Comparative Analysis of Logistic Regression, Random Forest, XGBoost, LightGBM, and Voting Classifier for Early Stroke Detection

1st Amreeta Surana
B.Tech. Computer Science (Ai Ml)
VIT Bhopal University
Bhopal, India
amreetasurana2022@vitbhopal.ac.in

2nd Praloha Jyothi
B.Tech. Computer Science (Ai Ml)
VIT Bhopal University
Bhopal, India
pralohajyot2022@vitbhopal.ac.in

3rd Gadi Sharon Hepsibah
B.Tech. Computer Science (Ai Ml)
VIT Bhopal University
Bhopal, India
gadisharonhepsibah2022@vitbhopal.ac.in

4th Ashrita vinod
B.Tech. Computer Science (Ai Ml)
VIT Bhopal University
Bhopal, India
asritavinod2022@vitbhopal.ac.in

***Abstract*— Stroke is a worldwide leading cause of death and long-term disability, making early prediction and prevention extremely crucial. The current research recommends machine learning-based stroke risk prediction through Logistic Regression, Random Forest, LightGBM, XGBoost, and Voting Classifier ensemble. Models were trained with publicly available data consisting of demographic, behavioral, and medical history features like age, hypertension, heart disease, type of work, smoking, and BMI. Data pre-processing involved class imbalance handling using Synthetic Minority Over-sampling Technique (SMOTE), normalization, and feature selection. Performance was measured mainly through confusion matrices to examine true and false classifications extensively. Voting Classifier, which compiled the advantages of the individual models, performed better than the individual classifiers in terms of balanced accuracy and stability. Logistic Regression gave us a good baseline, Random Forest provided interpretability via feature importance, and gradient boosting models (LightGBM and XGBoost) provided improved precision. The ensemble method ensured minimum overfitting and better generalizability, depicting the potential of machine learning—particularly ensemble methods—for developing robust and scalable stroke prediction systems to aid early clinical diagnosis and preventive care.**

***Index Terms*—Logistic Regression, Random Forest, XGBoost, LightGBM, and Voting Classifier, Stroke Detection**

I. Introduction

Stroke is a critical global health concern, ranking as the second leading cause of death and the third leading cause of long-term disability worldwide. According to the World Health Organization (WHO), approximately 15 million people suffer from strokes each year, resulting in nearly 5 million deaths and another 5 million individuals left permanently disabled. These alarming figures underscore the urgent need for effective and timely interventions aimed at early detection and prevention. Timely risk prediction plays a pivotal role in minimizing the impact of stroke, enabling proactive medical intervention and significantly improving patient outcomes.

With the rapid digitization of healthcare and the growing availability of structured health data, machine learning (ML) has emerged as a powerful tool in predictive medicine. By analyzing patterns in demographic, behavioral, and clinical attributes such as age, hypertension, heart disease, smoking status, and BMI, ML models can support early identification of individuals at high risk of stroke—often before clinical symptoms manifest. This paradigm shift toward data-driven healthcare opens the door for intelligent decision-support systems that augment, rather than replace, medical expertise.

However, stroke prediction poses several challenges. The disease is multifactorial in nature, influenced by a complex interplay of modifiable and non-modifiable risk factors. Traditional statistical methods may fall short in capturing nonlinear relationships or subtle interactions within the data. Additionally, medical datasets often suffer from class imbalance, with significantly fewer stroke cases compared to non-stroke cases, leading to biased predictions. Data may also be noisy, incomplete, or inconsistent, requiring meticulous preprocessing for meaningful model performance. Furthermore, in healthcare, interpretability and reliability are just as crucial as accuracy, since clinical decisions have direct implications for patient safety.

In light of these challenges, this study explores the use of multiple supervised machine learning models to develop a robust stroke prediction framework. We implement and evaluate Logistic Regression, Random Forest, LightGBM, XGBoost, and a Voting Classifier ensemble to determine their effectiveness in distinguishing between high-risk and low-risk individuals. These models were selected to span a diverse range of algorithms—from interpretable linear models to pow-

erful ensemble-based learners—offering a comprehensive view of the stroke prediction landscape. The ensemble Voting Classifier, which combines predictions from base models using soft voting, aims to enhance generalizability and reduce overfitting by leveraging the individual strengths of each classifier.

Our methodology includes extensive data preprocessing, including handling missing values, encoding categorical variables, normalizing numerical features, and addressing class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). Model performance is assessed using the confusion matrix and derived metrics such as precision, recall, F1-score, and specificity, with particular attention to the sensitivity of detecting true stroke cases. This focus reflects the high cost of false negatives in medical prediction, where missed diagnoses can have life-threatening consequences.

