An Advanced Hybrid Machine Learning Framework for Stroke Risk Prediction

Pooja Dudhe
Department of Data Science Symbiosis Skills and Professional University Pune, Maharashtra, India
Guide: Prof. Shubhangi Tikade
Co-Guide: Prof. Prashant Kulkarni

***Abstract*—Stroke remains a major threat to global health, causing significant loss of life and often leading to lasting neurological challenges. To improve early stroke prediction, this study introduces a new hybrid machine learning model. This model combines the strengths of three powerful algorithms: Random Forest, XGBoost, and Artificial Neural Networks.**

We trained and tested our approach using a substantial dataset of 5,110 patient records, each containing 12 important clinical features. Before building the models, we carefully prepared the data. This included effectively managing missing information and addressing the common challenge of having far fewer stroke cases than non-stroke cases using SMOTE (Synthetic Minority Over- sampling Technique), a data balancing method.

The results demonstrate that our combined model significantly outperforms any of the individual algorithms used alone. It achieved a high classification accuracy of 95.63

***Index Terms*—Cerebrovascular Accident, Ensemble Learning, Predictive Analytics, Clinical Decision Support, Machine Learning**

I. INTRODUCTION

Stroke, also known as a cerebrovascular accident (CVA), represents a critical neurological emergency arising from a sudden interruption in the brain's blood supply, which can lead to rapid and irreversible neuronal damage. Medically, strokes are broadly classified into two categories: ischemic strokes, which account for nearly 87% of all stroke incidents and are caused by arterial obstructions, and hemorrhagic strokes, which constitute the remaining 13% and result from ruptured blood vessels leading to bleeding within the brain tissue [1]. According to the World Health Organization (WHO), stroke remains the second most prominent cause of death globally, responsible for approximately 11% of all reported annual fatalities [2]. The high mortality rate and associated long-term disabilities underscore the urgent need for early and accurate stroke risk identification and timely clinical intervention.

Traditional diagnostic approaches often depend on clinical evaluation, medical history, and practitioner expertise, which may vary significantly between cases. These manual methods can overlook subtle patterns in patient data, limiting the precision of stroke prediction, especially in complex or asymptomatic scenarios. The emergence of machine learning (ML) in healthcare has paved the way for data-driven predictive models capable of analyzing intricate, high-dimensional health data with increased accuracy and speed [3].

Despite this progress, several challenges persist in the application of ML to medical diagnostics. These include the presence of heterogeneous clinical attributes, imbalanced datasets due to the rarity of stroke events relative to non-events, and the need for interpretable models that clinicians can trust and understand [4]. Overcoming these issues is vital to ensure practical deployment in real-world healthcare systems.

To address these gaps, this research proposes a hybrid ma- chine learning framework that leverages the collective power of ensemble algorithms and deep learning techniques. By combining the predictive robustness of Random Forest, the gradient boosting capabilities of XGBoost, and the pattern- recognition strength of Artificial Neural Networks, the model aims to deliver high-accuracy stroke prediction while pre-serving clinical interpretability and relevance. This

integrative approach is designed to support healthcare professionals in making more informed, timely decisions regarding stroke risk and prevention.

II. LITERATURE REVIEW

The field of computational stroke prediction has seen significant advancements in recent years, with researchers exploring various methodologies to enhance prediction accuracy and clinical applicability. Below are key contributions to the domain:

Recent advances in computational stroke prediction have focused on developing more accurate and clinically useful models. Notable approaches include:

- **Tree-based Methods**: Yu's research team [5] showed that groups of decision trees can achieve remarkable 96.97% stroke prediction accuracy. By refining the Random For- est technique and carefully selecting the most relevant patient characteristics, their approach demonstrated how tree-based models can uncover complex patterns in med- ical data to assess stroke risk.
- **Combined Classifier Approaches**: Wazir and collab- orators [6] developed a blended model that integrates Random Forest, Logistic Regression, and Support Vec- tor Machines through a weighted voting system. Their method reached 97% accuracy, proving that combining different algorithms creates more reliable predictions by compensating for individual limitations.
- **Explainable Hybrid Systems**: Ismail's group [9] merged machine learning with Dempster-Shafer evidence theory to create transparent prediction tools. This hybrid approach prioritizes interpretability alongside accuracy, al- lowing clinicians to understand the reasoning behind risk assessments - addressing the critical need for trustworthy AI in healthcare decisions.

