. In general, examining data from several perspectives and synthesising it into significant information is called data mining. (sometimes termed data or knowledge discovery). Among the investigative tools accessible for data exploration are data mining packages. Users can categorise, analyse, and summarise the links found in the data from many dimensions or perspectives. One tool for data mining is Weka. It has many machine-learning algorithms. It offers the capability of classifying our data using different algorithms. With many applications, classification is a crucial data mining approach. Data of all kinds are classified by it. In every aspect of our lives, there is[1] *classification. Classification is utilised to place separate items in programmed data into one of a predetermined number of classes or groupings. Many classification algorithms are the subject of our study in this work. Using the Waikato Environment for Knowledge Analysis, the thesis compares various categorisation methods to determine which users are suitable for using haematological data. This paper investigates the application of Decision Trees (J48) and K-Nearest Neighbor (KNN) algorithms to improve medical diagnosis in healthcare. Decision Trees, represented by the J48 algorithm and KNN, are machine-learning techniques used to analyse patient data and assist in medical decision-making. Results of decision tree and k-nearest neighbor algorithm classifiers with genetic search and Chi-Square technique" are summarised. Comparison is based on precision, accuracy, recall, f-Measure and which concluded that, in terms of accuracy, "k-nearest neighbor classifier algorithm" with Genetic Search with 97.5% accuracy. In our study, we tried to find a better and more efficient classifier to classify stroke disease using data mining*

[1] Department of Computer Science, University of Chitral, Khyber Pakhtunkhwa, Pakistan. Email: malak_5116@uoch.edu.pk

[2] Institute of Business and Management Sciences, The University of Agriculture Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: ifranaz1611@gmail.com

[3] North West School of Medicines, Khyber Medical University Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: ayyas.luqman@gmail.com

[4] Kohat Institute Medical Sciences, Khyber Medical University Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: junaidakhoonzada@gmail.com

[5] Department Computer Science, Abasyn University Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: sahibjan5516@gmail.com

[6] Department of Sociology, University of Chitral, Khyber Pakhtunkhwa, Pakistan. Email: habib_soc@yahoo.com

techniques; researchers concluded that KNN is a better algorithm model than the Decision tree algorithm and can be well used to predict strokes in a particular patient.

KEYWORDS: *Data Mining, Knowledge Discovery in Database (KDD), Stroke, Genetic Search, K-Nearest Neighbor (KNN), Decision tree.*

INTRODUCTION: Stroke is a disease that is caused by haemorrhage or obstruction in the blood vessels of the brain, leading to ischemia. This may result in serious neurological injury that may result in sensory loss, cognitive problems, visuospatial dysfunction, aphasia, etc. As mentioned above, the severity of the impairments depends on the extent of the area involved. Stroke is one of the significant reasons behind disability and death. A stroke may be an ischemic stroke, a blood vessel carrying blood to the brain is blocked by a blood clot, or a hemorrhagic stroke due to bleeding inside the brain (a thin, weakened blood vessel or a burst/leak in the brain aneurysm).

Data mining is the process by which automatic experts extract new and valuable information from a vast amount of data. The role and influence of data mining have increased significantly with a timely increase in data. Data mining is a fundamental part of KDD, consisting of a series of transformation steps of data pre-processing to generate results. The primary functions of data mining include classification, association, and clustering (Roman et al., 2022). It entails exploring, choosing, and organising vast amounts of information. This activity has become a regular endeavour across all areas of medical science research. Data mining has led to uncovering valuable, hidden insights from large datasets. Issues related to data mining are typically addressed using a range of computer science techniques, including multi-dimensional databases, data visualisation, machine learning, and statistical methods such as hypothesis testing, classification, clustering, and regression analysis (Aljumah, Ahamad, & Siddiqui, 2013).

Decision Trees create a tree-like structure to classify patients based on their characteristics, while KNN identifies similar patients based on their features and makes predictions accordingly. By leveraging these algorithms, healthcare professionals can effectively interpret patient information, such as symptoms and test results, to diagnose diseases and recommend appropriate treatments. This research aims to demonstrate the utility of Decision Trees and KNN in healthcare settings, potentially leading to more accurate diagnoses, personalised treatment plans, and better patient outcomes.

STROKE DISEASE: Serious consequences may result from a disease in which ischemia in the brain occurs, which is caused by an obstruction or haemorrhage in brain tissues. The disease is known as a stroke and has severe effects on neurological behaviour. The patient suffering from stroke may have sensory loss, memory loss, inability to learn, paralysis of one or both sides of the body, etc. and eventually may result in death as well. The area of the brain involved in stroke affects the severity of the above-mentioned signs and symptoms.

Stroke is a significant factor in causing disability and death. Seventeen million people presented with stroke in 2010, and out of them, 5.9 million expired. Besides, in the same year, 33 million people worldwide survived the stroke attack. Most of the survivors are living from hand to mouth as a result of symptoms of the disease and are considered a great challenge for healthcare providers (Zhang, Liparulo, Panella, Gu, & Fang, 2015).