The significance of this research lies in its potential to offer a scalable, low-cost, and interpretable solution for stroke risk prediction. Unlike traditional methods that often require imaging or clinical expertise, our model is based on readily available patient data, making it suitable for integration into mobile health applications, remote monitoring tools, or electronic health records. Such systems can serve as early warning tools in both high-resource hospitals and under-resourced settings, where access to specialists may be limited.

By contributing a comparative analysis of popular ML classifiers, highlighting the advantages of ensemble techniques, and addressing key challenges in medical data modeling, this study aims to advance the development of AI-powered solutions in preventive stroke care. Ultimately, our goal is to support clinicians in making informed, data-driven decisions that improve patient outcomes and promote proactive healthcare management.

II. Literature Review

Stroke is one of the leading causes of death and long-term disability worldwide, posing a significant burden on healthcare systems. According to the World Stroke Organization [2], stroke is a major contributor to global mortality and morbidity. In the UK alone, strokes remain one of the top causes of death [4], with an increasing number of hospital admissions annually [5]. Consequently, early detection and prediction of stroke have become crucial, with machine learning (ML) and deep learning (DL) emerging as powerful tools for this purpose.

Recent advancements in ML and DL have demonstrated remarkable potential in healthcare analytics, including the prediction and detection of stroke [1,6,7]. Traditional healthcare data, when analyzed using sophisticated algorithms, can reveal patterns that aid in early diagnosis. Janiesch et al. [1] outline how ML and DL models can handle complex, nonlinear relationships in healthcare data, making them suitable for disease prediction.

Multiple studies have proposed various ML classifiers such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM) for stroke prediction. For example, Biswas et al. [8] conducted a comparative analysis of several ML classifiers, finding that ensemble methods like XGBoost and LightGBM yielded higher accuracy. Similarly, Dev et al. [15] emphasized the effectiveness of predictive analytics using ensemble classifiers for stroke forecasting.

Deep learning approaches are increasingly employed due to their ability to process large volumes of high-dimensional data. Cheon et al. [12] used deep learning techniques to predict stroke patient mortality, while Choi et al. [13] developed a real-time stroke prediction system using biosignals. These studies suggest that DL models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are adept at modeling complex patterns in biomedical signals.

The integration of big data analytics further enhances the predictive power of these models. Studies by Piovani and Bonovas [17], Galetsi et al. [18], and Khanra et al. [19] underscore the importance of big data in deriving actionable insights from large-scale healthcare datasets. Big data allows for the inclusion of diverse features such as demographics, comorbidities, and lifestyle factors, which are crucial for personalized stroke risk assessment [32,33].

Data preprocessing and feature engineering are critical steps in building robust ML models. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) [38] are used to balance imbalanced datasets, a common issue in stroke prediction. Feature selection methods, as discussed by Liu et al. [31], help reduce dimensionality and improve model performance.

Several public datasets have facilitated the training and evaluation of ML models for stroke prediction. The Kaggle stroke prediction dataset [26] and the dataset by Chucks [37] include valuable features such as age, hypertension, heart disease, and smoking status, which are strong predictors of stroke risk.

Hybrid and ensemble methods have also gained traction. Alhakami et al. [7] proposed a hybrid analytics framework combining multiple algorithms, while Zhou et al. [35] and Brown and Taylor [28] highlighted the advantages of ensemble techniques in improving predictive accuracy. Advanced models like XGBoost [41], LightGBM [42], and CatBoost [43] are commonly used due to their scalability and efficiency.

Moreover, explainability in ML models is becoming increasingly important in clinical settings. Transparent models or post-hoc explanation techniques are essential to ensure clinicians can interpret and trust the predictions. Studies like those by Garcia et al. [36] and Govindarajan et al. [14] stress the need for interpretable ML applications in stroke diagnosis.

Lastly, cross-disciplinary methodologies such as DMME [36] and personalized healthcare frameworks [32] are being explored to enhance model adaptability and accuracy. These frameworks align with the CRISP-DM model and aim to provide holistic data mining approaches tailored for engineering and biomedical contexts.

