International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ADVANCED HEALTHCARE SOLUTION TO PREDICT STROKE WITH MACHINE LEARNING APPROACHES

[1*]R. Muthulakshmi, [2]Dr V. Vasudevan

[1*]Research Scholar, Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil-626126, Tamil Nadu, India, skclakshmirml@gmail.com

[2]Senior Professor, Department of Computer Science & Engineering, Kalasalingam Academy of Research and Education, Krishnankoil-626126, Tamil Nadu, India, drvasudevan.srivi@gmail.com

ABSTRACT

An abrupt cessation of blood flow to a specific region of the brain induces a stroke. In the absence of a blood supply, cerebral neurons progressively perish, leading to impairments that are specific to the affected region of the brain. The timely identification of symptoms can yield substantial insights that are instrumental in prognosticating stroke and advocating for overall well-being. This research paper utilizes Machine Learning (ML) to construct and assess multiple models that contribute to the development of a resilient framework for predicting the long-term risk of stroke. Majority voting technique that achieves high performance, as validated by numerous metrics including precision, recall,F-measure, and accuracy, is the primary contribution of this study. Majority voting classification outperformed the alternatives, as demonstrated by the experiment results, which included an F-measure (98.09%), precision (98.4%), and recall (97.8%), and an accuracy of 98.1%.

Keywords – Stroke prediction, Machine Learning, healthcare, classification.

INTRODUCTION

Stroke prevalence has been on the rise worldwide, making it a prominent contributor to mortality and impairment. Timely intervention is essential in reducing the long-term disability and mortality linked to stroke. Conventional approaches to forecasting the likelihood of a stroke, on the other hand, can require a significant amount of time and are susceptible to mistakes. Machine Learning algorithms have recently demonstrated significant potential in properly forecasting the likelihood of stroke by analyzing several clinical risk factors. By utilizing these algorithms, medical professionals can detect patients at a greater risk and take action at an early stage, thereby decreasing the occurrence of stroke-related problems and enhancing patient results.

Furthermore, there is an increasing demand for clarity and comprehensibility in machine learning models used in the healthcare industry. Utilizing an interpretable machine learning model can offer doctors significant insights into the elements that contribute to a patient's risk of stroke, thereby

Received: August 01 ,2025

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

assisting in making treatment decisions. According to the World Stroke Organisation, over 13 million individuals worldwide suffer from a stroke annually, resulting in 5.5 million deaths [1].

Stroke has a profound impact on several elements of a patient's life, encompassing their family, social surroundings, and employment. It is also recognized as one of the leading causes of death and disability worldwide [1], [2]. There is a widespread misunderstanding that stroke only affects specific demographics, such as the elderly or individuals with pre-existing health conditions. Indeed, anyone can be affected, irrespective of their age, gender, or physical well-being. A stroke is an acute and severe interruption in the circulation of blood to the brain, resulting in a deprivation of oxygen to brain cells.

There are two types of this condition: ischemic and hemorrhagic. Strokes of moderate to severe intensity might result in either permanent or temporary harm, contingent upon their severity. Hemorrhagic strokes are few; yet, they occur due to the rupture of a blood vessel in the brain. The predominant kind of stroke occurs when an artery becomes obstructed or constricted, impeding the circulation of blood to the brain [3,4]. Individuals who are over the age of 55, have previously experienced a stroke or transient ischemic attack (TIA), have arrhythmia, high blood pressure, carotid stenosis caused by atherosclerosis, smoke, have high blood cholesterol, diabetes, obesity, are physically inactive, use estrogen therapy, have blood clotting disorders, use cocaine or amphetamines, or have heart conditions such as infarction and cardiac arrest, are all at an increased risk for stroke [5-7].

Strokes can manifest abruptly, and the symptoms they produce can vary and catch one by surprise. The primary manifestations of a stroke encompass unilateral paralysis, facial, arm, or leg numbness, speech and gait impairment, vertigo, impaired vision, cephalalgia, emesis, facial drooping, and, in extreme instances, unconsciousness and coma. These sensations can occur abruptly or gradually, and in certain uncommon instances, they might lead to heightened consciousness [8-10]. Stroke can have a detrimental effect on both males and females, reducing their overall well-being and placing a burden on public health resources. The scientific community places great importance on developing predictive models for strokes in order to prevent them, and artificial intelligence (AI) plays a crucial role in this effort as it is widely used for illness prevention. Multiple studies have been conducted to develop models for diagnosing strokes [11-13], predicting treatment outcomes and patient responses, and designing personalized rehabilitation strategies [14-16].