The underdeveloped countries are still striving for data regarding stroke and its effects on healthcare providers. Most of the patients are present with memory problems. Besides, due to the increase in incidences of stroke, it has now become the third most common reason for death worldwide (Colak et al., 2020).

STROKE DISEASES CLASSIFICATION

The disease can be listed into two main groups.

- Stroke due to Ischemia or Ischemic Stroke:
  Ischemia is decreased or no blood availability to the tissues, which may be due to a clot. This may occur in the brain, and the blood vessels in the brain may clog up.

- Stroke due to haemorrhaging or Hemorrhagic Stroke:
  Hemorrhage is for bleeding. The reason it is called hemorrhagic is because of bleeding in brain tissue, e.g., as in Berry aneurysm.

Literature Review

Roman et al. (2022) proposed a machine-learning model using KNN and Fuzzy-KNN classifiers with Symmetric Uncertainty plus Genetic Search as a data cleaner and preprocessor for cardiovascular patients' predictions. Uses Waikato Environment for Knowledge Analysis tools for data analysis. The result shows that pre-processing techniques have an excellent effect on the algorithms' performance. Between the KNN and Fuzzy-KNN algorithms, the KNN gives 95% results with Symmetric uncertainty (used as a preprocessor).

Saeidi et al. (2021) tried their study on patients' emotional ECGs. ECGs for a total of 6 emotions, i.e., happiness, fear, anger, surprise, sadness, and disgust, were collected. He collected data from patients with Right Brain and left brain damage and from regular patients. Signals were collected and transformed into the time-frequency domain with the help of the Wavelet Packet Transform (WPT) method. A mix of stimuli, such as audio and video programs, induces emotions. Probabilistic Neural Network (PNN) and K-nearest Neighbor (KNN) were two classifier models that were used to compare all the feelings mentioned above and showed a staggering result of 82.32% accuracy for patients with left brain damage (LBD).

Mohamad and Mohamed (2020) explored the role of genetic algorithms (GAs) in feature selection for stroke prediction using data mining techniques. Their study demonstrated how GAs enhance feature selection processes, improving machine learning classifier accuracy in stroke prediction models.

Islam et al. (2017) came up with the idea of highlighting the factors with potential risks before designing a system. The factors that couldn't result in stroke were collected as data sets from a medical institute in Bangladesh. The obtained data was categorised with the help of Fuzzy C-means and Fuzzy Inference System. A fuzzy rule was predicted by adapting the Neuron Fuzzy Interference System to make a better prediction. This system had much better and improved results and is thought to be helpful for experts in the medical field and all populations in general.

Yoon et al. (2017) took account of the Behavioral Risk Factor Surveillance System (BRFSS) and collected a specific data set for it. It was, by then, the most extensive survey on health issues published by The Centers for Disease Control and Prevention (CDC), United States. The data collected, though, was not approved by the institutional board. It was massive data and contained information on about 451075 BRFSS patients. From these patients, 19603 were evaluated and identified as those who had previous attacks of stroke. Data was prepared with the application of SAS, and later, it was analysed with the help of Weka v3.7 to get a model for the association of disability. Patients were categorised with a mean age of 66.5(SD=15.2), and about 62% of the patients found were female. Whites topped the list among the respondents, 76%. Blacks were second on the list with 11%, and in the last 5% were Hispanics.

Arslan et al. (2016) investigated various data mining approaches for predicting ischemic stroke. He collected his data sets from TurgutOzal Medical Center, Inonu University, Malatya, Turkey, and it contained 112 healthy patients, 80 stroke patients, and 17 attributes. Penalised Logistic Regression (PLR), Support Vector Machine (SVM), and Stochastic Gradient Boosting (SGB) methods with ten cross-validations were applied for data mining. The model's performance was counterchecked with the help of Stochastic Gradient Boosting (SGB), Sensitivity, ROC curve Area, and accuracy, showing 97% accuracy.

Krishnaiah et al. (2016) conducted a study to establish the role of data mining techniques in forecasting cardiac diseases and to find a more favourable mining technique for classifying cardiac pathologies. Data mining algorithms compared were Fuzzy-KNN, K-Nearest Neighbor, C4.5, J48, Neuro-Fuzzy, K-means and Neural Network etc., for forecasting cardiac pathologies. Fuzzy-KNN was observed to have better and more accurate results than others.

Sedghi et al. (2016) considered migraine diagnosis and used data mining and text to find the difference between migraine and stroke patients. A heterogeneous technique consisted of patients' numeric attributes like blood pressure, age, etc., accompanied by tirage main complaints and some specialised final impressions. The data set obtained was very imbalanced and consisted of only 6% cases of migraine. Due to its imbalanced nature, the data was hard to tackle, and rearranging it was challenging to retrieve helpful information. 80% sensitivity and 75% specificity were observed, contrary to 15.7% sensitivity and 97% specificity in the case of using the pure imbalanced data set for classifier building.

