Prediction of Brain stroke using Machine Learning Techniques

Sai deepak Pemmasani, Kalyana Lakshmi, Diveesh poli ,Dr. Dakshinamurthy

Abstract :

Stroke occurs when the flow of blood to a specific region of the brain is abruptly halted, leading to the gradual death of brain cells and potentially resulting in various degrees of disability, contingent upon the affected area. Early identification of symptoms is paramount for both predicting and preventing strokes, thereby fostering a healthier lifestyle. In this research endeavor, machine learning (ML) techniques are leveraged to construct and assess multiple models aimed at formulating a robust framework for the long-term prediction of stroke occurrence.In our research, several classification models were rigorously evaluated for their efficacy in predicting stroke occurrence like XGBoost Classifier,CatBoost Classifier ,LightGBM Classifier, Bagging Classifier , ExtraTrees Classifier.The accuracy of ExtraTrees Classifier is 99.8 % , The accuracy of Bagging Classifier is 98.5% ,The accuracy of CatBoost Classifier is 97.7% ,The accuracy of XGBoost Classifier is 96.3% ,The accuracy of LightGBM Classifier is 91.8%.

Keywords: Stroke, machine Learning, Prediction, Classifiers.

1. Introduction:

Understanding how our body works is crucial for our overall well-being. Unfortunately, as we age, one significant threat we face is the risk of strokes, especially for those over 55. Just like how heart attacks affect our heart's function, strokes can have serious consequences for our brain.

Strokes, whether they're caused by a blockage or bleeding, have serious effects on our health. Our brain relies on a steady supply of oxygen and nutrients delivered through our bloodstream. Any disruption to this supply can result in severe complications. That's why it's crucial to understand what causes strokes, recognize their symptoms, and know the risk factors associated with them. This knowledge is vital for preventing strokes and ensuring effective treatment when they occur.

Stroke, according to the World Stroke Organization, affects around 13 million people every year, causing about 5.5 million deaths. It's the top reason for both death and disability worldwide, impacting all areas of life. Stroke doesn't just affect individuals, but also their families, friends, and workplaces. Contrary to what many think, it can happen to anyone, no matter their age, gender, or health. This highlights the importance of understanding and dealing with the challenges of stroke for everyone.

Several factors contribute to an increased likelihood of experiencing a stroke. Individuals with a history of prior strokes or transient ischemic attacks (TIAs), often referred to as mini-strokes, are at elevated risk. Additionally, those affected by heart conditions such as myocardial infarction (heart attack), heart failure, or atrial fibrillation face heightened susceptibility. While age is a significant factor, with individuals over 55 being more prone to strokes, it's essential to recognize that strokes can occur at any age, even among children.

Other risk factors include hypertension (high blood pressure), carotid artery narrowing due to atherosclerosis, smoking, high blood cholesterol levels, diabetes, obesity, leading a sedentary lifestyle, excessive alcohol consumption, blood clotting disorders, estrogen therapy, and the use of stimulant drugs like cocaine and amphetamines. Each of these factors plays a role in increasing the overall risk of stroke occurrence.

Information and communication technologies (ICTs), particularly artificial intelligence (AI) and machine learning (ML), have become pivotal in early disease prediction across various health conditions, including diabetes, hypertension, cholesterol levels, COVID-19, chronic obstructive pulmonary disease (COPD), cardiovascular diseases (CVDs), acute liver failure (ALF), sleep disorders, hepatitis C, and chronic kidney disease (CKD). In this study, we focus on stroke prediction, an area that has seen significant attention in machine learning research.

Our research introduces a methodology for developing effective binary classification ML models for predicting stroke occurrence. Recognizing the importance of class balancing in stroke prediction, we applied the synthetic minority over-sampling technique (SMOTE) to address this issue. Subsequently, we developed, configured, and evaluated various models using the balanced dataset. These models included CatBoost Classifier , LightGBM Classifier, Bagging Classifier , ExtraTrees Classifier, XGBoost Classifier.

Our experiments demonstrated the effectiveness of the ExtraTrees Classifier compared to individual models . ExtraTrees Classifier achieved high performance metrics including area under the curve (AUC), precision, recall, F-measure, and accuracy, highlighting its potential for improving stroke prediction accuracy.

