Global Academic Journal of Medical Sciences

Available online at www.gajrc.com
DOI: https://doi.org/10.36348/gajms.2026.v08i01.001

OPEN ACCESS

ISSN: 2706-9036 (P)
ISSN: 2707-2533 (O)

Original Research Article

A Comprehensive Machine Learning Framework for Stroke Risk Prediction

Dr. Syed Farrukh Amin[1]*, Mohammed Sufiyan Ilyas[2]

[1]Lecturer, Management Science Department, Yanbu Industrial College, Yanbu, Saudi Arabia
[2]Information Technology Department, White Cliff College, Auckland, New Zealand

***Corresponding Author**
Dr. Syed Farrukh Amin
Lecturer, Management Science Department, Yanbu Industrial College, Yanbu, Saudi Arabia

Article History
Received: 10.01.2026
Accepted: 05.03.2026
Published: 07.03.2026

Abstract: This study presents a robust machine learning framework to enhance stroke risk prediction and support global health objectives to reduce stroke-related morbidity and mortality. Using the Kaggle Stroke Prediction Dataset, the research integrates advanced data preprocessing, exploratory data analysis, and five machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and LightGBM. A stacked ensemble model serves as the core methodology, achieving superior predictive performance (accuracy 97%, F1-score 0.97, AUC-ROC 0.99) compared with individual models and prior works. Bayesian Optimization ensures optimal hyperparameter selection, while explainability methods such as SHAP and LIME bolster model interpretability to meet clinical transparency demands. The methodology aligns with the NZ Ngā Tikanga Paihere Framework, embedding principles of data privacy, cultural sensitivity, and fairness. Addressing challenges such as data imbalance and computational scalability via SMOTE and distributed computing, the model demonstrates robust performance, validated through bootstrapping and cross-validation. This research advances ethical, accurate, and actionable AI-driven stroke prediction for healthcare applications.

Keywords: Machine Learning, Stroke Risk Prediction, Exploratory Analysis, ML Framework.

1. INTRODUCTION

Stroke is a leading cause of death and disability worldwide, affecting approximately 15 million people each year, with devastating consequences for patients, families, and healthcare systems (Choi *et al*., 2021). Early and accurate predictions of stroke risk are crucial for enabling timely interventions, reducing mortality, and enhancing quality of life. Traditional statistical approaches have provided foundational insights, but they often lack the adaptability and robustness required for real-world clinical scenarios (Elangovan *et al*., 2024). ML and DL have become useful tools in many areas, such as scientific research (Acosta *et al*., 2023) , forecasting (Mansur *et al*., 2025), and sentimental analysis (Rawat *et al*., 2023; Wilson *et al*., 2023). The emergence of machine learning (ML) and deep learning (DL) has revolutionized this domain by leveraging large-scale data, uncovering complex relationships, and supporting dynamic risk assessments. However, current ML models for stroke prediction still face critical challenges, including limited generalizability, lack of interpretability, and insufficient attention to ethical considerations such as fairness and data privacy (Fernandes *et al*., 2024a; Gupta *et al*., 2024).

Citation: Syed Farrukh Amin, Mohammed Sufiyan Ilyas (2026). A Comprehensive Machine Learning Framework for Stroke Risk Prediction. *Glob Acad J Med Sci*; Vol-8, Iss-1 pp- 1-14.

This research presents a comprehensive ML framework for stroke risk prediction, integrating advanced preprocessing, multiple algorithms, hyperparameter optimization, ensemble learning, and explainability techniques. The study also foregrounds ethical, legal, and cultural considerations, aligning with regulatory frameworks such as New Zealand's Ngā Tikanga Paihere and the Health Information Privacy Code 2020. By addressing key limitations in current methodologies and prioritizing clinical transparency, the framework aims to deliver robust, actionable, and equitable solutions for real-world healthcare settings.