Zhang et al. (2015) considered cases of autonomous stroke rehabilitation, and a Fuzzy Kernal Motion Classifier was suggested. The major component of this model was the implication of Principal Component Analysis (PCA) on already processed data sets, and the classifier gave 95% accuracy.

Deolekar et al. (2015) used an artificial neural network tabla stroke to propose a classifier model for it. The audio was recorded, and from the file, 13 features were extracted as output units to output layers, whereas 62 were input in input layers. Dimension reduction served as the initial classifier, and later, it was classified without dimension reduction. Principal Component Analysis (PCA) was used for dimension reduction, and 62 to 28 features were reduced. Two sets, each comprising 650 tabla strokes, were used for experimentation. More than 98% accuracy was observed when classifying instances for both cases.

Bhuvaneswari and Therese (2015) went for a hybrid model of the Genetic k-nearest neighbor (GKNN) Algorithm, a genetic algorithm, and k-nearest neighbor for the identification of lung malignancy at very early stages, and MATLAB tools were applied for the implementation of this model. Data available at Rajarajeswari College, Bangalore, was selected and pre-processed. The Gabor filter was utilised for feature extraction. KNN gets the best feature, and a genetic algorithm optimises it. The algorithm then tested cancerous and non-cancerous C.T. lung images of the 5 data sets, and 90% accuracy was observed.

Colak et al. (2015) proposed a classifier model for stroke ailments. Dataset records of 297 patients were obtained from the Emergency Department, Turget Medical Centre, Turkey. Data processing was done with the help of the T2 test, which was based on Cramer's V test, and for classification techniques, SVM and ANN were applied, which showed 84.62% and 85.9% accuracy, respectively.

Beyan and Ogul (2008) used a fuzzy K-nearest method and microarray gene expression data to diagnose cancer. Their study compared the two classification algorithms, i.e., Fuzzy K-nearest and K-nearest neighbors, to cancer diagnosis based on gene expression. Data was collected from gene-system.org. A total of 6 sets were collected for diagnostic purposes. From analysis, Fuzzy K-nearest neighbors (K-NN) were proved to be more effective than the K-nearest neighbor(KNN) classifier with improved accuracy.

Omar et al. (2014) tried a model for predicting acute ischemic stroke. Data available at the Association of Malaysia (NASAM) were collected as test datasets. Three groups were formed, i.e., Early Group (E.G.), Intermediate Group (I.G.), and Advance Group (A.G.), and then data was tested with K-Nearest Neighbor (K-NN) classification algorithm and showed 85% accuracy in the prediction of stroke groups.

Prasertsakul et al. (2014) took data from stroke patients and presented a classification model for rehabilitation. Data available at the Edinburgh University website were selected and processed through the Matlab tool to implement Decision Tree (D.T) and Artificial Neural Networks (ANN). Neural Networks (ANN) was 96.4% accurate, and Decision Tree (D.T) was 98.2 % correct.

On the other hand, Jabbar et al. (2013) used a k-nearest classifier to suggest a method for classifying cardiac pathologies. The evaluator used chi-square, which was also utilised to remove the least ranked attribute. The data obtained was used to apply Chi-squares to the attributes, and the evaluated attributes were further analysed using KNN algorithms. The suggested model proved 95% accurate and had 2.25% improved results compared to other models.

Rajini and Bhavani (2013) used texture features and segmentation to identify ischemic stroke. Rajah Muthiah College Hospital (RMMCH) provided them with C.T. images, each size 512x512. For the model implementation technique, classification algorithms, i.e., Decision Tree (D.T), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were used. Classifier accuracy was 92% for Decision Tree (D.T), 96% for Artificial Neural Networks (ANN), 97% for K-Nearest Neighbors (KNN), and 91% for Support Vector Machine (SVM).

Dangare and Apte (2012) suggested an intelligent system that used patients' historical databases to classify and evaluate cardiac diseases. Different mining techniques, i.e., Neural Networks, Naive Bayes, and decision trees, were used to analyse heart diseases. WEKA tools were used for model production. UCI database provided them with Statlog and Cleveland heart data, which contained information for 270 and 303 patients. Smoking and obesity were added as an additional attribute in their proposed models. The data was pre-processed, and Neural Networks, decision trees, and Naive Bayes classifications were implicated. In their results, Neural Networks proved 1.30% and 0.40% more accurate than Naïve Bayes and Decision Tree, respectively.

Yeh et al. (2011) produced a classifier model that utilised a back propagation neural network, Bayesian classifier, and decision tree. Data regarding cardiac diseases was collected from different areas of Taiwan. After screening the attributes and pre-processing, the newly constructed model was implemented on the WEKA tool. The decision tree model proved good at forecasting cardiac diseases based on accuracy, sensitivity, and classification efficiencies. Researchers estimated 93.59% sensitivity and 93.48% accuracy for the decision tree.

