ARYAVART SHODH VIKAS PATRIKA
RNI TITLED NO. UPBIL04292
RNI REG. NO. UPBIL/2014/66218 PP. NO.- 344-350

ISSN NO.-2347-2944 (Print)
e-ISSN NO.-2582-2454 (Online)
Vol.-15, No.- II, Issues-24, YEAR- April-June-2022

Ensembled Machine Learning Model for Advanced Stroke Risk Prediction and Healthcare Enhancement

1. Faizaan Ahmada
2. DR. Vijay Kumar Verma

1. Research scholar, P.G Department of MATHEMATICS, 2. Professor & Head, P.G Department of PHYSICS, Magadh University, Bodh Gaya, (Bihar) India

Received- 21.06.2022, Revised- 25.06.2022, Accepted- 29.06.2022 E-mail: aaryvart2013@gmail.com

Abstract:** A stroke occurs when large brain parts lose blood flow, potentially leading to death. Given the increase in stroke occurrences, it is vital to understand the variables causing them. It is vital to create a system that can accurately predict whether or not a person will have a stroke. Several machine-learning methods, including Decision Trees (DT), Extra Trees (ET), Random Forest (RF), and Voting Classifiers (VC), are being investigated for their capacity to predict stroke risk. Furthermore, the research demonstrates that although cer- tain parameters, such as smoking status, age, and gender, are significant, others, such as domicile, are not and may be controlled by utilizing feature selection techniques. Principal Component Analysis is a dimensionality reduction approach that can be used together with a class balancing approach like the Synthetic Minority Over- sampling Technique (SMOTE). Previously, the dataset was unequal, with 95 percent of cases expressing non- stroke risk and the remaining cases suggesting stroke risk. To balance the data, the SMOTE oversampling approach is used, which involves replicating nearby samples. Each algorithm's Receiver Operating Character- istic score is determined, and ET, RF, and VC have a curve area larger than 0.95. When many performance criteria, such as overall accuracy, precision score, recall score, and f1 score, are considered, the Voting En- semble method outperforms current stroke detection algorithms. Most hypertensives will have a stroke. Stroke risk is greatest with hypertension. Most patients with cardiovascular disease experience a stroke. 95% of pa- tients without cardiac disease are stroke-free. 5% of heart disease patients suffer strokes.

Key Words: : Stroke, Machine Learning, Smart healthcare, Stroke prediction, Voting Ensemble method.

The healthcare sector is somewhat complex and evolving over time, which means that the system's activity varies over time and a comprehension of its component parts alone is insufficient to provide a complete picture. This system is more complicated than the educational system, banking industry, or military. No other business or industry has as many mechanisms, moving elements, a body with complex and varied requirements, and options and actions to meet those needs as the healthcare industry [1]. Numerous healthcare procedures must be modified due to the unpredictability of patient presentation. There are numerous stakeholders in the healthcare industry, each with their own responsibilities and interests, as well as rules of varying stringency that govern particular issues while neglecting others. There are numerous practicable combinations of care, activities, events, interac- tions, and outcomes. A stroke is a medical condition that can occur when portions of the brain do not receive enough blood or when their blood supply is cut off. In the absence of essential nutrients and oxygen, cells begin to perish. Strokes are regarded as medical emergencies requiring immediate treatment. It depends on a variety of health-related factors, such as a person's smoking and alcoholic behaviors, body mass index, glucose level, etc. The ability to predict strokes is crucial in the struggle against fatalities caused by strokes. This study included elevated blood pressure, body mass index (BMI), coronary artery disease, and average glucose levels as prospective risk factors for stroke. Moreover, ML has the potential to play a crucial role in stroke prediction[2] [3].

Brain MRI and CT are the main ways that neurodevelopmental prognosis is done when a stroke is evaluated. Other studies have demonstrated that bio-signals such as brain waves, muscle activity, and electrocardiograms may be utilized to detect and preclude stroke diseases [11]. For stroke detection, imaging methods like CT and MRI have recently been popular. However, they still have drawbacks in the examination and diagnostic process due to

Corresponding Author / Joint Authors

hypersensitive responses to contrast agent medication penetration, radiation exposure, and claustrophobia in a confined environment.