This research work presents a binary classification model for stroke prediction based on ML.This work recognizes that class balancing is extremely important for better prediction of stroke disease and employs Synthetic Minority Over-sampling Technique (SMOTE) [28] for achieving the same. This work utilizes classifiers such as k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), Naïve Bayes (NB), MultiLayer Perceptron (MLP), Decision Tree (DT), Stochastic Gradient Descent (SGD) and Random Forest (RF). Additionally, stacking and majority voting techniques are also employed for further analysis.

The remaining portion of the paper is structured in the following manner. Section 2 provides an overview of the pertinent literature related to the subject being discussed. Next, Section 3 provides a detailed explanation of the dataset and presents an analysis of the approach that was used. Furthermore, Section

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4 provides a detailed account of the experimental configuration and a comprehensive analysis of the obtained study findings. Section 5 provides an outline of the conclusions and future directions.

Related Works

The authors of [17] obtain a stroke dataset from the Sugam Multispecialty Hospital in India and employ mining and machine learning techniques to categorize the different types of strokes. The support vector machine and ensemble (bagged) models achieved an accuracy of 91%, however the artificial neural network trained using the stochastic gradient descent approach outperformed other algorithms, with a classification accuracy exceeding 95%.

In [18], the authors employ a modified random forest algorithm to make predictions for strokes. This was utilized to assess the levels of risk associated with strokes. According to the authors' suggestion, this approach is reported to have outperformed the existing algorithms. This specific research is limited in its applicability to only a select few types of strokes and cannot be extended to encompass any novel stroke forms that may arise in the future.

The authors of study [19] investigated the efficacy of machine learning (ML) techniques in accurately identifying patients who are suitable candidates for thrombolysis within the recommended time frame. This identification was based on the analysis of diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) magnetic resonance imaging. Three models were created utilizing three distinct machine learning algorithms: linear regression, support vector machine, and random forest. These models aimed to recognize strokes based on sensitivity and specificity within a given time frame. The sensitivity and specificity of ML algorithms outperformed human readers by a large margin. However, as the approach was only evaluated on a small amount of patients' data, the projected time it takes to identify a stroke may vary.

Jenna R.S. and Dr. Sukesh Kumar [20] conducted a study where they evaluated the performance of several kernel functions in the Support Vector Machine approach for stroke prediction. They concluded that the Support Vector Machine algorithm demonstrated excellent performance. The Kernel Linear Function gave the highest results in the experiments, achieving an accuracy of 91.7%, followed by the Polynomial function with an accuracy of 89.0%. However, this research has only utilized a limited amount of datasets, specifically 350 datasets for both training and testing purposes.

In [22], the authors utilized four machine learning algorithms to examine an individual's physical condition and medical report data in order to ascertain the specific type of stroke that may be imminent or has already taken place. They possessed a substantial assortment of medical records that they utilized to resolve their problem. The classification outcome indicated that the result was satisfactory and suitable for use as a real-time medical report. They hypothesized that machine learning algorithms could enhance comprehension of ailments and serve as an exceptional healthcare collaborator. However, the distribution of the dataset among all classes was not perfectly symmetrical, which could potentially lead to inaccuracies when trying to differentiate between stroke rates based on the number of patients.

Received: August 01 ,2025

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In a study conducted by Adam et al. [22], the objective was to determine the optimal method for categorizing ischemic strokes. The classification of ischemic stroke was performed using two models: a K-Nearest Neighbor method and a decision tree algorithm. Their study revealed that the decision tree algorithm was more user-friendly for medical professionals.

The authors employ a Machine learning system in [23] to forecast occurrences of strokes. This work utilized many physiological parameters to train five distinct models with the aim of achieving precise prediction. The algorithms used in this analysis are Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes Classification. The Naive Bayes Classification achieves superior performance, attaining an accuracy rate of 82%. According to the author, the research will utilize Neural Network training to broaden its coverage.

Harshitha et al. [24] evaluated the performance of five distinct machine learning techniques in predicting the probability of stroke. The models employed were Random Forest, Logistic Regression, K Nearest Neighbor, Decision Tree, and Support Vector Machine. The Random Forest model achieved the highest accuracy rate of 95.5% and was selected as the top model because to its exceptional accuracy and little occurrence of false negatives.

Dev et al. [25] conducted a study to identify crucial factors for predicting strokes using statistical approaches. The performance of neural networks, decision trees, and random forests was assessed in three different scenarios: using the original features, using PCA-transformed data from the first two principal components, and using PCA-transformed data from the first eight components. The study identified the key indicators for diagnosing stroke, and after testing various methods, it was found that a perceptron neural network with four crucial attributes achieved the best accuracy rate and the lowest rate of misdiagnosis.