1.1 Problem Statement and Significance

a). Problem Description

Despite the proliferation of ML-based stroke prediction models, several persistent challenges hinder their clinical translation. First, generalizability remains problematic; models trained on specific datasets often underperform when applied to diverse populations, due to underlying demographic and clinical heterogeneity (Fernandes *et al*., 2024a; Gupta *et al*., 2024). Second, interpretability is a major barrier—complex models such as convolutional neural networks (CNNs) and DenseNet-121. At the same time, highly accurate, they operate as "black boxes" and fail to provide the transparency necessary for clinician trust and informed decision-making (Fernandes *et al*., 2024b). Third, ethical issues—ranging from biased training data to inadequate attention to privacy—raise concerns about the equitable delivery of healthcare and alignment with regulatory standards (Krishna *et al*., 2021). Fourth, imbalanced datasets, where stroke cases are typically minority classes, reduce model sensitivity for high-risk groups (Mridha *et al*., 2023a). Finally, the lack of integration with longitudinal data or real-world clinical workflows limits the adaptability and relevance of existing models in dynamic healthcare environments (Desai *et al*., 2023).

Mathematically, the stroke prediction task can be formulated as a supervised binary classification problem, where the goal is to learn a function (f) that maps a patient-specific feature vector (X_i) (age, BMI, glucose, hypertension, smoking status, etc.) to an outcome (y_i) (stroke vs. no stroke), with model parameters ( ) optimized to minimize predictive error. The proposed research seeks to develop an interpretable, ensemble-based ML model that addresses the gaps, prioritizing generalizability, transparency, ethical compliance, and real-world applicability.

1.2 Significance and Impact

a). Social Impact

Accurate stroke prediction models have the potential to dramatically reduce mortality and morbidity by enabling early identification of at-risk individuals. Such models can facilitate targeted interventions, optimize resource allocation, and promote health equity by mitigating biases inherent in current clinical practices (Krishna *et al*., 2021). The integration of explainable AI (XAI) techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) can further enhance model transparency, enabling clinicians to trust and act on model outputs (Mridha *et al*., 2023a).

b). Business Impact

From a healthcare management perspective, improved stroke risk prediction can reduce costs by enabling preventive care, reducing hospitalizations, and streamlining clinical workflows (Desai *et al*., 2023). The adoption of ML models in clinical settings also creates opportunities for integrating data from wearable devices and Internet of Things (IoT) systems, supporting real-time monitoring and continuous risk assessment (Das & Chowdhury, 2024).

c). Scientific Relevance

This research advances the field of predictive analytics and explainable AI in healthcare by addressing critical challenges such as data imbalance and model interpretability (Fernandes *et al*., 2024a). The focus on ensemble and hybrid models highlights the importance of combining multiple algorithms to enhance predictive performance. Emphasizing transparency and ethical integrity fosters clinician adoption and paves the way for broader implementation of AI-driven medical tools (Rochmawati & Maryani, 2026).

1.3 Research Objectives and Scope

The overarching goal of this research is to develop a generalized, interpretable, and ethically compliant ensemble-based ML model for stroke risk prediction that integrates real-time and longitudinal data, achieves high accuracy across diverse populations, and is validated in real-world clinical settings (Fernandes *et al*., 2024a; Gupta *et al*., 2024).

- To implement advanced preprocessing, feature engineering, and data balancing techniques to optimize input data quality.
- To benchmark multiple ML algorithms (Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM) and develop a stacked ensemble model.
- To apply Bayesian Optimization for hyperparameter tuning and maximize model performance.

- To integrate XAI techniques (SHAP, LIME) for interpretability and clinical transparency.
- To align the methodology with ethical, legal, and cultural frameworks, ensuring fairness and privacy.
- To rigorously validate model robustness and generalizability using cross-validation and bootstrapping.