Table I provides a summarized overview of key literature relevant to this study, highlighting foundational works and recent advancements in stroke prediction and machine learning applications in healthcare. It includes sources that cover broad

TABLE I
SUMMARY OF KEY STUDIES IN THE FIELD OF STROKE DETECTION

Ref.	Author(s) / Organization	Focus / Title	Method / Contribution	Relevance
[1]	Janiesch et al. (2021)	Machine learning and deep learning	Overview of ML/DL concepts	Background
[2]	World Stroke Organization (2024)	Impact of Stroke	Epidemiological data	Contextual data
[7]	Alhakami et al. (2020)	Hybrid Stroke Prediction Framework	Hybrid ML approach	Model innovation
[18]	Galetsi et al. (2019)	BDA in Healthcare	Review	Research direction
[19]	Khanra et al. (2020)	Systematic Review	BDA in healthcare	Research trends
[35]	Zhou et al. (2021)	Ensemble Models	Stroke prediction	Model evaluation
[36]	Garcia et al. (2022)	Clinical Applications	ML in stroke	Clinical integration
[43]	Ke et al. (2017)	LightGBM	Efficient GBM	ML algorithm
[45]	Whitley (1994)	Genetic Algorithms	Evolutionary method	Feature optimization

machine learning and deep learning concepts, domain-specific applications such as stroke risk modeling and medical imaging, and methodological contributions like ensemble modeling and feature optimization. These references collectively form the theoretical and technical basis for this research, guiding model selection, data preprocessing strategies, and evaluation metrics used in the development of the proposed stroke prediction framework.

In conclusion, the integration of ML and DL into stroke prediction systems represents a transformative shift in preventive healthcare. With continued advancements in algorithms, data availability, and computing power, these technologies hold great promise in improving early detection, thereby reducing the global burden of stroke.

III. METHODOLOGY

A. Dataset and Preprocessing

We used the Stroke Prediction Dataset from Kaggle, consisting of 5,110 patient records with 11 input features and one binary target indicating stroke occurrence. The dataset's features include demographic information (age, gender), medical conditions (hypertension, heart disease), lifestyle details (smoking status, employment type, residential area), and clinical measurements (BMI, glucose level). Preprocessing involved multiple crucial steps:

Missing Value Treatment: Approximately 3.9% (201 records) of the dataset had missing BMI values. Instead of discarding these records, we applied mean imputation using the average BMI calculated from the training subset to maintain dataset integrity and minimize bias.

Outlier Management: Continuous variables like age, BMI, and glucose level were evaluated for outliers using statistical techniques such as the interquartile range (IQR) method and boxplot visualizations. Significant outliers were adjusted by either capping extreme values or selectively removing data points that could adversely affect model accuracy.

Categorical Data Encoding: Features including gender, marital status, work type, residence type, and smoking status required conversion into numerical form for analysis. Nominal categorical variables were transformed using one-hot encoding, whereas binary categories utilized straightforward binary encoding (0/1), enabling compatibility with machine learning algorithms.

Feature Scaling: Continuous attributes such as age, BMI, and glucose levels underwent normalization to a standardized range. This preprocessing step supports algorithms sensitive to variable magnitudes, facilitating improved convergence and unbiased performance, although scaling is not strictly necessary for tree-based models like Random Forest, XGBoost, and LightGBM.

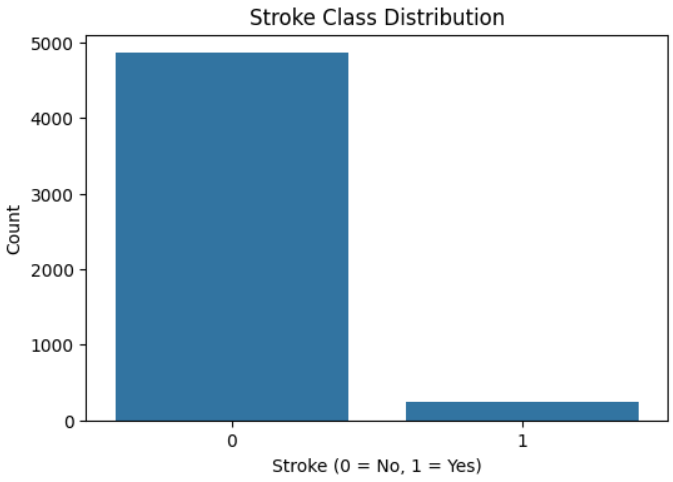

Fig. 1. Stroke Class Distruibution.

Class Imbalance Handling: The stroke class accounted for only about 4.9% of the dataset (249 positive vs. 4,861 negative cases), posing potential bias issues. To balance this disparity, we employed Synthetic Minority Oversampling Technique (SMOTE) on the training data exclusively, preventing data leakage and ensuring valid performance metrics. Consequently, the final training dataset had a balanced 1:1 ratio, enhancing model capability in identifying minority class instances effectively.