2.Related Work :

Ref_id	Author names	Algorithms used	Algorithm best performed	Accuracy
1	Mohamed Hanifa, Kasmir Raja S.V	Support Vector Machine	Support Vector Machine	98%
2	Jeena RS, Kumar S	Support Vector Machine (SVM) with different kernals	Support Vector Machine-linear kernal	90%
3	Dritsas E. , Trigka M.	stacking method	stacking method	98%
4	Sailasya G , Kumari G. L. A.	Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors, Support Vector Machine, Naïve Bayes Classification	Naïve Bayes Classification	82%
5	Tavares J.A.	Decision Tree, Logistic Regression, Naïve Bayes Classification, K-Nearest Neighbors, Random Forest, Neural networks Support Vector Machine, XGBoost Classifier	Decision Tree, XGBoost Classifier	90.11% 91.52%
6	Shoily T.I., Islam T. Jannat S, Tanna S.A. Alif T.M, Ema, R.R	Naive Bayes, J48, k-NN, Random Forest classifier	J48, k-NN, Random Forest classifier	99.8% 99.8% 99.8%
7	Govindarajan P, Soundarapandian R.K, Gandomi A.H., Patan R. Jayaraman P, Manikandan R.	artificial neural networks, support vector machine, boosting and bagging , random forests	artificial neural networks trained with a stochastic gradient descent algorithm	95%
8	Rahman S, Hasan M Sarkar	Extreme Gradient Boosting (XGBoost), Ada Boost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K Neighbors, SVM - Linear Kernel, Naive Bayes, deep neural networks (3-layer and 4-layer ANN)	Random Forest classifier Three layer deep neural network (4-Layer ANN)	99% 92.39%
9	Hung CY, Lin CH, Lan TH, Heng GS, Lee CC	DNN and gradient boosting decision tree (GBDT), logistic regression (LR) , support vector machine (SVM)	DNN and gradient boosting decision tree (GBDT)	
10	K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin, M. F. Mridha	K Nearest Neighbors, Logistic Regression, Support Vector Machine, Random Forest, XGB Classifier, Naive Bayes	Random Forest XGB Classifier	90.3% 89.02%

11	B.Akter, A.Rajbongshi, S. Sazzad, R. Shakil, J. Biswas U. Sara	Random Forest , Support Vector Machine , Decision Tree	Random Forest	95.30%
12	Tusher A. N., Sadik, M. S., Islam, M. T.	Logistic Regression, Classification and Regression Tree, K-Nearest Neighbor, Support Vector Machine	K-Nearest Neighbor	97%
13	Adam SY, Yousif A, Bashir MB	K-Nearest Neighbor	K-Nearest Neighbor	99%
14	N.S. Adi, N.R. Farhany, R.Ghina H.Napitupulu	Naive Bayes, Decision Tree, Random Forest	Random Forest Decision Tree	94.78% 91.9%
15	Biswas N, Uddin K. M. M, Rikta S. T, Dey S. K	Support Vector Machine, Random Forest, K-nearest Neighbor, Decision Tree, Naïve Bayes, Voting Classifier, AdaBoost, Gradient Boosting, Multi-Layer Perception, Nearest Centroid	Support Vector Machine	99.99%
16	Bandi V., Bhattacharyya D , Midhunchakkravarthy	Decision Tree , Gaussian Naïve Bayes, Logistic Regression , Linear SVM , Poly SVM , RBG SVM , Random Forest , AdaBoost AdaBoost with SGD	Decision Tree	93.75%
17	S.Gupta , S. Raheja	Gaussian Naive Bayes, Logistic Regression, Decision Tree Classifier, K-Nearest Neighbours, AdaBoost Classifier, XGBoost Classifier, Random Forest Classifier.	Random Forest Classifier, XGBoost, AdaBoost	97% 96% 95%
18	Puranjay SavarMattas	KNeighborsClassifier Support Vector Classification (SVC) Decision Tree Classifier Random Forest Classifier Multi-layer Perceptron classifier Stacking Classifier	Random Forest Classifier,	99.6%

3. Materials and Methods :

3.1. Dataset Description:

Our research utilized the McKinsey & Company healthcare hackathon dataset, with the Electronic Health Record (EHR) managed as the primary data source (McKinsey Analytics, 2018). This dataset is freely accessible for download from a repository and comprises 43,400 patient records, each containing 12 common attributes. Among these attributes, 10 serve as input variables, defining the operational features of the model.

The remaining attribute determines the likelihood of a stroke occurrence based on the combined input parameters, thus serving as the model's output variable. The input attributes, listed in their order of input, are as follows:

1. id: Patient ID
2. Gender: Male/Female
3. Age: 1-100
4. Hypertension: 1(True)/0(False)
5. Heart Disease: 1(True)/0(False)
6. Ever Married: 1(True)/0(False)
7. Average Glucose Level: 55-292
8. Body Mass Index(BMI): 10-100
9. WorkType:Govt/Never-Worked/Private/Self-Employed/Children
10. Residence Type: (0) Rural/ (1) Urban
11. Smoking Status: Formerly Smoked/Never Smoked/Smokes

To ensure our model accurately predicts the likelihood of a stroke, we need to carefully consider all the factors mentioned. This means we must remove any missing values and review data entries that could affect the model's accuracy from the dataset of 43,400 records. It's crucial to identify and fix any issues with the data before we use it in the model.