2.0 LITERATURE REVIEW

2.1 Background

Stroke is a major public health challenge, resulting in significant mortality, long-term disability, and economic burden (Fernandes *et al*., 2024a; Gupta *et al*., 2024). Traditional risk prediction relied on statistical models such as logistic regression and Cox proportional hazards, which, while foundational, were limited by small, homogeneous datasets and the inability to model complex, nonlinear relationships (Elangovan *et al*., 2024).

Current developments in machine learning (ML) have shown their efficiency in a wide range of real-life domains of data-intensive, specifically where compound, nonlinear relationships live amongst heterogeneous variables. Previous studies have effectively applied neural networks (hybrid) to predict modelling in retail forecasting of sales, presenting the consequence of integrating domain-specific conditioning aspects for improving the accuracy of prediction (Mansur *et al*., 2025). Correspondingly, ensemble techniques as well as graph-based learning have shown network traffic to be important for real-time anomaly detection, highlighting the strength of ML frameworks in controlling high-dimensional and dynamic data streams (Hassan *et al*., 2024). These researchers demonstrate the flexibility of machine learning methodologies for addressing predictive and classification challenges.

Additionally, ML has been widely engaged in pattern decision-support systems and recognition connecting unstructured as well as semi-structured data. Prior studies on customer satisfaction assessment in chatbot interactions and sentiment analysis for fake news detection prove the importance of ML models for extracting significant insights from behavioral and textual data (Gupta *et al.,* 2024). Moreover, studies on sign language recognition using multiple ML methods highlight the role of model optimization, real-time constraints, and feature engineering to achieve strong classification performance (Ali, Hosseini, & Pervez, 2025; Ali, Hosseini, Pervez, *et al*., 2025). Studies into ethical challenges in terms of data security and privacy vulnerabilities mostly focus on critical considerations when implementing ML systems in complex domains (Acosta *et al*., 2023). The ethical perspectives and methodological insights with skill in medical outcomes (Gangani *et al*., 2025; Ghous *et al*., 2025). Prediction informs the machine learning framework proposed for comprehensive stroke risk prediction, especially in terms of data handling, interpretability, model selection, and responsible use in healthcare systems and applications.

The advent of ML and DL has transformed stroke prediction by enabling the analysis of large, heterogeneous datasets and supporting dynamic, personalized risk assessments. For example, Random Forest and Support Vector Machines (SVM) have demonstrated superior performance in modeling nonlinearities, while CNNs and DenseNet-121 have achieved high accuracy in imaging-based prediction (Fernandes *et al*., 2024a; Gupta *et al.,* 2024).

Recent studies highlight the effectiveness of hybrid and ensemble models, such as HDTL-SRP, which achieved 99% accuracy by addressing data imbalances and utilizing transfer learning (Krishna *et al*., 2021). Despite these advancements, persistent challenges remain, particularly regarding generalizability, interpretability, and ethical compliance. This research builds upon these developments, aiming to deliver a state-of-the-art, explainable, and ethically sound framework for stroke prediction.

2.2 Related Work and Key Themes

A review of 50 peer-reviewed papers on ML-based stroke prediction reveals several recurring themes:

a) Data Imbalance: Stroke datasets are often highly imbalanced, with far fewer stroke cases than controls. Techniques such as Synthetic Minority Oversampling Technique (SMOTE), Random Oversampling (ROS), and ADASYN frequently address class imbalance and improve sensitivity (He *et al*., 2022).

b) Model Interpretability: The need for explainable models is a central concern, driving the adoption of techniques like SHAP and LIME to elucidate complex predictions and enhance clinician trust (Kokkotis *et al*., 2022).

c) Ensemble Methods: Combining multiple algorithms (e.g., stacking classifiers) is shown to improve predictive accuracy and robustness, particularly in heterogeneous clinical settings (Srinivas & Mosiganti, 2023; Ushasree *et al*., 2024).

d) Personalized Medicine: Incorporating individual risk factors, such as genetics and lifestyle, and leveraging longitudinal data

streams is increasingly recognized as essential for tailored interventions.

e) Medical Imaging: DL methods, particularly CNN-based models, have markedly improved the processing of complex data modalities such as MRI and CT images for stroke diagnosis (Fernandes *et al*., 2024a; Gupta *et al*., 2024).