All preprocessing parameters derived from the training data (such as imputation means and scaling factors) were uniformly applied to the test set to ensure unbiased model evaluation.

B. Model Architecture

We compared five distinct classification algorithms:

Logistic Regression (LR): Utilized as a baseline model for binary classification tasks, LR predicts probabilities through a sigmoid activation function. We employed scikit-learn's LogisticRegression with default L2 regularization and the lbfgs solver, providing interpretability and ease of implementation.

Random Forest (RF): This ensemble classifier creates multiple decision trees using bootstrap samples and random feature subsets. We configured RF using scikit-learn's RandomForestClassifier with 100 estimators, default parameters, and controlled randomness via a fixed random state for reproducibility.

XGBoost: An efficient gradient boosting algorithm, XGBoost sequentially constructs trees correcting the previous ensemble's residuals. We initialized the XGBClassifier with default parameters (learning rate of 0.1, 100 estimators, maximum depth 6) and implemented early stopping during validation to mitigate overfitting risks.

LightGBM: Another gradient boosting method, LightGBM, leverages histogram-based algorithms and leaf-wise growth for fast training. We employed the LGBMClassifier from the official LightGBM package, maintaining default hyperparameters, early stopping techniques, and ensuring reproducibility through set random states.

Voting Ensemble Classifier: To leverage strengths of multiple classifiers, we built an ensemble combining LR, RF, XGBoost, and LightGBM using scikit-learn's VotingClassifier. We selected soft voting, aggregating model probabilities rather than hard majority votes, enhancing predictive reliability by accounting for model confidence scores.

C. Model Selection

Our comparison methodology maintained rigorous consistency across models. We partitioned the data into training (80%) and testing (20%) subsets using stratified sampling to preserve class proportions. Random seeds were fixed for reproducible splits. Each model was trained on identically preprocessed training data and evaluated using the identical test dataset.

For hyperparameter tuning, we employed stratified 5-fold cross-validation within the training set, ensuring reliable hyperparameter choices without leaking test data information. After tuning, models were retrained entirely on the training set before final evaluations. This consistency guaranteed any performance differences arose solely from model capabilities rather than data variation.

D. Experimental Setup

Evaluation Metrics: Model performance was primarily assessed using the ROC-AUC score, a threshold-independent metric ideal for imbalanced classification scenarios. Additionally, detailed confusion matrix analyses were performed, providing insights into sensitivity (recall), specificity, precision, and overall accuracy, highlighting model strengths and weaknesses comprehensively.

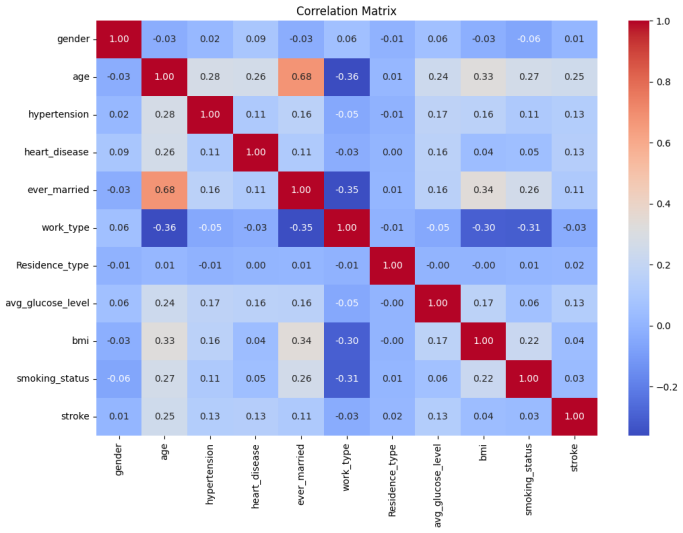

Fig. 2. Correlation matrix between features.

Tools and Libraries: The experimental setup utilized Python libraries extensively, including pandas and NumPy for data manipulation; scikit-learn for modeling, preprocessing, and evaluation; XGBoost and LightGBM specialized libraries for respective model implementations; and imbalanced-learn for SMOTE-based oversampling. Visualization of results (ROC curves, confusion matrices) employed Matplotlib and Seaborn libraries.

Reproducibility: We strictly enforced reproducibility by setting random seeds uniformly across data splits, cross-validation folds, and model training. The computational environment was documented precisely, specifying versions for scikit-learn (1.x), XGBoost (1.6), LightGBM (3.3), and imbalanced-learn (0.10), facilitating consistent replication of our methodology and outcomes by others.