3.2. Data preprocessing:

Improving the quality of raw data is crucial for enhancing the accuracy of our predictions. This involves addressing issues like missing values and noisy data through a process called data preprocessing. This process includes reducing redundant values, selecting relevant features, and organizing the data for better analysis .Part of data preprocessing involves ensuring that there's a balanced representation of different classes in our dataset. In our case, we used a technique called SMOTE to address the imbalance between stroke and non-stroke cases. Specifically, we oversampled the minority class (stroke) to ensure an equal distribution of participants across both classes.

3.2.1. The proportion of stroke among gender:

Gender has a minimal impact on the model, as shown in Figure 1 below, where the Female gender dominates with 59.2% of the demographic compared to the Male gender's 40.8%. This small difference has led us to assume that gender has a very small influence in what causes strokes.

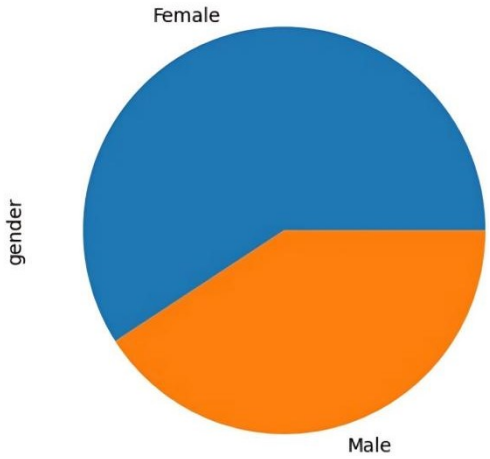

Fig. 1 -stroke among genders

3.2.2. The proportion of stroke based on marital status:

We were surprised to discover that unmarried participants accounted for only 35.6% of the total, while a significant majority, comprising 64.4%, were married. This finding suggests that marital status may indeed be a factor worth considering when assessing the likelihood of a stroke. It appears that one's marital status could play a significant role in the risk factors associated with stroke. The chart below in Figure 2 illustrates these proportions visually.

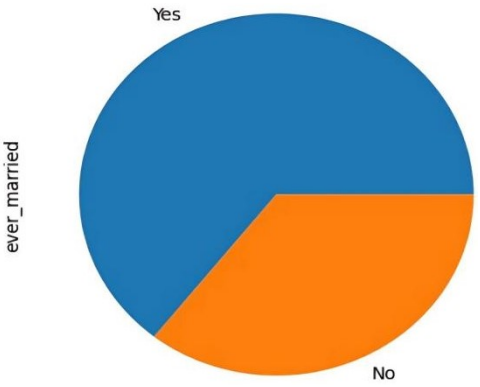

Fig. 2 -stroke based on marital status

3.2.3. The proportion of stroke based on work type:

The plot shown in Figure 3 provides valuable insights into the correlation between job status and the likelihood of experiencing work-related pressure. Notably, individuals in private employment show a relatively higher stroke rate of 57.2% within our dataset. In contrast, those employed in government roles exhibit a stroke rate of 12.5%, while self-employed individuals demonstrate a rate of 15.6%. Children, on the other hand, represent a rate of 14.1%, and people who never worked represent a minimal portion of 0.6%.By considering job status as a significant factor guiding our analysis, this data contributes significantly to our predictive efforts.

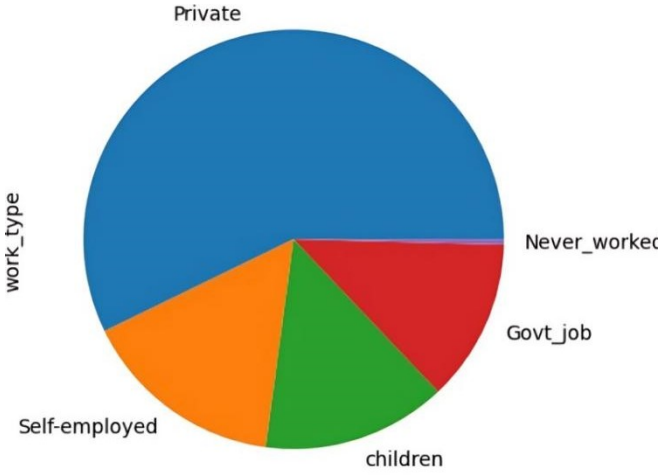

Fig. 3 -stroke based on work type

3.2.4.The proportion of stroke based on residence type:

Analyzing whether a participant resides in an urban or rural area may seem relevant to their likelihood of experiencing a stroke. However, upon reviewing the chart shown in Figure 4, we observe that 50.1% of urban participants and 49.9% of rural participants experienced strokes. This data suggests that the participant's place of residence does not significantly impact the likelihood of a stroke occurrence.