SUMMARY: The literature is rich in studies and research on the implementation of decision tree algorithms and k-nearest neighbors, and it comprehensively gives us an idea about the use of these algorithms in predicting stroke in certain patients. We have seen that genetic search is significant in selecting the best attributes to give us better results for our data. However, it is also observed that the pre-processing data techniques are applied to the results of decision tree algorithms and KNN, and their accuracy can be further improved. From going through the literature in detail, we can conclude that data mining techniques can prove to be very helpful in making a particular decision while performing a test to diagnose patients with stroke.

RESEARCH METHODOLOGY: The WEKA (The Waikato Environment for Knowledge Analysis) tool is used to construct a model for stroke disease forecasting. Stroke disease data were collected from Hayatabad Medical Complex Hospital in Peshawar. The dataset is in Microsoft Excel for WEKA readable format. After data pre-processing, two datasets were made, one for training and the second for model testing. Genetic search and the Chi-Square process are used separately for best and optimal feature selection in pre-processing. After loading the dataset into WEKA, to train the model, decision tree (D.T.) and K-nearest neighbor (KNN) classification algorithms are applied individually to check and test the accuracy of the models. The model was trained first, and then test data was used to calculate the accuracy of the proposed model.

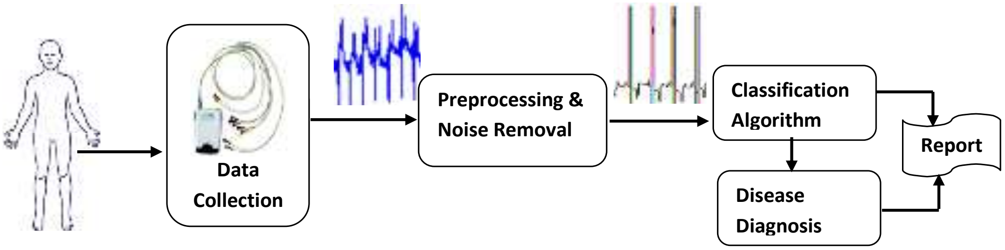


Figure: Architecture of Research Methodology (Roman et al., 2022).

STROKE DATA COLLECTION: The stroke disease patient's data is collected from Hayatabad Medical Complex Hospital Peshawar, Pakistan. The dataset consists of 12 attributes. The dataset includes age, gender, blood pressure, fasting blood sugar, diabetes, serum cholesterol, smoking, family history, arrhythmias, FND (Focal Neurological Deficit), stroke type, and Class Label.

PRE-PROCESSING: Data reliability is enhanced in this stage. Data cleaning, integration, reduction, and removal of outliers' tasks are performed at this stage.

GENETIC SEARCH: The adaptive heuristic search algorithms known as genetic algorithms (GAs) are founded on crossover, mutation, and selection in natural evolution. Because GAs are frequently employed for optimisation tasks in high-dimensional, complicated data settings, they are ideal for use in healthcare applications such as stroke risk prediction.
Mathematical Representation:

- **Fitness Function:** $f(x)$
- **Selection Probability:** $P(x) = f(x) / \Sigma f(x)$
- **Crossover Operator:** child = α * parent1 + (1-α) * parent2 (α is a random number between 0 and 1)
- **Mutation Operator:** mutated_gene = gene ± random_value

By iteratively applying these steps, genetic algorithms can effectively explore the solution space and find high-quality solutions to complex optimisation problems (Jin et al., 2006).

A GA can optimise the feature selection process in stroke prediction by determining the most essential elements from an extensive patient data collection, such as demographics, medical history, lifestyle factors, and biomarkers. Researchers and medical professionals can determine which feature combination produces the best-predicted accuracy for stroke by employing a GA.

TABLE I: ATTRIBUTE SELECTION BY GENETIC SEARCH

S. No.	**Attribute**
1	Gender
2	Blood Pressure
3	Cholesterol
4	Find (Focal Neurological Deficit)
5	Stroke Type

CHI SQUARE: The chi-squared test, also known as “Pearson’s chi-squared test,” is a statistical hypothesis test used when a test statistic’s sample distribution is chi-squared. Squared errors or sample variance are often used to construct a chi-squared test. Test statistics based on the assumption of independent customarily distributed data follow a chi-squared distribution, which the Central Limit Theorem also supports (Saeidi et al., 2021). This test can also be applied to rejecting a null hypothesis to prove that the data are independent.

A chi-squared test also implies that by making the sample size large enough, the sampling distribution can approximate a chi-squared distribution as closely as assumed or desired if the null hypothesis is true. Also, this test can be helpful if a significant difference exists between the observed frequencies and the expected frequencies in one or more variables.