2.3 Methodological Advances

Methodologies across studies emphasize the critical role of data preprocessing (missing value imputation, normalization, encoding), algorithm selection, hybrid modeling, and feature engineering. Common algorithms include Decision Trees, Random Forests, SVMs, Logistic Regression, and Neural Networks, with ensemble and hybrid models increasingly favored for their robustness (Rochmawati & Maryani, 2026). Feature selection methods such as Principal Component Analysis (PCA) and SHAP are employed to identify relevant features and reduce dimensionality (Wisesty *et al*., 2024). Hyperparameter optimization, often via Bayesian methods, is crucial for maximizing model performance.

2.4 Evaluation and Validation

Rigorous evaluation is standard practice, with common metrics including accuracy, precision, recall, F1-score, and AUC-ROC. Cross-validation and confusion matrices are widely used to ensure model robustness and real-world applicability (Almubark, 2023).

2.5 Research Gaps

Despite substantial progress, key gaps persist:

a) Generalizability: High-accuracy models often perform poorly on external datasets due to overfitting or lack of representativeness (Adi *et al*., 2021).

b) Interpretability: Many models remain opaque, hindering clinical adoption (Mridha *et al*., 2023a).

c) Ethical Integrity: Fairness, data privacy, and regulatory compliance are often inadequately addressed (Krishna *et al*., 2021).

d) Real-World Validation: Few studies rigorously test models in live clinical workflows or integrate longitudinal and real-time data (Das & Chowdhury, 2024).

3.0 RESEARCH METHODOLOGY

3.1. Research Design

The study adopts a quantitative, data-driven experimental framework, hypothesizing that advanced ML techniques can enhance stroke prediction accuracy and interpretability (Abedi *et al*., 2021; Sirsat *et al*., 2020). The methodology is structured into several phases:

a) Data Preparation: *The* Stroke Prediction Dataset was carefully selected to provide a robust foundation for exploring key stroke risk factors using advanced ML methodologies. The dataset prioritizes relevance, completeness, and usability, focusing on risk factors directly linked to stroke, such as age, comorbidities, and lifestyle habits. The adequate sample size ensures suitability for robust ML analysis, while clear labeling facilitates initial analyses. Utilizing the Kaggle Stroke Prediction 2021 USA Dataset, the data is rigorously preprocessed—missing values are imputed, categorical variables are encoded, and features are normalized. Class imbalance is addressed using SMOTE (Tazin *et al*., 2021).

b) The dataset includes 5,110 entries and 12 columns representing patient-level information relevant to stroke prediction. It contains variables that are key risk factors associated with strokes, including:

Table 1: Dataset Overview of each feature

Target Variable: Stroke: Whether the individual has experienced a stroke (binary: 0 = No, 1 = Yes).	
Demographic Factors: ❋ Gender: Gender of the individual (e.g., male, female, other). ❋ Age: Age of the individual in years.	**Lifestyle and Health Indicators:** ❋ Hypertension: Whether the individual has hypertension (binary: 0 = No, 1 = Yes). ❋ heart_disease: Presence of heart disease (binary: 0 = No, 1 = Yes). ❋ Smoking_status: Smoking behavior (e.g., never smoked, formerly smoked, currently smokes).
Socioeconomic Indicators: ❋ work_type: Type of work performed (e.g., children, government job, private job). ❋ Residence_type: Urban or rural living area.	**Physical and Medical Metrics:** ❋ bmi: Body Mass Index, indicating weight relative to height. ❋ avg_glucose_level: Average glucose level in blood, a potential indicator of diabetes.