In [26], the authors employed the Healthcare Dataset Stroke to evaluate the effectiveness of distributed machine-learning systems in forecasting stroke. The stroke prediction model was constructed with Apache Spark, a renowned big data platform, in combination with the MLlib package. The classification techniques utilized were Decision Tree, Support Vector Machine, Random Forest Classifier, and Logistic Regression. In order to enhance the results, the process of cross-validation and hyperparameter optimization were implemented. The Random Forest Classifier outperformed the other models, achieving a 90% accuracy across all performance metrics including Accuracy, Precision, Recall, and F1-measure.

The authors of the paper [27] examine the difficulties and inherent prejudices of deep learning algorithms in the field of medical picture analysis. The authors suggest tactics to enhance the comprehensibility and reliability of these algorithms, such as employing visualization approaches, feature attribution methods, and interpretable models. The essay offers useful perspectives on the significance of guaranteeing transparency and comprehensibility of deep learning algorithms in medical picture analysis for both medical professionals and patients.

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Motivated by these existing stroke prediction techniques, this article aims to present a ML based stroke prediction technique that could prove minimal False Positive (FP) and False Negative (FN) rates. The proposed work is elaborated in the following section.

PROPOSED MACHINE LEARNING BASED STROKE PREDICTION SYSTEM

The primary contribution of this research is the utilization of diverse algorithms on a publicly accessible dataset (obtained from the Kaggle website) to compare and determine the most effective strategy for predicting the occurrence of a stroke. This study utilizes machine learning algorithms as classification tools to predict the presence of stroke disease based on several associated factors.

Various performance metrics, including accuracy, precision, recall, F-1 score, and AUC curve, are calculated and compared among different classification models to see which one provides more accurate predictions on the dataset. The experimental results of the proposed approaches were compared with the previous works to demonstrate the originality of the work. The complete work is organized into three phases, such as data collection, data pre-processing, description of machine learning models and class label prediction. All these phases are presented in the forthcoming sections.

3.1 Data Collection

The study was conducted utilizing the stroke prediction dataset accessible on the Kaggle website [29]. The dataset consists of 5110 rows and 12 columns. The dataset has 11 characteristics and 1 target variable. The stroke of the output column is either 1 or 0. A stroke risk assessment was conducted upon detecting the value 1, whereas no such assessment was conducted when the value 0 was presented. We narrowed our focus on participants who were 18 years of age or older from this dataset. The total number of participants was 3254. Detailed descriptions of all attributes (10 for input to ML models and 1 for the target class) are provided below.

- Age (in years): This attribute denotes the participants' age in years who have surpassed the minimum age of 18.
- Gender: This attribute pertains to the gender of the participant. 1984 constitutes the female population, while 1260 males are present.
- Hypertension: This attribute denotes the presence or absence of hypertension in the participant. 12.54 percent of the participants are diagnosed with hypertension.
- Heart disease status: This attribute indicates whether or not the individual in question is afflicted with heart disease. 6.33 percent of the participants are afflicted with cardiovascular disease.
- This attribute signifies the marital status of the respondents, of which 79.84% are presently married.
- Work type comprises four categories that represent the work status of the participant: private (75.02%), self-employed (19.21%), government-employed (15.67%), and never-worked (0.1%).
- The residence type attribute denotes the residing status of the participant and is divided into two categories: rural (48.86%) and urban (51.14%).

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Average glucose level (mg/dL): This parameter measures the average glucose level of the participant.
- The body mass index (Kg/m2) is a metric utilized to determine the weight of the participants.
- Smoking Status: This attribute records the smoking status of the participant and is divided into three categories: current smokers (22.37%), never smokers (52.64%), and former smokers (24.90%).
- Stroke: This attribute indicates whether or not the participant has experienced a stroke in the past. 5.53 percent of the participants have experienced a cerebrovascular accident.

The dataset consists of twelve attributes, as stated previously. Initially, the column id is disregarded as its presence does not influence the formation of the model. The dataset is then examined for null values and populated accordingly. The null values in the BMI column are populated with the "most frequent" value from the data column. Label encoding transforms the string literals in the dataset into integer values that are comprehensible to the computer. The strings must be converted to integers due to the fact that computers are generally programmed to process numerical data.