TABLE II: ATTRIBUTE SELECTION BY CHI-SQUARE

S. No.	**Attribute**
1	FND (Focal Neurological Deficit)
2	Stroke Type
3	Blood Pressure
4	Cholesterol
5	Family History
6	Gender
7	Smoke
8	Fasting Blood Sugar

CLASSIFICATION: Creating a model capable of classifying a group of items to predict their class labels or the characteristics of unknown future objects is referred to as classification. The classification method presents a model that can efficiently and swiftly determine the class of items with unknown labels. The testing data is compared or analysed with the training or sample data (having class labels in advance). Algorithms like neural networks, Bayesian classification, Decision trees, K-Nearest neighbor, etc., are used in classification (Han & Kamber, 2006).

K-NEAREST NEIGHBOR: The k-nearest neighbor algorithm is also known as the lazy learning classification algorithm. The groundwork of this process lies in contrasting the given sample or test data with the training data. This comparison is executed using a similar approach. The trained dataset is assessed against the sample or test data to analyse them. When unfamiliar data is given, the K-NN classifier searches for a similar data item or tuple in the expert / trained data set. K-nearest classifier finds the next-door and nearby record and transfers that nearest record class label to the unfamiliar tested record. This exact tuple is the adjacent neighbor for the unknown data item (Han & Kamber, 2006).

Many distance algorithms are used; these are:

- Manhattan distance formula d= $\sum_{i=1}^{n} |X_i - Y_i|$
- Euclidean distance formula d= $\sqrt{\sum_{i=1}^{n} (Xi - Yi)^2}$

Several distance formulas, including t wski, and Euclidean distance formulas, are used to calculate the closeness in terms of the Cartesian distance metric. The most popular application of the Euclidean distance formula is to determine the separation between two points (the training tuple and the sample). The following Euclidean distance formula is used to determine the distance between the first and second instances, assuming that the first instance points are (a1, a2, - - - an) and the second instance points are (b1, b2, - - - bn) (Han & Kamber, 2006).

Euclidean distance = $\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots\ldots\ (a_n - b_n)^2}$

DECISION TREE: The decision tree approach is a well-liked data mining technique for developing classification schemes centred on several variables or forecasting algorithms for a target variable. The tree diagram schematically determines the behaviour or provides statistical probability. This approach divides an inhabitant (population) into branch-like portions, resulting in an inverted tree with leaf, internal, and root nodes. Imagine you have patient data with features like age, blood pressure, smoking status, and cholesterol level. These factors can influence the likelihood of a stroke.

Entropy(S) = $-\sum p_i \log_2(p_i)$, Where p_i is the probability of each class (e.g., stroke or no stroke).

Information Gain (S, Feature) = $\text{Entropy}(S) - \sum (|S_j| / |S|) * \text{Entropy}(S_j)$

S is the original dataset, and S_j is the subsets after a split (Han et al., 2011).

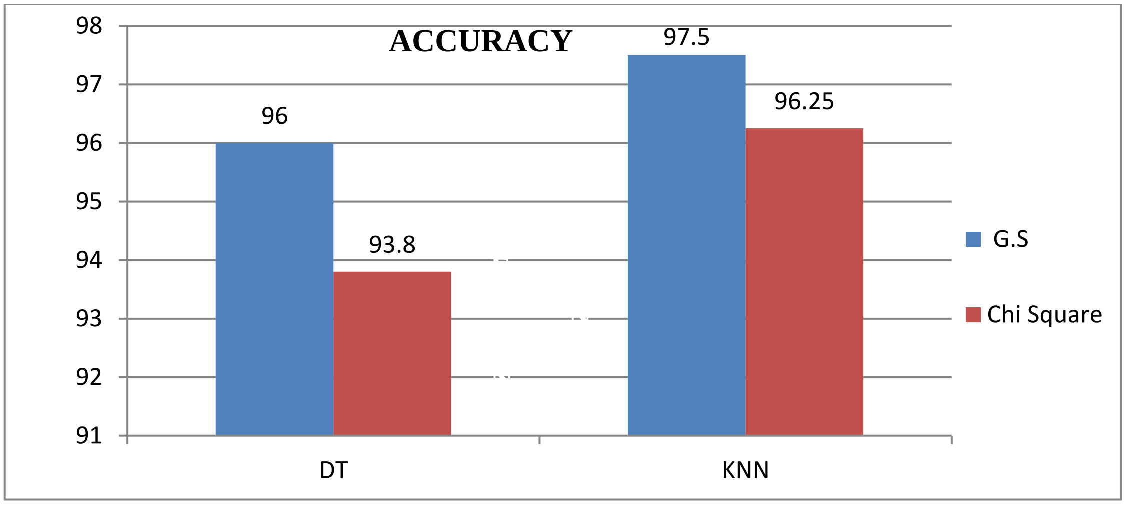
RESULTS: Results of "decision tree and k-nearest neighbor algorithm classifiers with genetic search and Chi-Square technique" are summarised. Comparison is based on precision, accuracy, recall, f-Measure, and ROC curves, which concluded that, in terms of accuracy, the "k-nearest neighbor classifier algorithm" with Genetic Search with 97.5% accuracy, 97.5% precision,97.5% recall, and 97.5% f-measure, is far better than "decision tree algorithm."