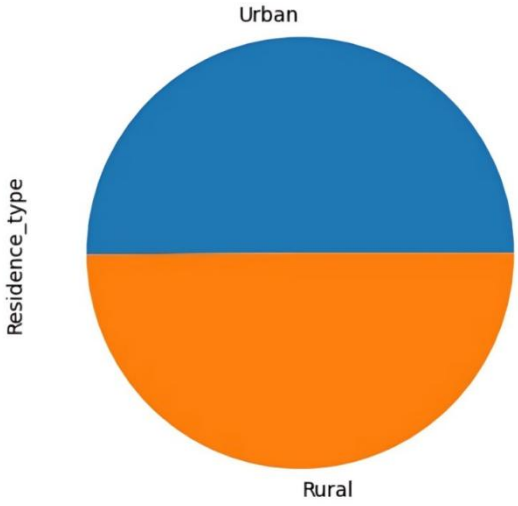

Fig. 4 -stroke based on residence type

3.2.5. The proportion of stroke based on smoking status:

The graph shown in Figure 5 reveals a rather unexpected trend, 55.3% of individuals who have never smoked are found to have experienced a stroke, compared to 29.6% of former smokers and only 15.1% of current smokers.But we need to be careful when we analyze this data because there's a lot of information missing in this column.

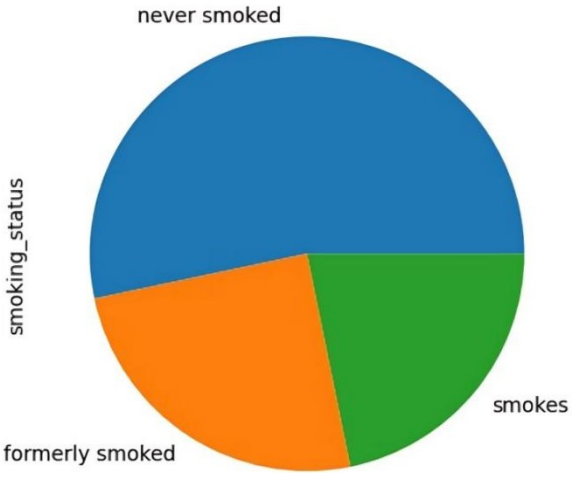

Fig. 5 -stroke based on smoking status

3.3. Machine Learning Models:

3.3.1. XGBoost Classifier:

We considered the XGBoost Classifier, which is a powerful gradient boosting technique known for its effectiveness in predictive modeling tasks. XGBoost optimizes predictive performance by iteratively constructing an ensemble of weak learners, with each subsequent learner aiming to correct the errors of its predecessors.

For a new subject *i* with feature vector *fi* classification into class *c using* XGBoost involves maximizing the probability *P(c|fi1,...,fin).*

The prediction process in XGBoost can be conceptualized as:

$$\hat{c} = \text{argmax} \sum_{i=1}^{N} [\text{Loss}(\text{actual}_i, \text{predicted}_i) + \Omega(f_m)]$$

where *Loss* represents the loss function to be minimized, actual *i* and predicted *i* denote the actual and predicted values for the *i*-th observation,respectively, Ω(*fm)* is the regularization term penalizing model complexity, and *fm* represents the *m*-th tree in the ensemble.

The accuracy of the XGBoost for our dataset is 96.3%.

3.3.2.CatBoost Classifier:

we explored the CatBoost classifier, which is a state-of-the-art gradient boosting algorithm designed for high-performance machine learning tasks. CatBoost is particularly renowned for its ability to handle categorical variables effectively and efficiently.

The mathematical representation of CatBoost Classifier is

$$F(x) = F_0(x) + \sum_{m=1}^{M} \sum_{i=1}^{N} f_m(x_i)$$

F(x) represents the overall prediction function that CatBoost aims to learn. It takes an input vector x and predicts the corresponding target variable y.

$F_0(x)$ is the initial guess or the baseline prediction. It is often set as the mean of the target variable in the training dataset. This term captures the overall average behavior of the target variable.

$\sum_{m=1}^{M}$ represents the summation of the ensemble of trees. M denotes the total number of trees in the ensemble.

$\sum_{i=1}^{N}$ represents the summation of the training samples. N denotes the total number of training samples.

$f_m(x_i)$ represents the prediction of the m-th tree for the i-th training sample. Each tree in the ensemble contributes to the overall prediction by making its own prediction for each training sample.

The accuracy of the CatBoost Classifier for our dataset is 97.7%.

3.3.3.LightGBM Classifier:

LightGBM represents a specialized implementation of gradient boosting, renowned for its efficiency and scalability in handling large datasets. It introduces innovative features such as a histogram-based learning approach and a leaf-wise tree growth strategy, distinguishing it from traditional gradient boosting implementations.