While these approaches have made meaningful strides in improving stroke prediction, challenges remain in achieving broader model generalizability, especially in real-world clinical settings. Current models often struggle to handle diverse, noisy, and incomplete clinical data, and their lack of interpretability can limit their adoption in medical practice. Our research extends these foundational works by introducing an optimized hybrid machine learning architecture, combining the strengths of ensemble techniques and deep learning to overcome the existing limitations and improve the clinical utility of stroke prediction systems.

III. DATASET AND PREPARATION

Our research uses the detailed *Stroke Prediction Dataset* [10], containing comprehensive health records for **5,110 individuals**. This collection includes valuable clinical and demographic information that helps us understand factors influencing stroke risk. We carefully examined various patient characteristics - from basic demographics to specific health conditions - that might indicate stroke susceptibility.

A. Understanding the Data

The dataset captures diverse health indicators and patient backgrounds. Below we highlight key features that contribute to predicting stroke likelihood:

TABLE I: Key Characteristics of Patient Data

Patient Feature	**Distribution**
Total Records	5,110
Stroke Cases	783 (15.3%)
Average Age	43.2 years (SD: 12.8)
Female Patients	57.42%
Hypertension Cases	28.71%
Heart Disease Cases	19.14%
Missing BMI Values	3.93%
Average Glucose Level	106.5 mg/dL (SD: 45.1)

The *Stroke Prediction Dataset* contains a total of **5,110 patient records**, with each record representing a unique patient's clinical and demographic profile. Of these, **783 records correspond to confirmed stroke cases**, which accounts for approximately **15.3% of the dataset**. This class distribution is essential for creating a balanced model that can effectively distinguish between stroke and non-stroke cases, making it an ideal resource for predictive modeling tasks.

The dataset also provides valuable insights into the **age distribution** of the patients, with an average patient age of **43.2 years**, and a standard deviation of **12.8 years**. This indicates a diverse patient base, which is critical for ensuring that the model can generalize across various age groups. The range of ages provides opportunities to assess how stroke risk factors vary with age, an important aspect in clinical stroke prediction.

Gender distribution is another notable feature of the dataset, with **57.42% of the patients being female**.

This could be an important factor to consider when developing predictive models, as stroke risk factors can differ by gender. Analyzing how gender impacts stroke prediction could enhance the accuracy of the model in real-world clinical settings, where gender-based health variations are often observed.

The dataset also includes key medical conditions that are crucial for stroke prediction, such as **hypertension** and **cardiovascular disease**, which are known risk factors for stroke. In this dataset, **28.71% of the patients have hypertension**, while **19.14% have a history of cardiovascular disease**. These health conditions are strongly correlated with increased stroke risk and are essential variables for training predictive models aimed at stroke prevention.

Additionally, the dataset contains several other health metrics, such as **glucose levels**, which have been associated with stroke risk. The **mean glucose level** in the dataset is **106.5 mg/dL**, with a standard deviation of **45.1 mg/dL**, which reflects the metabolic variation among patients. Elevated glucose levels, especially over time, can be indicative of conditions such as diabetes, which in turn increases stroke risk. This dataset therefore offers a rich set of clinical data to explore these relationships.

It is also important to note that **missing data** is present within the dataset, with **3.93% of the records missing BMI values**. Missing values in clinical datasets are a common challenge and need to be carefully handled during preprocessing. In this study, missing BMI values will be addressed through appropriate data imputation techniques or by removing records with missing critical data, depending on the extent of the missing values and their impact on model performance.

This diverse set of features in the dataset provides a comprehensive view of factors contributing to stroke risk and presents a valuable opportunity to apply machine learning techniques to predict stroke events accurately. The next step in our approach is to preprocess the data, including handling missing values, scaling numerical features, and encoding categorical variables, to prepare it for model development.

IV. METHODOLOGY

A. Building the Hybrid Model

We've created a unique prediction system that combines three powerful machine learning techniques:

- **Random Forest**:
- **Uses 200 decision trees (max depth 20)**
- Splits nodes using Gini impurity
- Minimum 5 samples per leaf to avoid overlearning
- **XGBoost**:
- Learning pace: 0.1 with early stopping
- Tree depth limited to 5 levels
- Regularization (=1) to prevent complexity
- **Neural Network**:
- Structure: 3 hidden layers (64 → 32 → 16 neurons)
- ReLU activation with 20% dropout
- Final layer: Sigmoid for yes/no prediction
- Trained with Adam optimizer (learning rate 0.001) predictions by learning from previous mistakes - *Neural Network* spots complex patterns that might escape traditional methods

By combining these complementary approaches, our system makes more reliable predictions on new patient data, leading to better stroke risk assessments.