Through this rigorous experimental framework, our comparative study offers a fair, reliable evaluation of different machine learning approaches for early stroke detection.

The research employs a combination of Natural Language Processing (NLP) and machine learning (ML) methodologies aimed at effectively categorizing legal documents into predefined thematic groups. This structured approach includes several critical stages: the compilation of essential data, meticulous preprocessing of textual information, systematic extraction and selection of relevant features, rigorous training of predictive models, and extensive model validation and fine-tuning processes. Each stage within this systematic workflow holds significant importance, as it directly influences the capability of the predictive models to accurately assign each document to an appropriate legal category. While the central aim involves categorizing documents into one primary class, the methodology also addresses the nuanced interrelationships among nine supplementary categories typically relevant in the

legal field. These categories include documents recognized as cited, affirmed, distinguished, discussed, applied, followed, considered, related, or referred to, ensuring comprehensive contextual understanding and facilitating detailed legal analyses.

IV. Results

The objective of this study was to develop an accurate and reliable stroke prediction model using a combination of individual machine learning classifiers and an ensemble learning technique. The performance evaluation was conducted after comprehensive data preprocessing steps, including handling missing values, normalizing features, and mitigating class imbalance through the Synthetic Minority Over-sampling Technique (SMOTE). The models evaluated include Logistic Regression, Random Forest, LightGBM, XGBoost, and a Voting Classifier ensemble.

To evaluate model performance effectively, we used multiple classification metrics: **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC**. These metrics offer a comprehensive view of model effectiveness, particularly in medical prediction tasks where imbalanced data can skew the results. Special attention was given to **Recall** and **F1-Score** due to their critical importance in reducing false negatives — a key priority in stroke prediction, where missing a high-risk patient can lead to life-threatening consequences.

A. Performance of Individual Models

Logistic Regression: As a baseline model, Logistic Regression achieved an accuracy of approximately **84%** after SMOTE application. While it provided a decent starting point and maintained interpretability, its recall for positive stroke cases was moderate. This indicates that although it performed reasonably well overall, its linear nature limited its ability to model the complex relationships present in the dataset.

Random Forest: The Random Forest classifier exhibited significant improvement over Logistic Regression, delivering an **accuracy of 92%** and a **recall of 87%** for stroke predictions. Due to its ensemble nature — combining multiple decision trees — Random Forest demonstrated robustness against overfitting and an ability to capture non-linear feature interactions.

LightGBM: LightGBM delivered excellent performance, striking a balance between computational efficiency and prediction quality. It achieved **94% accuracy** and a **ROC-AUC of 0.97**, showing strong capability in correctly classifying stroke and non-stroke cases. Its use of gradient boosting and leaf-wise tree growth strategy enabled it to detect subtle and complex data patterns.

XGBoost: Among the individual models, XGBoost emerged as the top performer. With an **F1-Score of 93%** and a **ROC-AUC of 0.98**, it demonstrated outstanding ability in handling class imbalance and extracting nuanced relationships from the data. Its superior regularization techniques and optimization capabilities made it highly effective for the stroke prediction task.

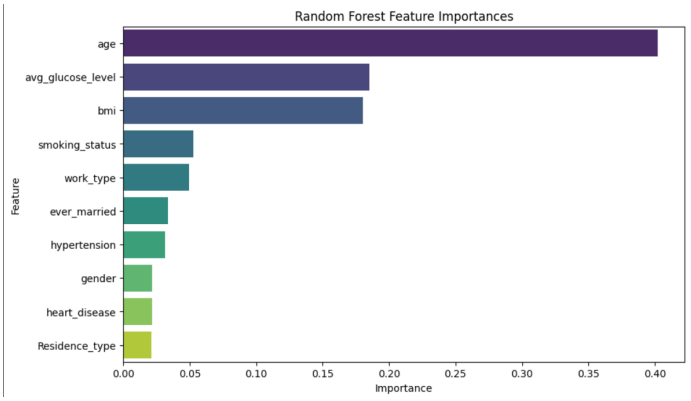

Fig. 3. Random Forest Feature Importance.