TABLE III: COMPARISON OF RESULTS BOTH OF DECISION TREE AND K-NEAREST NEIGHBOR WITH GENETIC SEARCH AND CHI-SQUARE

Algorithms	Pre-processing	Accuracy	Precision	Recall	F-Measure	ROC Area
Decision Tree	Genetic Search	96 %	0.963	0.963	0.967	0.974
	Chi-Square	93.75 %	93.8	93.8	93.8	93.3
KNN	Genetic Search	97.5%	0.975	0.975	0.975	0.995
	Chi-Square	96.25 %	96.2	96.3	96.2	94.8

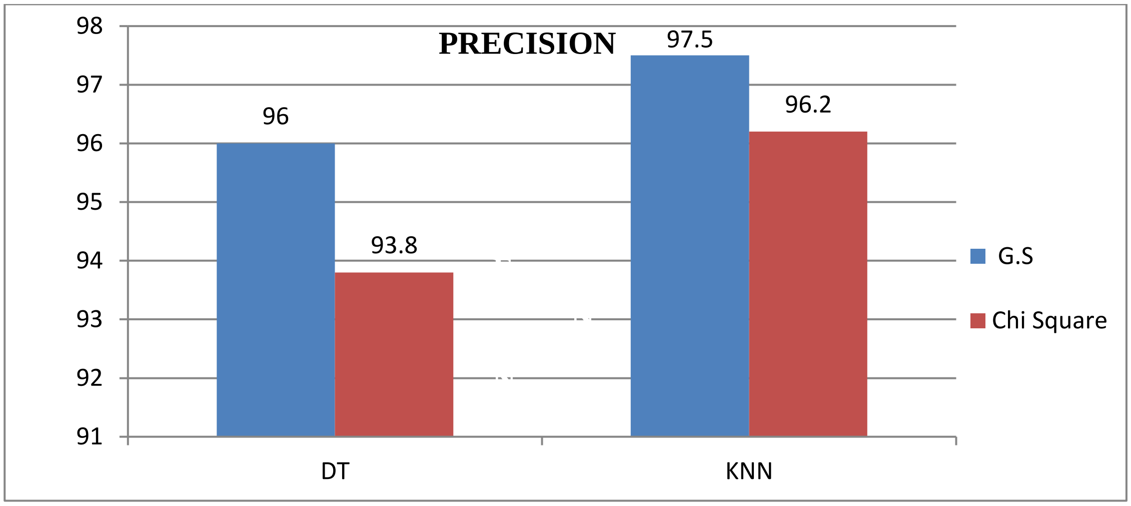
ACCURACY: Shows graphical results of “decision tree and k-nearest neighbor algorithm classifiers with genetic search and Chi-Square technique in terms of accuracy, which shows that k-nearest neighbor algorithm with the genetic search performed better with the accuracy of 97.5% than decision tree with and without using genetic search.

FIG I: COMPARE RESULTS OF DT & KNN WITH G.S AND CHI-SQUARE IN TERM OF ACCURACY



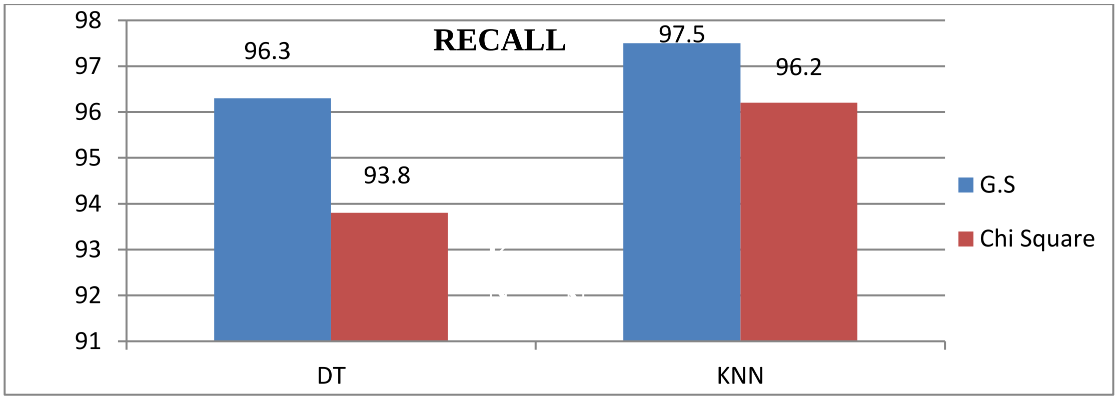
PRECISION: shows graphical results of “decision tree and k-nearest neighbor algorithm classifiers with genetic search and Chi-Square technique in terms of accuracy, which shows that k-nearest neighbor algorithm with the genetic search performed better with the accuracy of 97.5% than decision tree with and without using genetic search.