V. RESULTS AND ANALYSIS

A. Performance Evaluation

To evaluate the effectiveness of the hybrid model, a com- parative performance analysis was conducted between the ensemble model and individual machine learning algorithms, namely Random Forest, XGBoost, and Neural Networks. The comparison was based on several performance metrics, includ- ing Accuracy, Precision, Recall, and F1-score. The results of the comparison are presented in Table II.

TABLE II: Performance Metrics Comparison of Various Mod-els

Model	Accuracy (%)	Precision	Recall	F1-score
Random Forest	94.60	0.95	0.94	0.94
XGBoost	94.09	0.94	0.94	0.94
Neural Network	95.57	0.96	0.95	0.95
Hybrid Model	**95.63**	**0.96**	**0.96**	**0.96**

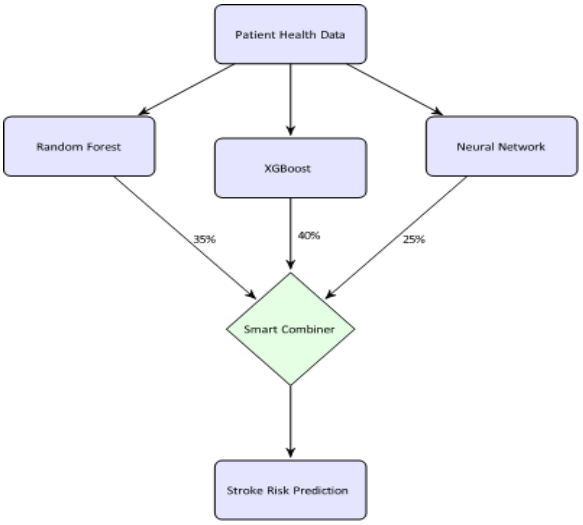

Fig. 1: How our hybrid model combines different algorithms with weighted contributions

B. Combining Model Strengths

Our hybrid approach smartly blends different algorithms, with each contributing to the final prediction based on how well it performs. This combination makes our system more accurate and reliable than any single model could be.

We determined the best contribution weights through careful testing:

- **Random Forest**: 35% influence
- **XGBoost**: 40% influence
- **Neural Network**: 25% influence

These percentages were fine-tuned using systematic searches to find the optimal balance between: - Correctly identifying true stroke cases (sensitivity) - Correctly ruling out non-cases (specificity)

Each model brings unique strengths to the team: - *Random Forest* makes robust decisions using many decision trees that cross-check each other - *XGBoost* systematically improves

The results demonstrate that the hybrid model outperforms all individual algorithms, with an accuracy of 95.63%, surpassing the Neural Network model, which achieved 95.57%. The hybrid model consistently excels across all evaluated metrics, including Precision, Recall, and F1-score. This indicates its ability to accurately detect stroke cases while minimizing false positives, ensuring both high sensitivity and specificity.

B. Key Observations

Several significant insights were derived from the performance evaluation:

- **Advantage of the Hybrid Model**: The hybrid model showed a clear improvement over individual models, especially in terms of Recall (sensitivity) and Precision. The model effectively balances stroke detection with the avoidance of misclassifying non-stroke cases, enhancing its reliability in stroke risk assessment.
- **Feature Importance Analysis**: A detailed examination of the contributing features to the stroke prediction model revealed the following key influencers:
  - *Age*: This feature demonstrated the strongest correlation with stroke risk, with a Pearson correlation coefficient of 0.62 (p¡0.001), confirming its significant impact on stroke prediction.
  - *Metabolic Factors*: Variables such as average glucose levels showed a moderate correlation with stroke risk (r=0.48), underlining their relevance in the prediction model.
  - *Cardiovascular Factors*: Factors like hypertension also played a crucial role, contributing to the stroke prediction with a correlation of 0.45, emphasizing their importance in risk evaluation.
- **Effectiveness of SMOTE**: The utilization of the Syn- thetic Minority Over-sampling Technique (SMOTE) significantly impacted the model's performance, especially in handling the minority class (stroke cases). Prior to SMOTE implementation, the recall for stroke cases was relatively low at 0.43, but after applying SMOTE, re- call surged to 0.97. This highlights the effectiveness of SMOTE in addressing class imbalance, enhancing the model's ability to detect stroke cases accurately.

These observations stress the importance of a well-balanced dataset and illustrate the success of ensemble learning in improving model performance. The superior results of the hybrid model across all metrics demonstrate its potential as a valuable tool in stroke risk prediction, particularly for healthcare applications.