The acquired dataset comprises five columns, namely smoking status, work type, ever married, gender, and residence type, each containing data of the string variety. All strings are encoded and the entire dataset is converted to a set of numbers during label encoding. The dataset used to forecast stroke is notably unbalanced. The -e dataset comprises 5110 cells, of which 249 indicate the probability of a stroke and 4861 demonstrate its absence. Although the use of such data to train a model at the machine level may result in improved accuracy, alternative metrics of accuracy such as recall and precision are inadequate. Failure to properly manage such uneven data would result in erroneous conclusions and render the forecast useless. Hence, prioritizing the resolution of this irregular data is imperative for the development of an effective model.

To rectify the inequitable distribution of participants between the stroke and non-stroke groups, SMOTE [28] was implemented. To be more precise, the'stroke' minority class was oversampled to ensure that the distribution of participants was uniform. Furthermore, no missing or null values were present, so neither data imputation nor discarding was performed. The significance of features is a fundamental element in classification analysis that enables the construction of precise and high-fidelity machine learning models. The classifiers' accuracy increases until the optimal number of features is taken into account.

It is possible for the performance of machine learning models to decline when irrelevant features are incorporated during the training process. The act of feature ranking entails the allocation of a numerical value to every feature present in a given dataset. This approach prioritizes the most significant ones, specifically those that have the potential to make a substantial contribution to the objective variable and improve the accuracy of the model, which is achieved by Information Gain Ratio (IGR).

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

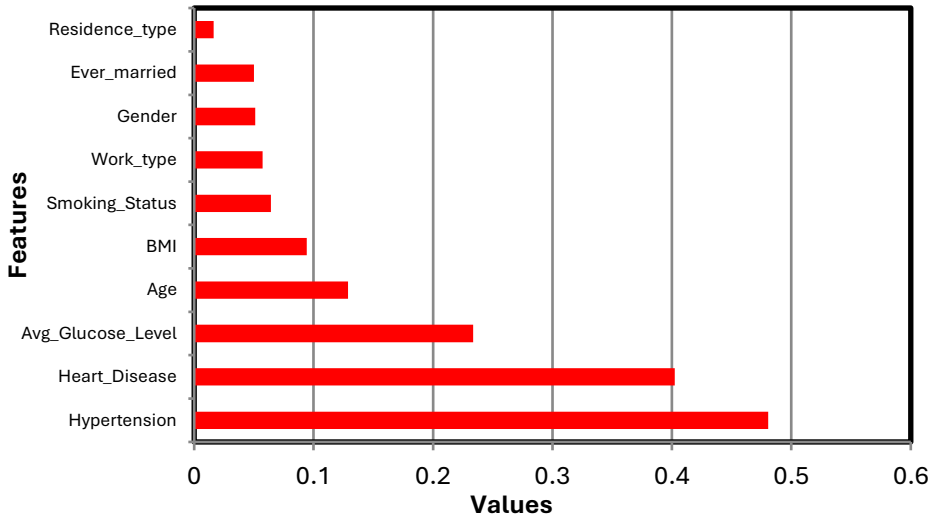

Fig.2. Ranked features by IGR

IGR focuses on non-zero values and is an enhancement of Information Gain (IG). The main merit of IGR is that it takes all the distinct values into account. The required information for value classification is calculated by

$$IG_{DS} = -\left[\frac{o.f(cl_1,DS)}{|DS|}\right] log_2 \left[\frac{o.f(cl_1,DS)}{|DS|}\right] \quad (1)$$

Where $\left[\frac{o.f(cl_1,DS)}{|DS|}\right]$ is the probability of the occurrence frequency of a sample, which is a part of class cl_1. Consider that a feature Ft consists of x number of different values $\{Ft_1, Ft_2, Ft_3 \ldots . Ft_x\}$, such that the training dataset can be divided into x subclasses $\{cl_1, cl_2, cl_3 \ldots . cl_x\}$. In this case, x is 2 and the information gain of a feature is calculated by

$$IG(Ft) = \left[\frac{|Ft_i|}{|Ft|}\right] \times IG(Ft_i) \quad (2)$$

The information gain ratio is calculated as follows

$$IGR(Ft) = \left[\frac{IG_{DS} - IG(Ft)}{IG_{DS} + IG(Ft)}\right] \times 100 \quad (3)$$

The IGR of all features are calculated by the above equation and the features are ranked by IGR. The most important features of this dataset are found to be heart_disease and hypertension and the least important features are residence type and ever_married. The ranks attained by the features are represented in graphical format in figure 2. As observed from figure, all the features prove a positive score such that the ML model can perform effectively in class prediction.

3.2 ML Models

The utilized ML models for stroke prediction are explained in this section.