FIG II: COMPARE RESULTS OF DT & KNN WITH G.S AND CHI-SQUARE IN TERMS OF PRECISION



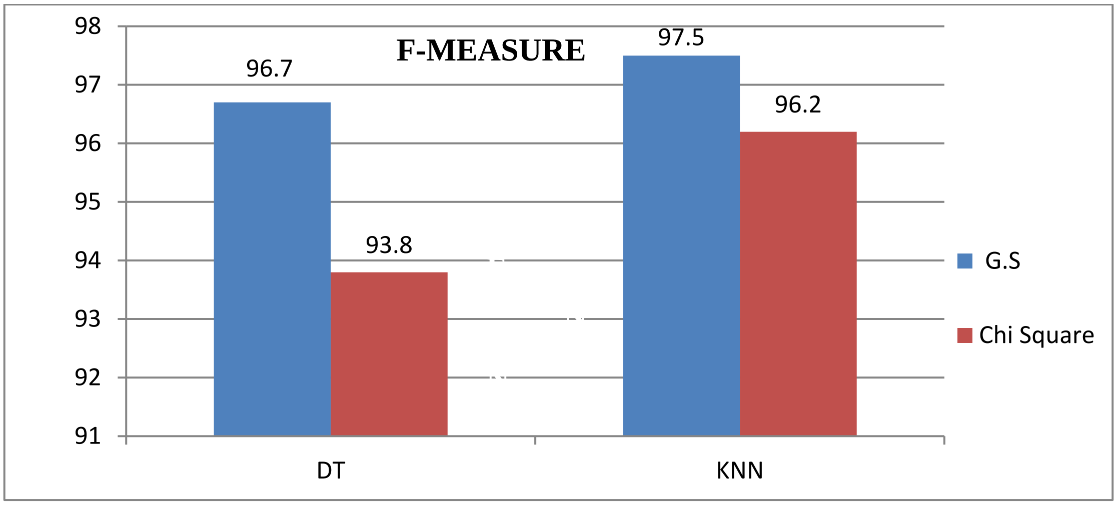
RECALL: Shows graphical results of “decision tree and k-nearest neighbor algorithm classifiers with genetic search and Chi-Square technique in terms of recall, which shows that k-nearest neighbor algorithm with the genetic search performed better with the accuracy of 97.5% than decision tree with and without using genetic search.

FIG III: COMPARE RESULTS OF DT & KNN WITH G.S AND CHI-SQUARE IN TERMS OF PRECISION



F-MEASURE: shows graphical results of “decision tree and k-nearest neighbor algorithm classifiers with genetic search and Chi-Square technique in terms of accuracy, which shows that k-nearest neighbor algorithm with the genetic search performed better with the accuracy of 97.5% than decision tree with and without using genetic search.

FIG IV: COMPARE RESULTS OF DT & KNN WITH G.S AND CHI-SQUARE IN TERMS OF F-MEASURE



CONCLUSION

The above-presented experiment and its results state that the k-nearest neighbor classifier and the decision tree algorithm with genetic search give much better results than the same algorithms with Chi-Square and are best defined in precision, accuracy, and recall. Though simple, the k-nearest neighbor classifier algorithm performs far better than other advanced classification algorithms. Our results also showed that the k-nearest neighbor classifier is more accurate and faster than the decision tree. In our study, we tried to find a better and more efficient classifier to classify stroke disease by using data mining techniques. We concluded that KNN is a better algorithm model with genetic search (preprocessor) than the Decision tree algorithm and can be well used to predict stroke in a particular patient.

RECOMMENDATIONS

1. Considering the results of our studies, it can be stated that the suggested model can be extended for the evaluation and classification of other commonly prevalent diseases related to the heart, kidneys, liver, etc.
2. A hybrid model can also be made to compare different classifiers.
3. Further improvement can be incorporated by designing automated software for easy data handling.
4. The results, though, reveal more efficiency and more accuracy than previously used models. However, more work in this area is needed, and clustering may further be incorporated, in the future, into the proposed model to achieve speedier processing of data and increase accuracy.

References

1. Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences, 25(2), 127–136. https://doi.org/10.1016/j.jksuci.2012.10.003
2. Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. Computer Methods and Programs in Biomedicine, 130, 87–92.