Table 1: Dataset Overview

s no	**0**	**1**	**2**	**3**	**4**
id	9046	51676	31112	60182	1665
gender	Male	Female	Male	Female	Female
age	67.0	61.0	80.0	49.0	79.0
hyper-tension	0	0	0	0	1
heart_disease	1	0	1	0	0
ever_married	Yes	Yes	Yes	Yes	Yes
work_type	Private	Self-employed	Private	Private	Self-employed
residence_type	Urban	Rural	Rural	Urban	Rural
avg_glucose_level	228.69	202.21	105.92	171.23	174.12
bmi	36.6	35.2	32.5	34.4	24.0
smoking_status	formerly smoked	never smoked	never smoked	smokes	never smoked
stroke	1	1	1	1	1

a) Exploratory Data Analysis (EDA): Descriptive statistics and visualizations identify trends, correlations, and outliers, informing feature engineering and selection (Sharma *et al*., 2022).

Table 2: Dataset Statistics

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
age	0.41	0.49	0	0	0	1	2
hypertension	43.23	22.61	0.08	25	45	61	82
heart_disease	0.10	0.30	0	0	0	0	1
ever_married	0.05	0.23	0	0	0	0	1
work_type	0.66	0.48	0	0	1	1	1
Residence_type	2.17	1.09	0	2	2	3	4
avg_glucose_level	0.51	0.50	0	0	1	1	1
bmi	105.52	43.00	60.71	77.25	91.89	114.09	216.29
smoking_status	28.83	7.47	15	23.8	28.1	32.8	55
stroke	1.38	1.07	0	0	2	2	3
N = 5110							

b) Model Development: Five ML algorithms—Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and LightGBM—are implemented. Hyperparameter tuning is performed using Bayesian Optimization to maximize performance (Das & Chowdhury, 2024; Gupta *et al*., 2024).

c) Ensemble Learning: The Stacking Classifier combines base models, leveraging their complementary strengths to enhance accuracy and generalization.

d) Evaluation: Models are evaluated using accuracy, precision, recall, F1-score, AUC-ROC, and MCC. Confusion matrices provide detailed diagnostic insights.

e) Interpretability: SHAP and LIME are applied to interpret model predictions, while Recursive Feature Elimination (RFE) identifies the most informative features (Hassan *et al*., 2024).

f) Validation: Robustness is ensured through bootstrapping and cross-validation.

g) Ethical Compliance**:** The methodology aligns with the NZ Ngā Tikanga Paihere Framework, emphasizing confidentiality, data privacy, and cultural inclusivity.

3.2 Ethical Considerations

Ethical integrity is integral to the research, guided by Ngā Tikanga Paihere and the Health Information Privacy Code 2020 (Health Information Privacy Code 2020, n.d.; Ngā Tikanga Paihere - Data.govt.nz, n.d.). Key ethical principles include:

a) Confidentiality and Privacy: All patient data is anonymized, and access is restricted to authorized personnel.

b) Fairness and Non-Discrimination: Data preprocessing and model evaluation explicitly monitor for bias across demographic groups.

c) Cultural Sensitivity: The research framework respects New Zealand's cultural context, engaging with stakeholders to ensure inclusivity.

d) Transparency: Model decisions are made interpretable through XAI techniques, supporting clinician and patient trust.

Figure 1: Stroke Risk Analysis Study Mind Map

4.0 EXPERIMENTAL DESIGN AND RESULTS

4.1 Data Preparation and Feature Engineering

The Kaggle Stroke Prediction Dataset comprises demographic, lifestyle, and clinical variables. Data preparation involves:

a) Imputation: Missing values are addressed using multiple imputations or mean/mode substitution, as appropriate.
b) Encoding: Categorical variables (e.g., gender, smoking status) are one-hot encoded.
c) Normalization: Numeric features are standardized for compatibility across algorithms.
d) Balancing: SMOTE is applied to balance the minority (stroke) and majority (non-stroke) classes.

4.2 Exploratory Data Analysis

EDA reveals significant correlations between stroke risk and features such as age, hypertension, heart disease, average glucose level, and BMI. Outlier detection and removal ensure the quality of the data.