` **MOTIVATION-** Early stroke recognition, which is a critical step on the way to successful treatment, may benefit greatly from ML. Machine learning (ML) is a cutting-edge technology that can help doctors make clinical decisions and pre- dictions in all fields of life including financial sector [12], e-learning [13][14][15], recommendation systems [16][17], etc. This allows them to attain the previously stated aim. Few studies have been conducted over the past several decades to improve stroke diagnosis with the use of ML in terms of accuracy and speed. Given this, the current work categories some of those studies based on their commonalities, assesses each categorization in a systematic way, and provides useful information on the application of ML-based techniques in brain stroke.

CONTRIBUTION- In this work, the data description, machine learning algorithms and matrices, and experimental methodologies are each broken out into their respective areas.

- Approach 1: The first contribution is to use the testing set to test the models after training them with the original data (after SMOTE resampling).
- Approach 2: The second one is on the testing set, constructing new data with the assistance of the predictions generated by the basic models used in Approach 1.

The principal component analysis was performed on each of the test sets. The PCA transformations of the data were used to produce new data, which were then handled as new features. These new data were produced using the predictions of the data. The original models are going to be trained on the new data that was used for the training set, and then the new test set is going to be tested using a voting classifier, an ensemble that includes the original model as its basis.

Table 1. List of Abbreviations

Abbreviations	Term
ML	Machine Learning
MRI	Magnetic Resonance Imaging
CT	Computerized Tomography
BMI	Body Mass Index
SMOTE	Synthetic Minority Oversampling Technique
PCA	Principal Component Analysis
RF	Random Forest
DT	Decision Trees
ET	Extra Trees
KNN	K-Nearest Neighbour
SVM	Support Vector Machine
VC	Voting Classifiers
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent

The random forest (RF) algorithm is a supervised learning technique that utilizes the bagging method to con- struct an ensemble of decision trees (DT).

In the RF algorithm, the construction of trees occurs in an unpredictable manner. Unlike a standard DT, where the selection of the most informative feature is prioritized for node splitting, RF considers a diverse range of features during this process. This diversity results in the generation of a more robust model. Therefore, when splitting a node in RF, only a randomly selected subset of features is considered. The introduction of random thresholds to each feature further enhances the randomness of the trees, surpassing the pursuit of finding the optimal thresholds as in a standard DT.

ORGANIZATION- The research paper has been divided into five main sections. In the first section, a quick explanation of the study's purpose and scope is discussed. In Section 2, the relevant research is analysed, and a comparison study is also presented in tabular form. The methods that were applied in those investigations are examined in Section 2, which is titled "Materials and Methods." This section discusses a variety of methodologies, provides an explanation of the data that was selected, examines ML methodologies, and examines evaluation metrics. The results of the correlation and performance analysis are discussed in the next section, which is section 3, along with additional information that is more in-depth. The conclusion of this study may be found in section 4, which contains the concluding notes as well.

MATERIALS AND METHODS- LONG-TERM RISK OF STROKE- In order to evaluate the long-term risk of stroke, the initial dataset was divided into a training set and a test set. Each instance in the dataset was assigned a binary variable y to indicate its class label, representing either "Stroke" or "Non-Stroke." The objective of the subsequent analysis was to develop machine learning models that demon- strate high recall (or sensitivity) and area under the curve, ensuring accurate prediction of stroke cases. The pro- posed methodology for stroke prediction encompassed several steps, which are elaborated upon below. Within the realm of classification analysis, feature importance assumes a critical role in facilitating the development of precise and reliable machine learning models. In this section, we present the models that will be utilized in the classification framework for stroke occurrence. For this purpose, various types of classifiers are employed.

NAIVE BAYES CLASSIFIERS- First, the naïve Bayes (NB) classifier was explored, which maximises probability if the characteristics are substantially independent [50]. A new subject i with characteristics vector xi is classified in the class y with the highest P(y|x_{i1},..., x_{in}). The conditional probability is expressed as follows:

$$P(y|\, x_{i1}, \ldots, x_{in}) = \frac{P(x_{i1}, \ldots, x_{in}|y)P(y)}{P(x_{i1}, \ldots, x_{in})} \tag{1}$$

Where $P(x_{i1}, \ldots, x_{in}|y) = \prod_{j=1} P(x_{ij}|y)$ is the features probability given the class, P(y) is the prior proba- bility of the class, and P(xi1,..., xin) is the prior probability of the features. The following optimisation problem was created in order to maximise the numerator of equation (1).