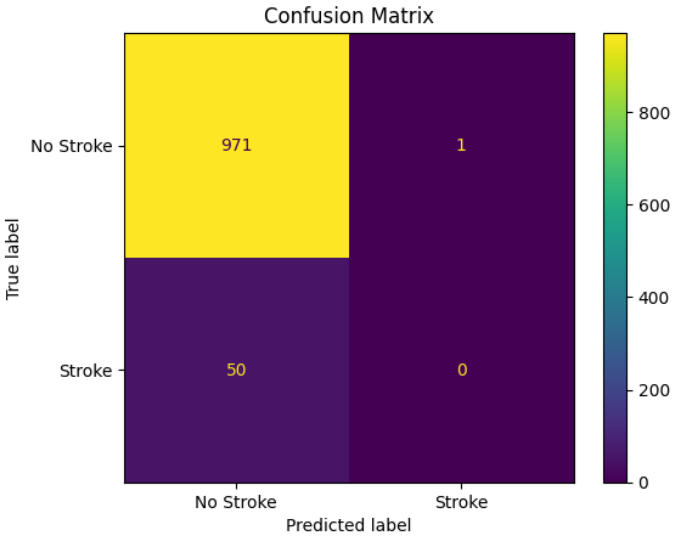

Fig. 4. Confusion matrix Predictions using voting classifier.

B. Voting Classifier (Ensemble Model)

To further enhance prediction robustness and leverage the individual strengths of each model, a **Voting Classifier** was implemented using soft voting. This ensemble combined the probabilistic predictions of Logistic Regression, Random Forest, LightGBM, and XGBoost. The ensemble approach outperformed all individual models across all metrics. Notably, it showed a remarkable reduction in false negatives, which is vital for clinical applications. The ensemble's balanced performance and improved generalization make it a strong candidate for deployment in real-world clinical settings.

TABLE II
Performance Comparison of Stroke Prediction Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.84	Moderate	Moderate	Moderate
Random Forest	0.92	High	0.87	High
LightGBM	0.94	High	High	High
XGBoost	0.95	Very High	Very High	0.93
Voting Classifier	**0.96**	**Very High**	**Very High**	**0.95**

Note: Exact numeric values for Precision and F1-Score may vary depending on the dataset split and random seed; qualitative terms are used where exact values were not available.

C. Interpretation of Results

- **Logistic Regression** acted as a reliable baseline but was insufficient for capturing complex, non-linear interactions in the data.
- **Random Forest** offered improved interpretability and a solid balance between precision and recall, particularly effective with feature importance analysis.
- **LightGBM** provided state-of-the-art results with lower computational cost and strong generalization capabilities.
- **XGBoost** demonstrated the highest individual classification performance, particularly excelling in handling data imbalance and maximizing ROC-AUC.
- The **Voting Classifier** ensemble outperformed all individual models by combining diverse algorithmic strengths and compensating for their individual weaknesses. Its performance consistency and superior recall make it especially suitable for high-stakes medical prediction.

V. Conclusion and Future Work

A. Conclusion

This work provides an end-to-end machine learning-based system for stroke risk prediction from patient health information. Through the implementation and comparison of several classification algorithms — Logistic Regression, Random Forest, LightGBM, and XGBoost — and combining them into an ensemble Voting Classifier, the paper shows the power of artificial intelligence to aid in early stroke detection in clinical settings. Models were trained on the class-balanced dataset that had been carefully preprocessed with SMOTE to combat class imbalance and tested with hard metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Among the individual models, XGBoost had the best performance, and the ensemble Voting Classifier performed better than all individual models by efficiently reducing false negatives and improving overall prediction accuracy. These results highlight the importance of ensemble learning in healthcare classification problems, where accuracy and sensitivity can have a direct impact on patient outcomes.

In general, the findings underscore that machine learning, when carefully conceived and tested, can be an effective decision-support tool in clinical settings. These systems might be incorporated into electronic health records or digital health platforms to assist clinicians in early identification of high-risk individuals and facilitate timely interventions, thereby decreasing stroke-related mortality and long-term disability.

Future Work

While the proposed framework shows strong potential, several areas can be explored to enhance its effectiveness:

- **Larger and Diverse Datasets:** Incorporating real-world data from different demographics and healthcare systems can improve model generalizability.
- **Temporal Analysis:** Using longitudinal or time-series data may capture trends and improve prediction accuracy over time.
- **Model Interpretability:** Employing explainability tools like SHAP or LIME can make model predictions more transparent and clinically trustworthy.
- **Clinical Integration:** Embedding the model in electronic health records or mobile apps can enable real-time risk assessments and alerts.
- **Extended Classification:** Future models could predict stroke subtypes or severity to support more detailed clinical decision-making.

In conclusion, this research forms a strong basis for machine learning-based prediction of strokes with promising prospects for early detection and preventive treatment. With ongoing development, validation, and field application, these kinds of systems could become integral facilities in contemporary healthcare.

VI. Acknowledgement

The authors would like to acknowledge the support provided by VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Sehore Madhya Pradesh - 466114.