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2.1 k-NN

The K-NNs classifier is a method that calculates the similarity or difference between two occurrences in the dataset being analyzed using distance metrics such as Euclidean or Manhattan. The Euclidean distance is the most straightforward and often employed method. Let ft_{new}be the feature vector of a new sample that will be identified as either a stroke or non-stroke candidate. The KNN classifier identifies the 'K' vectors that are closest to ft_{new}. Next, the value of ft_{new} is allocated to the class that the majority of its neighbors belong to.

3.2.2 SVM

A Support Vector Machine (SVM) is a supervised learning algorithm that classifies unknown data based on labelled data [3]. The expression of decision boundaries in this context relies on the utilization of decision planes or hyperplanes [19]. A hyperplane is employed to partition the set of data objects into distinct classes. The Radial Basis Function (RBF) kernel was utilized with a parameter value of 1. The SVM algorithm aims to classify data by constructing a function that accurately assigns each data point to its correct label while minimizing errors and maximizing the margin between different classes.

3.2.3 NB

NB renders greater probability, in case of independent features. The probability of class 'Cl' is increased, when a new data sample 'DS' with feature vector 'ft_i' is classified, which is shown in eqn.(4).

$$PR(Cl|ft_{i1}, ft_{i2}, \dots ft_{in}) = \frac{PR(ft_{i1}, ft_{i2}, \dots ft_{in}|Cl)PR(Cl)}{PR(ft_{i1}, ft_{i2}, \dots ft_{in})} \quad (4)$$

$$PR(ft_{i1}, ft_{i2}, \dots ft_{in}|Cl) = \prod_{j=1}^{n} PR(ft_{ij}|Cl) \quad (5)$$

The prior probabilities of features and class are given by $PR(ft_{i1}, ft_{i2}, \dots ft_{in})$ and $PR(Cl)$ respectively. The maximization equation of (4) is as follows.

$$\widehat{Cl} = \arg max \; PR(Cl) \prod_{j=1}^{n} PR(ft_{ij}|Cl) \quad (6)$$

$$Cl \in \{'Stroke', 'Non - Stroke'\} \quad (7)$$

3.2.4 DT

Decision Tree (DT) classification is a versatile method that may be used for both regression and classification problems. This methodology employs a supervised learning algorithm and a target variable. It has arboreal traits. The decision tree consists of two main components: the decision node and the leaf node. The data is partitioned at the initial node and subsequently merged at the second node to produce the output. Due to its ability to replicate the stages of human creation of real-world objects, DT is easily comprehensible. The algorithm operates on the assumption that the presence or absence of a specific attribute in the dataset is influenced by other attributes, and it aids in classifying the target into a specific category.

3.2.5 SGD

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Stochastic Gradient Descent (SGD) is an optimization method that may be used to train different linear models and is not limited to a particular family of machine learning models. The approach is efficient as it computes the gradient using only one sample at each iteration. Minibatch functionality is enabled, making it well-suited for handling large-scale challenges.

3.2.6 RF

The Random Forest (RF) algorithm utilizes multiple independent decision trees and develops distinct subsets of instances through resampling in order to accomplish classification and regression tasks. Every decision tree produces its own categorization result, and the final class is determined by majority voting.

3.2.7 Majority Voting

When using a group of $'K'$ basic models, simple majority voting may be used to forecast the class label of an input instance. This can be done through either hard or soft voting methods. The former consolidates the votes corresponding to each class label and selects the one with the highest number of votes as the candidate class. The latter calculates the total of the expected probabilities for each class label and then predicts the class label with the highest likelihood. In this context, the method of hard voting was implemented. The equation below shows the overall function of the system.

$$max \sum_{k=1}^{K} P_{k,cl} \quad (8)$$

$P_{k,cl}$ prediction of model 'k' in class cl(stroke, non-stroke).

Hence, the ML models are explained in this section and the following section discusses the results proven by the models.

4. Results and Discussion

This section assesses the efficacy of the ML models by employing Python and Scikit-learn libraries. Experiments were conducted on a personal computer with the subsequent attributes: x64-based processor, Intel(R) Core(TM) i7-9750H processor at 2.60 GHz and 2.59 GHz, 16 GB of memory, Windows 10 Home 64-bit operating system. In our experimental setup, we utilized 10-cross validation to evaluate the performance of the models on the balanced dataset comprising 6148 instances. In order to construct the layering model, a fusion was performed on four base classifiers. To be more precise, the results obtained from Naïve Bayes, random forest andDT were incorporated into a logistic regression meta-classifier.