$$\hat{y} = arg\ max\ P(y) \prod^{n}{}_{j=1} P(x_{ij}|y) \tag{2}$$

K-NEAREST NEIGHBORS- The K-nearest neighbours (K-NN) classifier is an algorithm that relies on distance measurements, such as Euclidean or Manhattan distances, to assess the similarity or dissimilarity between two instances within a given dataset . Among these distance metrics, the Euclidean distance stands out as the most straightforward and widely adopted. Assuming xnew represents the feature vector of a new sample awaiting classification, which can be categorized as either a stroke or non-stroke case, the KNN classifier identifies the K nearest vectors (i.e., neigh- bours) to xnew. Subsequently, xnew is assigned to the class to which the majority of its neighbouring vectors belong.

RANDOM FOREST- Random Forest is an ensemble learning algorithm [19] that combines multiple decision trees to make predic- tions. It utilizes the concepts of bagging and random feature selection to improve accuracy and reduce overfitting. Its suitability for stroke prediction lies in its ability to handle high-dimensional data, handle missing values, and provide robustness against noise, ultimately leading to reliable predictions.

PROPOSED CLASSIFIER- The use of RF makes determining the relevance of variables simpler. Sklearn measures relevance by how much a feature reduces impurity across all trees in the forest. Because of their low predictive value, features based on significance may be eliminated. Overfitting happens largely when there are too many traits, although it may also occur in the reverse direction. The RF's hyper parameters are utilised to either speed up the model or increase its forecasting abilities. One of the most significant advantages of an RF is its versatility. It can be used to solve classification and regression problems, and it is evident how much weight the input characteristics get.

$$Entropy\ \Phi(str) = P_{str=1} * \log(P_{str=1}) - P_{str=0} * \log(P_{str=0}) \quad (4)$$

$$Gain(str, A) \equiv Entropy(str) - \sum_{v \in D_A} \frac{|S_v|}{|Str|} Entropy(S_v) \quad (5)$$

ET adopts a top-down approach to construct decision and regression trees. Unlike previous tree-based ensemble methods, Extra Trees utilizes the entire learning sample during tree construction. The hyperparameter for Extra Trees is the number of trees in the ensemble. Increasing the number of trees generally improves the overall performance of the model. While adding more trees may seem prone to overfitting, all three methods (bagging, RF, and Extra Trees) are stochastic and inherently resistant to overfitting.

Evaluation Metrics- During the process of evaluating the Machine Learning models that were taken into consideration, numerous performance indicators were recorded. In this particular analysis, we will focus on the one that is utilised the vast majority of the time in the relevant literature .

The predictive performance of a model may be summed up using the F-measure, which is the harmonic mean of the accuracy and recall statistics.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$F - Measure = 2\frac{Precision * Recall}{Precision + Recall} \quad (8)$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (9)$$

Here, TP: true positive, TN: true negative, FP: false positive and FN: false negative. Machine learning and statistical analysis use the area under the curve (AUC) to evaluate binary classification models. It measures the model's ability to differentiate stroke and non-stroke cases. A higher AUC indicates better stroke classification.

The AUC of a receiver operating characteristic (ROC) curve shows the trade-off between sensitivity (sensitivity) and specificity (1-specificity) for different categorization thresholds. An AUC of 1 indicates that the model properly assigns stroke cases higher probability or scores than non-stroke instances. If all strokes and non-strokes are misclassified, the AUC is 0, indicating no discriminating power. AUC evaluates model performance across categorization thresholds.