The mathematical representation of LightGBM Classifier is

Y = Base_tree(X) - lr * Tree1(X) - lr * Tree2(X) - lr * Tree3(X)

This illustrates the core mechanism of LightGBM in combining predictions from various trees to derive a final prediction.

Base Tree Prediction: Base_tree(X) The initial term denotes the prediction generated by a simple decision tree, often shallow in depth. This base tree provides a preliminary estimation of the target variable.

Correction Terms: (-lr * Tree1(X) - lr * Tree2(X) - lr * Tree3(X)) The subsequent terms represent the corrections applied by additional trees (Tree1, Tree2, Tree3, etc.). Each of these trees aims to rectify the errors made by its predecessors. The learning rate (lr) modulates the impact of these corrections, influencing their magnitude.

Final Prediction: (Y) The final prediction is the culmination of the base tree prediction and the collective corrections introduced by subsequent trees, each scaled by the learning rate. This iterative process results in a refined and comprehensive prediction, surpassing the predictive capability of an individual decision tree.

The accuracy of the LightGBM classifier for our dataset is 91.8%.

3.3.4. Bagging Classifier:

Bagging (or Bootstrap aggregating) is a type of ensemble learning in which multiple base models are trained independently and in parallel on different subsets of the training data. Each subset is generated using bootstrap sampling, in which data points are picked at random with replacement. In the case of the bagging classifier, the final prediction is made by aggregating the predictions of the all-base model using majority voting. In the models of regression, the final prediction is made by averaging the predictions of the all-base model, and that is known as bagging regression.

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

The accuracy of the Bagging Classifier for our dataset is 98.5%.

3.3.5.Extra Trees Classifier:

The Extra Trees Classifier, short for Extremely Randomized Trees Classifier, is a variant of the Random Forest algorithm that further increases the randomness of the trees. It's designed to reduce variance and increase computational efficiency while maintaining or even improving predictive performance.

ExtraTrees classifier performed better than other classifiers for our dataset with an accuracy of 99.8%

4. PROPOSED WORK:

The research workflow follows a systematic process for acquiring, preprocessing, analyzing, and evaluating brain stroke data. This structured approach ensures the reliability and robustness of the findings and contributes to the development of effective predictive models for stroke risk assessment or prognosis.

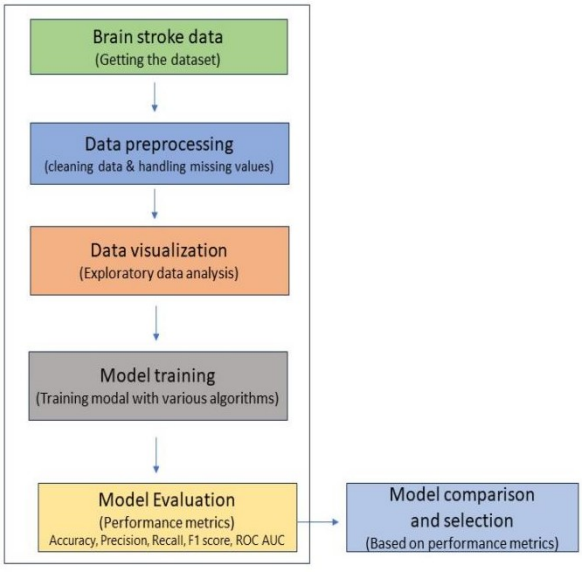

Fig.6-Proposed workflow

4.1. Acquiring Brain Stroke Data:

Data acquisition involves sourcing relevant datasets from reputable sources such as medical databases, research repositories, or healthcare institutions. In this study, we utilized a comprehensive healthcare dataset that includes information on patients' habits and lifestyles, alongside other pertinent medical data.

4.2. Data Preprocessing:

Preprocessing begins with cleaning and preparing the acquired data. Tasks include handling missing values, removing outliers, standardizing or normalizing numerical features, and encoding categorical variables. The dataset is split into training and testing sets to ensure unbiased model evaluation.

4.3. Data Visualization (Exploratory Data Analysis):

Exploratory Data Analysis (EDA) involves visualizing the data through histograms, box plots, scatter plots, and correlation matrices. These visualizations provide insights into the distribution of variables, relationships between features, and potential patterns or trends in the data.

4.4. Model Training:

Model training encompasses the utilization of diverse machine learning algorithms tailored for binary classification tasks. Specifically, we employed models such as CatBoost, LightGBM, XGBoost, Bagging classifier, ExtraTrees, etc. Each model was trained on the preprocessed dataset, with hyperparameters tuned to optimize performance.

4.5. Model Evaluation (Performance Metrics):

Model evaluation assesses performance using metrics including Accuracy, Precision, Recall, F1 score, and ROC AUC. Confusion matrices are utilized to visualize true positives, false positives, true negatives, and false negatives. The results provide insights into the model's predictive ability and generalization performance.