Various performance metrics were collected during the assessment of the ML models under consideration. For the purpose of this analysis, we will take into account the most frequently used ones from the relevant literature. Sensitivity, also known as recall (true positive rate), quantifies the proportion of participants who experienced a stroke and were accurately identified as positive, relative to all positive participants. Precision (P) and recall(Γ), on the other hand, are more appropriate for identifying errors within a mode.

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$P = \frac{T_p}{T_p + F_p} \quad (9)$$

$$\Gamma = \frac{T_p}{T_p + F_n} \quad (10)$$

$$\Phi = \frac{2(P * \Gamma)}{P + \Gamma} \quad (11)$$

$$A = \frac{T_p + T_n}{T_p + F_p + T_n + F_n} \quad (12)$$

The results shown by the proposed work is shown in the following figure 2 and tabulated in table 1 correspondingly.

Table 1. Performance comparison of classifiers

Classifiers/Perf.measures	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
k-NN	91.6	87.6	89.55	90.8
SVM	95.8	92.8	94.27	93.4
NB	93.2	88.4	90.73	92.1
DT	90.88	83.6	87.08	88.4
SGD	78.92	77.8	78.35	80.8
RF	96.8	91.6	94.12	94.3
Majority Voting	**98.4**	**97.8**	**98.09**	**98.1**

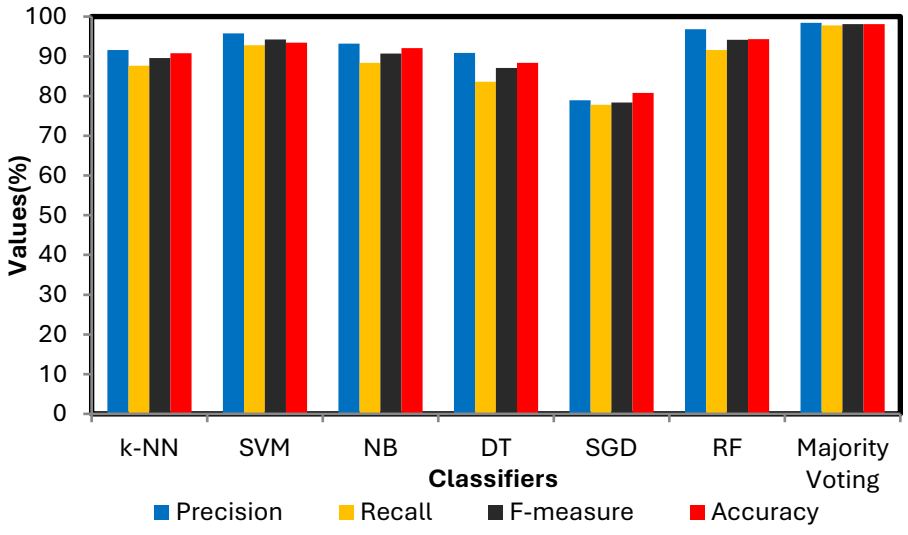

Fig.2. Performance of classifiers

In all considerations of metrics, the majority voting model implemented beneath the chosen foundation models was the most efficient. Exact values were also attained by the RF and SVM classifiers. The

Received: August 01 ,2025

International Journal of Applied Mathematics

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

majority voting model outperforms the SVM by 3.8% on the F-measure, and RF and NB by 3.9% and 7.3%, respectively. The results of the present investigation are shown compared with those of the study described in [35] using the identical dataset [34] in Table 2.

Table 2. Performance comparison against existing approaches

Techniques/Perf.measures	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
Proposed MV based + IGR	**98.4**	**97.8**	**98.09**	**98.1**
Proposed MV based	96.8	95.2	95.99	97.3
[17]	96.4	93.8	95.08	97.2
[23]	97.5	96.4	96.94	96.8

All proposed models, with DT and RF in particular, exhibit a substantial improvement over their respective performance in terms of recall, F-measure, and accuracy when compared to [17 and 23]. In summary, the majority voting method continues to exhibit superior performance and is the primary recommendation of our research.