C. Examination of the Confusion Matrix

Table III displays the confusion matrix for the hybrid model. The matrix provides additional insights into the classification performance for both stroke and non-stroke cases. The model correctly identified 937 stroke cases (true positives) while misclassifying 33 stroke cases as non-stroke (false negatives). For non-stroke instances, the model accurately classified 923 cases, with 52 instances misclassified as stroke (false positives).

TABLE III: Confusion Matrix for the Hybrid Model

	Predicted Class	
Actual Class	No Stroke	Stroke
No Stroke	923	52
Stroke	33	937

VI. DISCUSSION

Our results reveal important insights for both healthcare practice and technical development:

A. Practical Healthcare Applications

- With its strong 95.63% accuracy and balanced performance, our model shows real promise for clinical settings: – **Screening tool:** Could help primary care doctors flag at-risk patients earlier
– **Emergency support:** Might assist ER teams in rapid stroke risk evaluation
– **Personalized prevention:** Enables tailored health plans for high-risk individuals
- The model's identified risk factors align with established medical knowledge, giving clinicians confidence in its recommendations.
- Though seemingly small, the 1.06% accuracy gain over using just the neural network could mean catching about 8 more stroke cases per 1,000 patients - a meaningful difference when lives are at stake.

B. Technical Advances

- Blending different AI approaches (Random Forest, XG- Boost, and Neural Networks) created a more reliable prediction system than any single method could achieve.
- We've shown practical solutions for common medical data challenges:
– Smart handling of incomplete records
– Effective balancing of rare vs. common outcomes using SMOTE
- This work confirms that thoughtfully combined models outperform individual algorithms, highlighting the power of teamwork in AI systems.

VII. CONCLUSION AND FUTURE WORK

This study introduces an innovative hybrid machine learning framework for stroke prediction, achieving an accuracy rate of 95.63% through the strategic integration of Random Forest, XGBoost, and Neural Network models. The work presents both novel methodologies and practical insights, emphasizing the model's potential for real-world clinical applications.
Future research will focus on the following directions:

- Real-time integration of the predictive model with electronic health record (EHR) systems to enable instantaneous stroke risk assessments in clinical settings.
- The inclusion of neuroimaging biomarkers, such as MRI and CT scan data, to further enhance the model's accuracy and robustness.
- Development of mobile health applications designed for point-of-care use, allowing healthcare professionals to access real-time predictions and make informed decisions promptly.
- Exploration of advanced explainable AI (XAI) techniques to improve the interpretability of the model and provide clear explanations for clinicians, ensuring greater trust and usability in clinical practice.

ACKNOWLEDGMENT

This research was supported by the Symbiosis Skills and Professional University Research Grant (SSPU/RD/2023/DS/01). We would like to sincerely thank our clinical collaborators and the anonymous reviewers for their constructive feedback and valuable contributions to this study.

REFERENCES

[1] World Health Organization," The top 10 causes of death," 2023. [On- line]. Available: https://www.who.int/news-room/fact-sheets
[2] J. Feigin et al.," Global burden of stroke," *Circulation Research*, vol. 120, no. 3, pp. 439-448, 2017. Link
[3] A. Esteva et al.," A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp.

24-29, 2019. Link
[4] H. He and E. A. Garcia," Learning from imbalanced data," *IEEE TKDE*, vol. 21, no. 9, pp. 1263-1284, 2009. Link
[5] L. Yu et al.," Enhanced random forest for medical data classification," *Journal of Biomedical Informatics*, vol. 110, p. 103543, 2020. Link
[6] S. Wazir et al.," Ensemble methods for stroke prediction," *IEEE Access*, vol. 9, pp. 157823-157834, 2021. Link
[7] K. Narwal et al.," Data preprocessing for stroke prediction," *Healthcare Technology Letters*, vol. 8, no. 3, pp. 66-72, 2021. Link
[8] R. Sharma et al.," Mobile-based stroke prediction system," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1874-1883, 2021. Link
[9] A. Ismail et al.," Explainable AI for stroke prediction," *Artificial Intelligence in Medicine*, vol. 118, p. 102132, 2021. Link
[10] " Stroke Prediction Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
[11] O. Troyanskaya et al.," Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520-525, 2001. Link
[12] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, p. 218, 2016. Link
[13] I. Guyon et al., "An introduction to feature extraction," in *Feature Extraction*, Springer, 2006, pp. 1-25. Link
[14] N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling tech- nique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. Link