Table 2. Table 3: Description of Features of Stroke Dataset

S. No.	Attribute	Description	Column_ name	Sample [0]	Sample [1]	Sample [2]	Sample [3]
1.	Id	Unique identifier number of the patient.	'id'	9046	51676	31112	60182
2.	Gender	Male or Female or Other.	'gender'	M	F	M	F
3.	Age	Age of the patient.	'age'	67	61	80	49
4.	Hypertension	0 if patient is not hypertensive, 1 if the patient is hypertensive.	'hypertension'	0	0	0	0
5.	Heart disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease.	'heart_disease'	1	0	1	0
6.	Ever married	The marital status: yes or no.	'ever_married,	Yes	Yes	Yes	Yes
7.	Work type	work status: children, Government job, never worked, Private or Self-employed.	'work_type,	Private	Self-Employed	Private	Private
8.	Residence	The living status: rural or urban.	'Residence_type	Urban	Rural	Rural	Urban
9.	Level of Glucose	Average glucose level mg/dL present in blood of the patient.	'avg_glucose_level'	228.69	202.21	105.92	171.23
10.	BMI	Body mass index of the patient in Kg/m2	'bmi'	36.6	NA	32.5	34.4
11.	Smokes	Categories: formerly, never, smokes or unknown	'smoking_status'	Formerly Smoked	Never Smoked	Never Smoked	Smokes
12.	Stroke	1 if the patient had a stroke or 0 if not	'stroke'	1	1	1	1

IMPLEMENTATION & ANALYSIS- DATASET DESCRIPTION- The dataset for the study and estimation of stroke risk was obtained from the open-source database repository Kaggle. It's a healthcare dataset including sample rows on around 5110 persons. Table 3 describes all of the attributes gathered, which comprises information on 5110 people with various health conditions. The data was compiled from a variety of medical clinics in Bangladesh. The stroke dataset consisted of 12 metrics, including patients' demographic data (ID, Gender, Age, Marital Status, Type of Work, and Residence Type) and health records (Hypertension, Heart Disease, Average Glucose Level measured, Body Mass Index (BMI), Smoking sta- tus and experience of stroke). In Table 3, Column_name abbreviates all 12 mentioned features, and Sample[0], Sample[1], and Sample[2] presents the feature values.

From the dataset, the age, hypertension, heart disease, avg_glucose_level, BMI, and stroke features are the numerical values hence Table 4 presents a statistical analysis of such features from the dataset. Table 4 provides the total count, Mean (average), Min (minimum), Max (maximum), and SD (standard deviation) of the numerical data values presented in the stroke dataset.

Table 3. Statistical Analysis of Stroke Dataset

Par.	age	hyper-tension	heart_disease	avg_glucose_level	bmi	stroke
Count	5110	5110	5110	5110	4909	5110
Mean	43.22	0.097	0.054	106.147	28.89	0.0487
Min.	0.08	0	0	55.12	10.3	0
Max	82	1	1	271.74	97.6	1
SD	22.61	0.29	0.226	45.283	7.854	0.2153

EVALUATION- Following the training phase, the proposed ensemble model for classifying stroke risk was tested using a new test data set. Measurements are taken on things like accuracy, precision, recall, F1-score, confusion matrix, and Roc curve. The classification report provides a summary of the metrics that were reviewed.

The results have been obtained to evaluate the ET, DT, RF and VC method and flow that has been followed to train models in order to accomplish the found to outperform approach for estimating the final optimum performance. These results have allowed for an evaluation of the method and flow that has been followed.

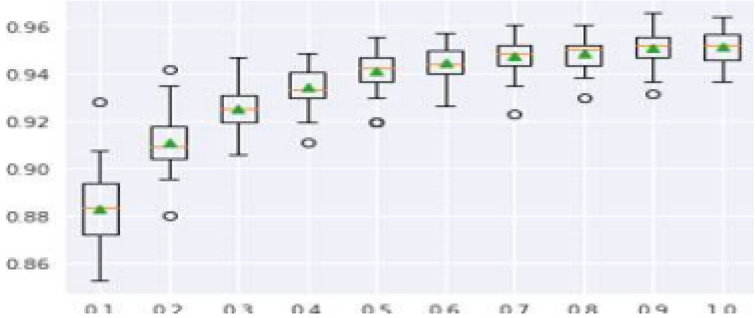

Fig.5. Random Forest results

When trained and evaluated on the additional features data, an ensemble model was utilized, and the results showed the amount of right and incorrect predictions. It has demonstrated the best outcomes compared to the other models that were examined before. It has been noticed that the number of wrong results is at its lowest. The ROC-AUC curve has been plotted for the proposed method. The curve has achieved 0.97 area of evaluate the nature of predicted classes. The model has outperformed in terms of area under curve, accuracy and other metrics measure to distinguish and classifying two different classes; stroke and not stroke.

Observation-

* Hypertension and high blood pressure: The majority of persons who have hypertension as a risk factor will

have a stroke. A stroke will not occur in around 90% of patients who do not have hypertension. The greatest major risk factor for stroke seems to be hypertension.