3. Beyan, C., & Ogul, H. (2008, January). A fuzzy k-NN approach for cancer diagnosis with microarray gene expression data. Paper presented at the International Symposium on Health Informatics and Bioinformatics, Istanbul, Turkey.
4. Bhuvaneswari, P., & Therese, A. B. (2015). Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm. Procedia Materials Science, 10, 433–440. https://doi.org/10.1016/j.mspro.2015.06.077
5. Colak, C., Karaman, E., & Turtay, M. G. (2015). Application of knowledge discovery process on the prediction of stroke. Computer Methods and Programs in Biomedicine, 119(3), 181–185.
6. Çolak, T., Yencilek, H. İ., Kalaycıoğlu, O., Çelik, K., & Tekten, B. Ö. (2020). Evaluation of Patients Diagnosed as Having Acute Stroke in the Emergency Department: Two-year Analysis. Turkish Journal of Neurology, 26(2), 142–148. https://doi.org/10.4274/tnd.2020.92231
7. Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44–48.
8. Deolekar, S., & Abraham, S. (2015). Classification of Tabla Strokes Using Neural Network. In Advances in intelligent systems and computing (pp. 347–356). https://doi.org/10.1007/978-81-322-2734-2_35
9. Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed.). Morgan Kaufmann.
10. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.
11. Islam, F., Shoilee, S. B. A., Shams, M., & Rahman, R. M. (2017). Potential risk factor analysis and risk prediction system for stroke using fuzzy logic. In Artificial intelligence trends in intelligent systems: Proceedings of the 6th Computer Science Online Conference 2017 (CSOC2017) (Vol. 16, pp. 262–272). Springer International Publishing. https://doi.org/10.1007/978-3-319-57261-1_26
12. Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart disease classification using nearest neighbor classifier with feature subset selection. Annals. Computer Science Series, 11(1), 47–53.
13. Jin, X., Xu, A., Bie, R., & Guo, P. (2006). Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In J. Li, Q. Yang, & A. H. Tan (Eds.), Data mining for biomedical applications. BioDM 2006. Lecture notes in computer science (Vol. 3916). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11691730_11
14. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2016). Heart disease prediction system using data mining techniques and intelligent fuzzy approach: A review. International Journal of Computer Applications, 136(2), 43–51.
15. Mohamad, H., & Mohamed, E. (2020). A genetic algorithm approach to feature selection in stroke prediction using data mining techniques. Procedia Computer Science, 170, 391–398.
16. Omar, W. R. W., Azman, A., Abdullah, S. A., & Nor, F. M. (2014). Brainwave classification for acute ischemic stroke group level using k-NN technique. In Proceedings of the 2014 5th International Conference on Intelligent Systems, Modelling and Simulation (pp. 117–120). IEEE. https://doi.org/10.1109/ISMS.2014.26
17. Prasertsakul, T., Kaimuk, P., & Charoensuk, W. (2014). Defining the rehabilitation treatment programs for stroke patients by applying neural network and decision trees models. In Proceedings of the 7th 2014 Biomedical Engineering International Conference (pp. 1–5). IEEE. https://doi.org/10.1109/BMEiCON.2014.7017422
18. Rajini, N. H., & Bhavani, R. (2013). Computer aided detection of ischemic stroke using segmentation and texture features. Measurement, 46(6), 1865–1874. https://doi.org/10.1016/j.measurement.2013.01.010
19. Roman, M. R., Nawab, H. U. N., Ahmad, S. M., Zaib, A. Z., Khan, N. W. K., Jan, M. S. J., Rahman, M. A. R., & Khan, I. A. K. (2022, November 1). K-Nearest Neighbor and Fuzzy K-Nearest Neighbor Algorithm Performance Analysis for Heart Disease Classification. Webology, 19(1), 8607–8619.
20. Saeidi, M., Karwowski, W., Farahani, F. V., Fiok, K., Taiar, R., Hancock, P. A., & Al-Juaid, A. (2021). Neural Decoding of EEG Signals with Machine Learning: A Systematic Review. Brain Sciences, 11(11), 1525. https://doi.org/10.3390/brainsci11111525
21. Sedghi, E., Weber, J. H., Thomo, A., Bibok, M., & Penn, A. M. (2016). A new approach to distinguish migraine from stroke by mining structured and unstructured clinical data sources. Network Modeling Analysis in Health Informatics and Bioinformatics, 5, 1–11.
22. Yeh, D., Cheng, C., & Chen, Y. (2011). A predictive model for cerebrovascular disease using data mining. Expert Systems With Applications, 38(7), 8970–8977. https://doi.org/10.1016/j.eswa.2011.01.114

23. Yoon, S., Patrao, M., Schauer, D., & Gutierrez, J. (2017). Prediction models for burden of caregivers applying data mining techniques. Big Data & Information Analytics, 2(3), 209–217. https://doi.org/10.3934/bdia.2017014
24. Zhang, Z., Liparulo, L., Panella, M., Gu, X., & Fang, Q. (2015). A Fuzzy Kernel Motion Classifier for Autonomous Stroke Rehabilitation. IEEE Journal of Biomedical and Health Informatics, 20(3), 893–901. https://doi.org/10.1109/jbhi.2015.2430524