4.6. Model Comparison and Selection:

Models are compared based on their evaluation metrics, considering factors like interpretability, computational complexity, and scalability. The best-performing model is selected for further analysis or prediction tasks, ensuring the most effective approach for stroke risk assessment or prognosis.

By adhering to this workflow, the research aims to contribute to the advancement of stroke prediction models, ultimately enhancing clinical decision-making and patient outcomes in the management of stroke-related conditions.

5.Working of ExtraTrees Classifier:

5.1. Basic Idea:

The basic idea behind Extra Trees is to build multiple decision trees using random subsets of the training data and random subsets of the features. The predictions from these trees are then combined to make the final prediction.

5.2.How Extra Trees Differs from Random Forests:

5.2.1.Random Splitting Points*:

In Random Forests, the algorithm selects the best split among a random subset of features for each node in each tree. In contrast, Extra Trees randomly selects split points for each feature at each node, without searching for the best possible split. This randomness allows Extra Trees to build trees much faster than Random Forests.

5.2.2.Bootstrapping:

Random Forests typically use bootstrapping (sampling with replacement) to create multiple datasets for training each tree. Extra Trees, however, does not use bootstrapping. Instead, it trains each tree using the entire training dataset.

5.3 Algorithm Steps:

5.3.1. Random Feature Selection:

For each tree in the ensemble, a random subset of features is selected.

5.3.2. Random Splitting Points:

At each node of the tree, instead of finding the optimal split point, Extra Trees randomly selects a feature and a split point.

5.3.3 Building Trees:

Trees are grown until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples per leaf.

5.3.4. Aggregation:

To make predictions, each tree in the ensemble independently predicts the target variable, and the final prediction is often the average (for regression) or the mode (for classification) of these individual predictions.

5.4.Mathematical Formulation:

5.4.1.Entropy and Gini Impurity:

Extra Trees, like decision trees, can use metrics like entropy or Gini impurity to measure the homogeneity of a set of samples.

Entropy

$$H(s) = -P_{(} + \log_2 P_{(}+) - P_{(}-) \log_2 P_{(-}$$

Gini impurity:

$$GI = 1 - \sum_{i=1}^{n} (p)^2$$

$$GI = 1 - \left[(P_{(}+))^2 + (P_{(}-) \right)^2 \Big]$$

5.4.2.Splitting Criterion:
Extra Trees randomly selects feature and split points at each node. The splitting criterion is not based on minimizing entropy or Gini impurity, as in traditional decision trees. Instead, it's purely random.

5.4.3. Prediction Aggregation:
The final prediction in Extra Trees is typically the average (for regression) or the mode (for classification) of predictions made by individual trees.

6.Evaluation Metrics:

		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (**TN**)	False Positives (**FP**) **Type I error**
	Positive +	False Negatives (**FN**) **Type II error**	True Positives (**TP**)

Fig.7-Confusion matrix

True Positive(TP): Actual Positive and Predicted as Positive.
True Negative(TN): Actual Negative and Predicted as Negative.
False Positive(Type I Error)(FP): Actual Negative but predicted as Positive.
False Negative(Type II Error)(FN): Actual Positive but predicted as Negative.

6.1.Accuracy :
Accuracy measures the proportion of correct predictions among the total number of predictions.

$$\text{Accuracy} = \frac{\text{True Positives + True Negatives}}{\text{All Samples}}$$

6.2.Precision Score:
Precision measures the proportion of true positive predictions among all positive predictions made.

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

6.3. Recall Score (Sensitivity):
Recall measures the proportion of true positive predictions among all actual positives.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

6.4. F1 Score:
F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall.

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

6.5.ROC AUC Score (Receiver Operating Characteristic - Area Under Curve):
ROC AUC Score measures the area under the receiver operating characteristic curve. It is a performance measurement for classification problems at various threshold settings.

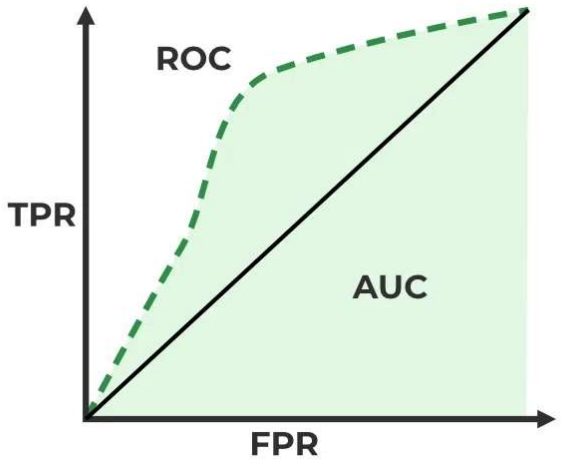

Fig.8- ROC AUC curve

TPR – True Positive Rate
FPR – False Positive Rate

7.Results and Discussion :

In evaluating the performance of an algorithm, various metrics are employed to provide a comprehensive understanding of its effectiveness. Accuracy, the simplest metric, measures the proportion of correct predictions out of the total predictions made. Sensitivity, also known as recall, focuses on the algorithm's ability to correctly identify positive instances from all actual positive instances. It highlights the algorithm's capability to avoid false negatives, crucial in scenarios where missing positive instances could have significant consequences, such as medical diagnoses.