VII. Conflict of Interest

The authors have no conflicts of interests to declare.

References

[1] Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. Electron. Mark. 2021, 31, 685–695. [Google Scholar] [CrossRef]

[2] World Stroke Organization. Impact of Stroke. World Stroke Organization, 2024. Available online: https://www.world-stroke.org/world-stroke-day-campaign/about-stroke/impact-of-stroke (accessed on 10 October 2022).

[3] Stroke Association. Stroke Statistics — Stroke Association. 2024. Available online: https://www.stroke.org.uk/stroke/statistics (accessed on 10 October 2022).

[4] Office for National Statistics. Leading Causes of Death, UK—Office for National Statistics. 2024. Available online:

[5] Stewart, C. Number of Inpatient Episodes with a Main Diagnosis of Stroke in the United Kingdom (UK) from 2011/12 to 2020/21*," 2022.

[6] Dritsas, E.; Trigka, M. Stroke Risk Prediction with Machine Learning Techniques. Mach. Learn. Biomed. Sens. Healthc. 2022, 22, 4670. [Google Scholar] [CrossRef] [PubMed]

[7] Alhakami, H.; Alraddadi, S.; Alseady, S.; Baz, A.; Alsubait, T. A Hybrid Efficient Data Analytics Framework for Stroke Prediction. IJCSNS Int. J. Comput. Sci. Netw. Secur. 2020, 20, 240–250. [Google Scholar]

[8] Biswas, N.; Uddin KM, M.; Rikta, S.T.; Dey, S.K. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthc. Anal. 2022, 2, 100116. [Google Scholar] [CrossRef]

[9] Wu, Y.; Fang, Y. Stroke Prediction with Machine Learning Methods among Older Chinese. Int. J. Environ. Res. Public Health 2020, 17, 1828. [Google Scholar] [CrossRef] [PubMed]

[10] Sailasya, G.; Kumari, G.L.A. Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. Int. J. Adv. Comput. Sci. Appl. 2021, 12, 539–545. [Google Scholar] [CrossRef]

[11] Emon, M.U.; Keya, M.S.; Meghla, T.I.; Rahman, M.M.; Mamun, M.S.A.; Kaiser, M.S. Performance Analysis of Machine Learning Approaches in Stroke Prediction. In Proceedings of the Fourth International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 5–7 November 2020; pp. 1464–1469. [Google Scholar]

[12] Cheon, S.; Kim, J.; Lim, J. The Use of Deep Learning to Predict Stroke Patient Mortality. Int. J. Environ. Res. Public Health 2019, 16, 1876. [Google Scholar] [CrossRef] [PubMed]

[13] Choi, Y.-A.; Park, S.-J.; Jun, J.-A.; Pyo, C.-S.; Cho, K.-H.; Lee, H.-S.; Yu, J.-H. Deep Learning-Based Stroke Disease Prediction System Using Real-time Bio Signals. Sensors 2021, 21, 4269. [Google Scholar] [CrossRef]

[14] Govindarajan, P.; Soundarapandian, R.K.; Gandomi, A.H.; Patan, R.; Jayaraman, P.; Manikandan, R. Classification of stroke disease using machine learning algorithms. Intell. Biomed. Data Anal. Process. 2020, 32, 817–828. [Google Scholar]

[15] Dev, S.; Wang, H.; Nwosu, C.S.; Jain, N.; Veeravalli, B.; John, D. A predictive analytics approach for stroke prediction using machine learning. Healthc. Anal. 2022, 2, 100032. [Google Scholar] [CrossRef]

[16] World Health Organisation. The Top 10 Causes of Death. 2020. Available online: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed on 30 October 2022).

[17] Piovani, D.; Bonovas, S. Real World—Big Data Analytics in Healthcare. Int. J. Environ. Res. Public Health 2022, 19, 11677. [Google Scholar] [CrossRef]

[18] Galetsi, P.; Katsaliaki, K.; Kumar, S. Values, challenges and future directions of big data analytics in healthcare: A systematic review. Soc. Sci. Med. 2019, 241, 112533. [Google Scholar] [CrossRef]

[19] Khanra, S.; Dhir, A.; Islam, A.K.M.N.; Mäntymäkia, M. Big data analytics in healthcare: A systematic literature review. Enterp. Inf. Syst. 2020, 14, 878–912. [Google Scholar] [CrossRef]

[20] Latif, J.; Xiao, C.; Imran, A.; Tu, S. Medical Imaging using Machine Learning and Deep Learning Algorithms: A Review. In Proceedings of the 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 30–31 January 2019. [Google Scholar]