* Cardiovascular Disease: The majority of persons with heart disease will have a stroke. Approximately 95% of patients with no heart illness are not at risk of having a stroke. Heart disease seems to be a major cause of stroke, since 5% of people with heart disease had a stroke.

It can be seen that "NB" has an accuracy of 0.84 which is better than "LR" with an accuracy of 0.79 and "NN" with an accuracy of 0.81. "SGD" has an accuracy of 0.88, precision of 0.79, recall of 0.79, and an F1-Score of 0.79. Only accuracy is good, but it has lowest recall, precision and F1-score. "MLP" has an accuracy of 0.92 and "ET" has an accuracy of 0.96 which is better than "DT" with an accuracy of 0.91. "RF" and "VC" has an accuracy of 0.95 and 0.96 respectively.

"Proposed Ensemble" has the highest values across all metrics with an accuracy of 0.97, precision of 0.96, recall of 0.98, and an F1-Score of 0.97. This suggests that the proposed ensemble model performs very well across these evaluation metrics.

CONCLUSION AND FUTURE WORKS- The healthcare industry needs to change into a learning system where everyone knows how it works, and there are strong feedback loops to keep change moving forward. The efficiency of our existing systems may be en- hanced if we develop a unified point of view and experiment with other ways of thinking.

It is time to stop making things more complicated, changing the boxes on the organization chart, and adding more kips; instead, it is time to do something more complex, which is why I would like to warn those who advocate for the most popular tactics of the present day. Performance metrics include accuracy, precision, recall, the f1-score, comparison bar graphs, and confusion matrices.

Thus, in the future, both features from CSV and image file (extraction using deep learning) can be combined and ML can be applied to predict the stroke with improved accuracy. Also, IoT-based data collection and online diagnostics can be made, and prescriptions can be recommended by Medical Experts based on the patient's data and his/her images through the hybrid model.

Also, the patient's data having all details along with CT scans/PET/MRI/FMRI can be considered to pass it into the Deep Learning model. The features extracted from flatten layer can be combined with other features of patients like age, other diseases, symptoms, etc., to get better models which can detect and recommend preventive measures to avoid strokes.

REFERENCES

1. M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. Al Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's dis- ease, Parkinson's disease and schizophrenia," Brain Inform, vol. 7, no. 1, 2020, doi: 10.1186/s40708-020-00112- 2.
2. M. Mahmud et al., "A Brain-Inspired Trust Management Model to Assure Security in a Cloud Based IoT Frame- work for Neuroscience Applications," Cognit Comput, vol. 10, no. 5, pp. 864-873, Oct. 2018, doi: 10.1007/s12559-018-9543-3.
3. S. Bhatia, S. Alam, M. Shuaib, M. Hameed Alhameed, F. Jeribi, and R. I. Alsuwailem, "Retinal Vessel Extraction via Assisted Multi-Channel Feature Map and U-Net," Front Public Health, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.858327.
4. "Ischemic stroke." [Online]. Available: https://www.mayoclinic.org/diseases-conditions/stroke/multime- dia/img-20116029

5. S.-L. Liew et al., "A large, open source dataset of stroke anatomical brain images and manual lesion segmentations," Sci Data, vol. 5, no. 1, p. 180011, Dec. 2018, doi: 10.1038/sdata.2018.11.

6. Y. Sun et al., "Risk factors for constipation in patients with acute and subacute ischemic stroke: A retrospective cohort study," Journal of Clinical Neuroscience, vol. 106, pp. 91-95, Dec. 2022, doi: 10.1016/ j.jocn.2022.10.014.

7. 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, p. 100032, Nov. 2022, doi: 10.1016/j.health.2022.100032.

8. T. D. Musuka, S. B. Wilton, M. Traboulsi, and M. D. Hill, "Diagnosis and management of acute ischemic stroke: speed is critical," Can Med Assoc J, vol. 187, no. 12, pp. 887-893, Sep. 2015, doi: 10.1503/ cmaj.140355.

9. Q. Song et al., "Long Sleep Duration and Risk of Ischemic Stroke and Hemorrhagic Stroke: the Kailuan Pro- spective Study," Sci Rep, vol. 6, no. 1, p. 33664, Sep. 2016, doi: 10.1038/srep33664.

*****