Specificity complements sensitivity by measuring the algorithm's proficiency in correctly identifying negative instances from all actual negative instances. It emphasizes the importance of avoiding false positives, which can lead to unnecessary interventions.

F1 score offers a balanced measure of a model's precision and recall, providing a single value that captures both aspects of performance. It's particularly useful when there's an uneven class distribution or when both false positives and false negatives are equally important. Finally, the Matthews Correlation Coefficient (MCC) offers a more holistic assessment by taking into account all elements of the confusion matrix, including true positives, true negatives, false positives, and false negatives. It's particularly valuable when dealing with imbalanced datasets and is considered a reliable measure for binary classification tasks, such as distinguishing between benign and malign classifications in medical diagnoses. These metrics collectively provide a nuanced evaluation of an algorithm's performance, enabling informed decisions regarding its deployment and optimization.

The performance metrics of each algorithm is mentioned below:

	Precision	Recall	F-measure	Mcc	AUC	Accuracy
XGBoost	0.928	1	0.962	0.925	0.992	0.961
CatBoost	0.956	1	0.977	0.955	0.997	0.977
LightGBM	0.862	0.988	0.921	0.84	0.967	0.915
Bagging	0.972	1	0.985	0.971	0.99	0.985
Extra Trees	0.997	1	0.998	0.997	1	0.998

Fig.8-evaluation metrics of machine learning algorithms.

8.Conclusion and Future Work:

In the realm of healthcare, particularly in addressing life-threatening conditions like strokes, the integration of AI and machine learning presents a promising avenue for early detection and prevention. By leveraging established models, clinical providers and decision-makers can tap into vast datasets to discern pertinent features or risk factors associated with stroke occurrence. These features, extracted from patient profiles encompassing various demographic, lifestyle, and medical factors, serve as crucial inputs for predictive models.

The study mentioned delves into assessing the efficacy of different machine learning algorithms in predicting strokes accurately. Through the analysis of diverse algorithms, ranging from traditional statistical methods to more advanced techniques like neural networks, researchers aim to identify the most suitable approach for stroke prediction. By evaluating the performance of these algorithms against benchmark metrics such as accuracy, sensitivity, specificity, F1 score, and MCC, researchers can determine which model yields the most reliable predictions.

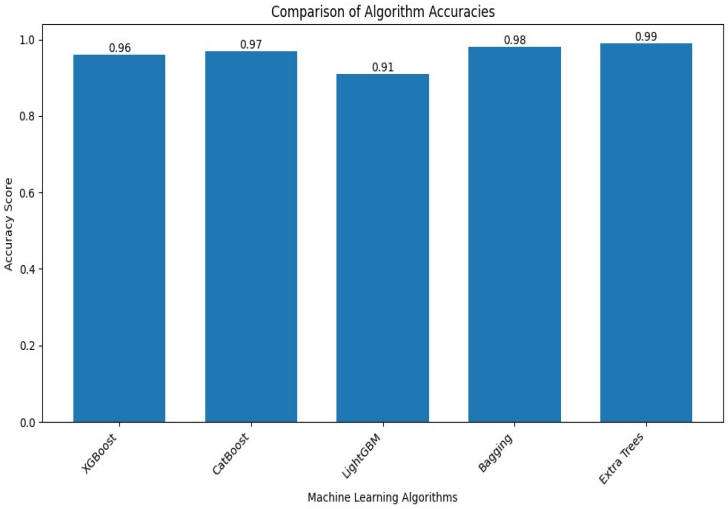

Fig.9-Performance analysis of classification models

Ultimately, the objective is to develop a robust predictive model capable of accurately forecasting stroke risk based on individual profiles. Such a model holds significant potential for early intervention and personalized preventive measures, thereby mitigating the severe consequences associated with strokes. By harnessing the power of machine learning, healthcare professionals can enhance their ability to preemptively identify individuals at risk, thus facilitating timely interventions and improving patient outcomes.