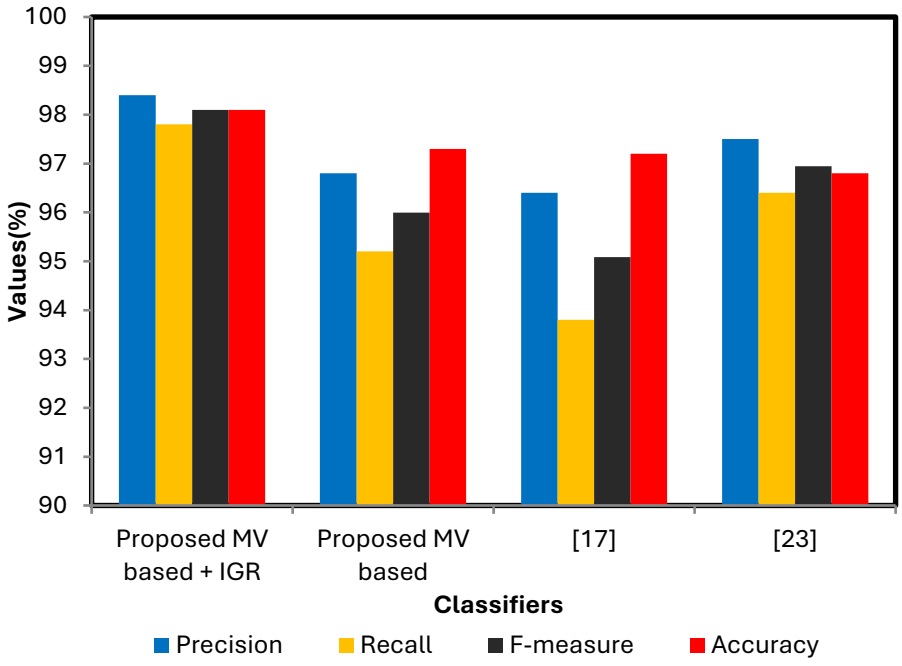

Fig.3. Comparative analysis against existing approaches

This study is limited by the fact that it was constructed using a publicly accessible dataset. These data possess distinct characteristics and dimensions in contrast to data obtained from medical facilities or academic institutions. Although the latter could provide more comprehensive information in the form of data models with numerous features that capture a detailed health profile of the participants, it is

Received: August 01 ,2025

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

typically difficult and time-consuming to obtain such data due to privacy concerns. From the experimental results, it is evident that the proposed work proves better performance with reasonable F-measure and accuracy rates.

Conclusions

A stroke is a life-threatening condition that necessitates prevention and/or treatment in order to avert unanticipated complications. In the current era, clinical providers, medical specialists, and decision-makers can utilize established models to identify the most pertinent characteristics (or, alternatively, risk factors) associated with the occurrence of strokes and to calculate the corresponding probability or risk. This research examines the efficacy of different ML algorithms in order to determine which one provides the most precise stroke prediction using a variety of profile features of the participants.

The assessment of the classifiers' performance, which incorporates accuracy, F-measure (a metric that combines precision and recall), is largely appropriate for interpreting the models and presenting their classification capabilities. Furthermore, they disclose the predictive capability and validity of the models with respect to the stroke category. With an accuracy of 98.1%, majority voting classification techniques outperforms the alternatives with an F-measure, precision, and recall rates of 98.04%, 98.4%, 97.8% respectively. Therefore, the majority voting classification is an effective strategy for identifying individuals, who are at a heightened risk of developing a chronic stroke. In future, this work can be extended by employing real-time dataset and deep learning techniques.

REFERENCES

[1] About stroke (Accessed: Dec 20, 2023). [Online]. Available:https://www.world-stroke.org/world-stroke-day-campaign/about-stroke

[2] T. Elloker and A. J. Rhoda, ‘‘The relationship between social support and participation in stroke: A systematic review,’’ Afr. J. Disability, vol. 7, pp. 1–9, Oct. 2018.

[3] M. Katan and A. Luft, ‘‘Global burden of stroke,’’ Seminar Neurol., vol. 38, no. 2, pp. 208–211, Apr. 2018.

[4] A. Bustamante, A. Penalba, C. Orset, L. Azurmendi, V. Llombart, A. Simats, E. Pecharroman, O. Ventura, M. Ribó, D. Vivien, J. C. Sanchez, and J. Montaner, ‘‘Blood biomarkers to differentiate ischemic and hemorrhagicstrokes,’’ Neurology, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021.

[5] X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, Y. Shen, and X. Li, ‘‘Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey,’’ J. Neurol., vol. 266, no. 6, pp. 1449–1458, Jun. 2019.

[6] A. Alloubani, A. Saleh, and I. Abdelhafiz, ‘‘Hypertension and diabetes mellitus as a predictive risk factors for stroke,’’ Diabetes Metabolic Syndrome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.

Received: August 01 ,2025

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[7] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, ‘‘Stroke risk factors, genetics, and prevention,’’ Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.