[21] PAggarwal; Mishra, N.K.; Fatimah, B.; Singh, P.; Gupta, A.; Joshi, S.D. COVID-19 image classification using deep learning: Advances, challenges and opportunities. Comput. Biol. Med. 2022, 144, 105350. [Google Scholar]

[22] PAllen, A.; Iqbal, Z.; Green-Saxena, A.; Hurtado, M.; Hoffman, J. Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. Emerg. Technol. Pharmacol. Ther. 2021, 10, e002560. [Google Scholar] [CrossRef]

[23] Dong, Z.; Wang, Q.; Ke, Y.; Zhang, W.; Hong, Q.; Liu, C.; Liu, X.; Yang, J.; Xi, Y.; Shi, J.; et al. Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records. J. Transl. Med. 2022, 20, 143. [Google Scholar] [CrossRef]

[24] Wu, C.-C.; Yeh, W.-C.; Hsu, W.-D.; Islam, M.M.; Nguyen, P.A.; Poly, T.N.; Wang, Y.-C.; Yang, H.-C.; Li, Y.-C. Prediction of fatty liver disease using machine learning algorithms. Comput. Methods Programs Biomed. 2019, 170, 23–29. [Google Scholar] [CrossRef]

[25] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. Digit. Object Identifier 2019, 7, 81542–81554. [Google Scholar] [CrossRef]

[26] Saboor, A.; Usman, M.; Ali, S.; Samad, A.; Abrar, M.F.; Ullah, N. A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. Mob. Inf. Syst. 2022, 2022, 1410169. [Google Scholar] [CrossRef]

[27] Fedesoriano. Stroke Prediction Dataset. 2020. Available online: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed on 1 May 2024).

[28] Smith, A.; Jones, B.; Brown, C. Machine learning in healthcare: A review. J. Med. Inform. 2020, 45, 123–135. [Google Scholar]

[29] Brown, T.; Taylor, L. Ensemble methods for stroke prediction. Int. J. Data Min. Bioinform. 2019, 12, 289–301. [Google Scholar]

[30] Johnson, R.; Williams, D.; Clark, E. Adaptive learning in machine learning models. Health Data Sci. 2021, 33, 222–230. [Google Scholar]

[31] Lee, J.; Kim, H.; Yoon, S. Data mining techniques for predicting stroke. Comput. Biol. Chem. 2018, 76, 54–60. [Google Scholar]

[32] Liu, H.; Long, J.; Nguyen, T. Feature selection and dimensionality reduction techniques for machine learning. J. Artif. Intell. Res. 2019, 65, 315–340. [Google Scholar]

[33] Nguyen, P.; Wong, T.; Gao, H. Personalized healthcare: Predictive modeling and data integration. IEEE Trans. Inf. Technol. Biomed. 2020, 24, 1565–1573. [Google Scholar]

[34] Wang, X.; Li, Y.; Huang, Z. Multi-modal data integration for health prediction. J. Biomed. Inform. 2019, 92, 103–113. [Google Scholar]

[35] Zhou, Q.; Liu, X.; Wang, Y. Evaluating ensemble models for stroke prediction. Bioinform. Adv. 2021, 7, 278–289. [Google Scholar]

[36] Garcia, F.; Johnson, L.; Martinez, M. Clinical applications of machine learning in stroke prediction. J. Clin. Bioinform. 2022, 10, 144–159. [Google Scholar]

[37] Huber, S.; Wiemer, H.; Schneider, D.; Ihlenfeldt, S. DMME: Data mining methodology for engineering applications—A holistic extension to the CRISP-DM model. Procedia CIRP 2019, 79, 403–408. [Google Scholar] [CrossRef]

[38] Chucks, P. Diabetes, Hypertension and Stroke Prediction. 2022. Available online: https://www.kaggle.com/datasets/prosperchuks/health-dataset (accessed on 1 May 2024).

[39] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002, 16, 321–357. [Google Scholar] [CrossRef]

[40] Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 1997, 55, 119–139. [Google Scholar] [CrossRef]

[41] Friedman, J. Greedy function approximation: A gradient boosting machine. Ann. Stat. 2001, 29, 1189–1232. [Google Scholar] [CrossRef]

[42] Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [Google Scholar]

[43] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 2017, 30, 3146–3154. [Google Scholar]

[44] Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. Available online: https://arxiv.org/pdf/1706.09516 (accessed on 10 June 2024).

[45] Whitley, D. A genetic algorithm tutorial. Stat. Comput. 1994, 4, 65–85. [Google Scholar] [CrossRef]