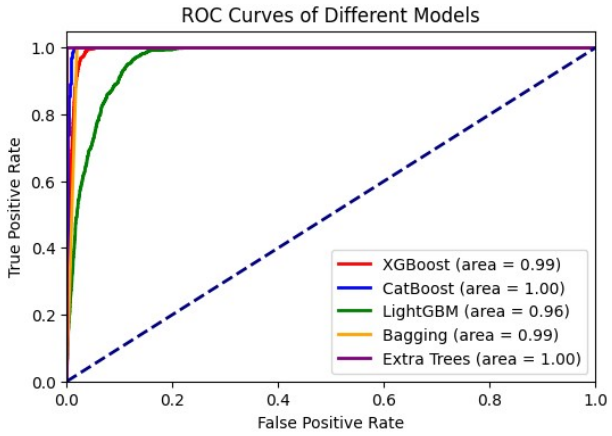

Fig.10-ROC Curves of different models

References:

[1] Hanifa SM, Raja SK. Stroke risk prediction through non-linear support vector classification models. Int. J. Adv. Res. Comput. Sci., 2010; 1 [1]

[2] Jeena RS, Kumar S. Stroke prediction using SVM, International Conference on Control. Instrumentation, Communication and Computational Technologies (ICCICCT), 2016: 600-602.[2]

[3] Dritsas E. , & Trigka M. (2022). Stroke risk prediction with machine learning techniques. Sensors, 22(13), 4670.[3]

[4] Sailasya G and Kumari G. L. A. Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science And Applications. 2021; 12(6): 539–545.[4]

[5] Tavares J-A. Stroke prediction through Data Science and Machine Learning Algorithms. 2021; doi: .13140/RG.2.2.33027.43040. [5]

[6] Shoily, T.I.; Islam, T.; Jannat, S.; Tanna, S.A.; Alif, T.M.; Ema, R.R. Detection of stroke disease using machine learning algorithms. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT),Kanpur, India, 6–8 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.[6]

[7] Govindarajan, P.; Soundarapandian, R.K.; Gandomi, A.H.; Patan, R.; Jayaraman, P.; Manikandan, R. Classification of stroke disease using machine learning algorithms. Neural Comput. Appl. 2020, 32, 817–828. [7]

[8] Rahman, S., Hasan, M. and Sarkar, A.K. 2023. Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. European Journal of Electrical Engineering and Computer Science. 7, 1 (Jan. 2023), 23–30.[8]

[9] Hung CY, Lin CH, Lan TH, Peng GS, Lee CC. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017: 3110–3113. [9]

[10] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin and M. F. Mridha, Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention, in IEEE Access, vol. 11, pp. 52288-52308, 2023, doi: 10.1109/ACCESS.2023.3278273.[10]

[11] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas and U. Sara, "A Machine Learning Approach to Detect the Brain Stroke Disease," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 897-901, doi: 10.1109/ICSSIT53264.2022.9716345. [11]

[12] Tusher, A. N., Sadik, M. S., & Islam, M. T. (2022, December). Early brain stroke prediction using machine learning. In 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 1280-1284). IEEE. [12]

[13] Adam SY, Yousif A, Bashir MB. Classification of ischemic stroke using machine learning algorithms. International Journal of Computer Application, 2016;149(10):26–31.[13]

[14] N. S. Adi, R. Farhany, R. Ghina and H. Napitupulu, Stroke Risk Prediction Model Using Machine Learning, 2021 International Conference on Artificial Intelligence and Big Data Analytics, Bandung, Indonesia, 2021, pp. 56-60, doi: 10.1109/ICAIBDA53487.2021.9689731.[14]

[15] Biswas, N., Uddin, K. M. M., Rikta, S. T., & Dey, S. K. (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics, 2, 100116.[15]

[16] Bandi, V., Bhattacharyya, D., & Midhunchakkravarthy, D. (2020). Prediction of Brain Stroke Severity Using Machine Learning. Rev. d'Intelligence Artif., 34(6), 753-761.[16]

[17] S. Gupta and S. Raheja, Stroke Prediction using Machine Learning Methods, 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 553-558, doi: 10.1109/Confluence52989.2022.9734197.[17]

[18] Puranjay Savar Mattasa aORCID ID: https://orcid.org/0000-0002-5314-5647, #129/2 Sharda University, New Delhi 110030,IndiaDOI:https://doi.org/10.55248/gengpi.2022.31211[18]

[19] Liuzzi, Piergiuseppe, Antonello Grippo, Silvia Campagnini, Maenia Scarpino, Francesca Draghi, Annamaria Romoli, Bahia Hakiki et al. Merging clinical and EEG biomarkers in an elastic-net regression for disorder of consciousness prognosis prediction. IEEE Transactions on Neural Systems and Rehabilitation Engineering 30 (2022): 1504-1513.

[20] Calesella F, Testolin A, De Filippo De Grazia M, Zorzi M. A comparison of feature extraction methods for prediction of neuropsychological scores from functional connectivity data of stroke patients. Brain Inform. 2021 Apr 20;8(1):8. doi: 10.1186/s40708-021-00129-1. PMID: 33877469; PMCID: PMC8058135.