[8] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, ‘‘Stroke symptoms and the decision to call for an ambulance,’’ Stroke, vol. 38, no. 2, pp. 361–366, Feb. 2007.

[9] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White, M. Eccles, and H. Rodgers, ‘‘Response to symptoms of stroke in the UK: A systematic review,” BMC Health Services Res., vol. 10, no. 1, pp. 1–9,Dec. 2010.

[10] L. Gibson and W. Whiteley, “The differential diagnosis of suspected stroke: A systematic review,” J. Roy. College Physicians Edinburgh, vol. 43, no. 2, pp. 114–118, Jun. 2013.

[11] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, “Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review,” J. NeuroInterventional Surgery, vol. 12, no. 2, pp. 156–164, Feb. 2020.

[12] Y. Zhao, S. Fu, S. J. Bielinski, P. A. Decker, A. M. Chamberlain, V. L. Roger, H. Liu, and N. B. Larson, ‘‘Natural language processing and machine learning for identifying incident stroke from electronic health records: Algorithm development and validation,’’ J. Med. Internet Res., vol. 23, no. 3, Mar. 2021, Art. no. e22951.

[13] B. McDermott, A. Elahi, A. Santorelli, M. O’Halloran, J. Avery, and E. Porter, ‘‘Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis,’’ PhysiologicalMeas., vol. 41, no. 7, Aug. 2020, Art. no. 075010.

[14] A. Bivard, L. Churilov, and M. Parsons, ‘‘Artificial intelligence for decision support in acute stroke—Current roles and potential,’’ Nature Rev. Neurol., vol. 16, no. 10, pp. 575–585, Oct. 2020.

[15] W.Wang, M. Kiik, N. Peek, V. Curcin, I. J. Marshall, A. G. Rudd, Y.Wang, A. Douiri, C. D. Wolfe, and B. Bray, “A systematic review of machine learning models for predicting outcomes of stroke with structured data,” PLoS ONE, vol. 15, no. 6, Jun. 2020, Art. no. e0234722.

[16] M. S. Sirsat, E. Fermé, and J. Câmara, ‘‘Machine learning for brain stroke: A review,’’ J. Stroke Cerebrovascular Diseases, vol. 29, no. 10, Oct. 2020, Art. no. 105162.

[17] Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3), 817-828.

[18] Bandi, V., Bhattacharyya, D., & Midhunchakkravarthy, D. (2020). Prediction of Brain Stroke Severity Using Machine Learning. Rev. d'Intelligence Artif., 34(6), 753-761.

[19] Lee, H., Lee, E. J., Ham, S., Lee, H. B., Lee, J. S., Kwon, S. U., ... & Kang, D. W. (2020). Machine learning approach to identify stroke within 4.5 hours. Stroke, 51(3), 860-866.

Received: August 01 ,2025

Volume 38 No. 5 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[20] Jeena, R. S., & Kumar, S. (2016, December). Stroke prediction using SVM. In 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 600-602). IEEE.

[21] Shoily, T. I., Islam, T., Jannat, S., Tanna, S. A., Alif, T. M., & Ema, R. R. (2019, July). Detection of stroke disease using machine learning algorithms. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[22] Adam, S. Y., Yousif, A., & Bashir, M. B. (2016). Classification of ischemic stroke using machine learning algorithms. International Journal of Computer Applications, 149(10), 26-31.

[23] Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).

[24] H. K. Gupta, ‘‘Stroke prediction using machine learning algorithms,’’ Int. J. Innov. Res. Eng. Manage., vol. 8, no. 4, pp. 6–9, Jul. 2021, doi: 10.21276/ijirem.2021.8.4.2.

[25] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, ‘‘A predictive analytics approach for stroke prediction using machine learning and neural networks,’’ Healthcare Analytics, vol. 2, Nov. 2022, Art. no. 100032.

[26] A. A. Ali, ‘‘Stroke prediction using distributed machine learning based on Apache spark,’’ Stroke, vol. 28, no. 15, pp. 89–97, 2019.

[27] T. Dhar, N. Dey, S. Borra, and R. S. Sherratt, ‘‘Challenges of deep learning in medical image analysis—Improving explainability and trust,’’ IEEE Trans. Technol. Soc., vol. 4, no. 1, pp. 68–75, Mar. 2023, doi:10.1109/TTS.2023.3234203.

[28] Maldonado, S.; López, J.; Vairetti, C. An alternative SMOTE oversampling strategy for high-dimensional datasets. Appl. Soft Comput. 2019, 76, 380–389.

[29] Stroke Prediction Dataset. Available online: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed on 20 Dec 2023).

Received: August 01 ,2025