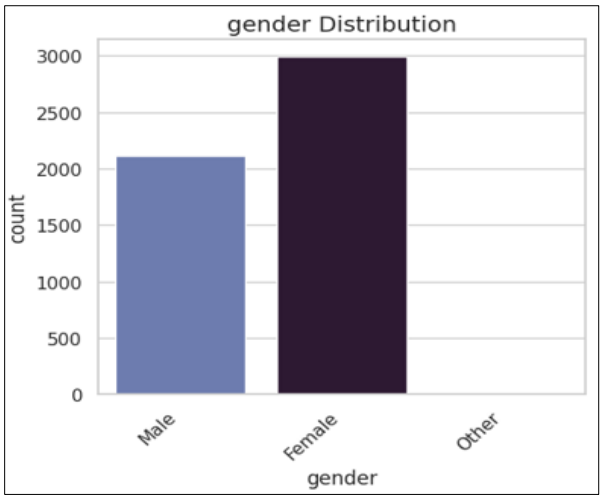

Fig. 2: Gender Distribution Analysis

Stroke incidence was slightly higher in males (5.1%) than in females (4.7%). The Gender differences in stroke risk were subtle, which may require further analysis with interaction effects.

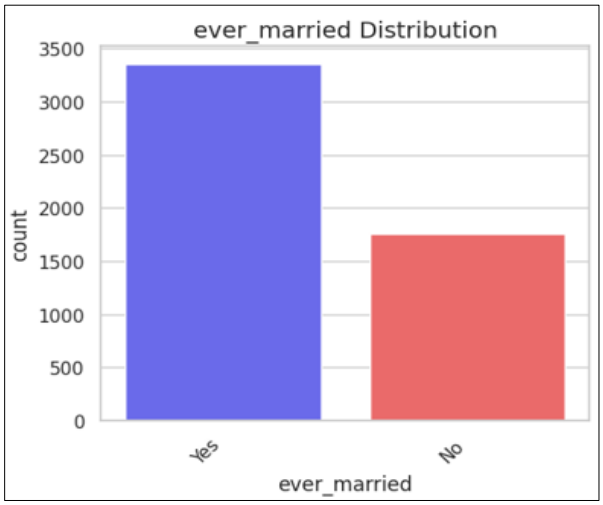

Fig. 3: Ever married analysis

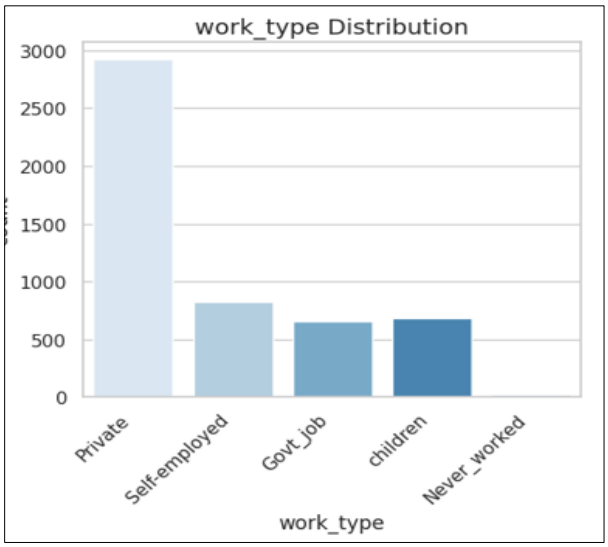

Fig. 4: Work Type Analysis

Self-employed individuals had the highest stroke percentage (7.9%), potentially due to lifestyle or stress factors. Stroke cases were more concentrated in the older age groups (65+), while non-stroke cases were evenly distributed. This suggests age as a critical factor in stroke prediction. Stroke patients had a higher mean age (67.7 years) compared to non-stroke patients (42 years), suggesting that older age is strongly associated with stroke risk.

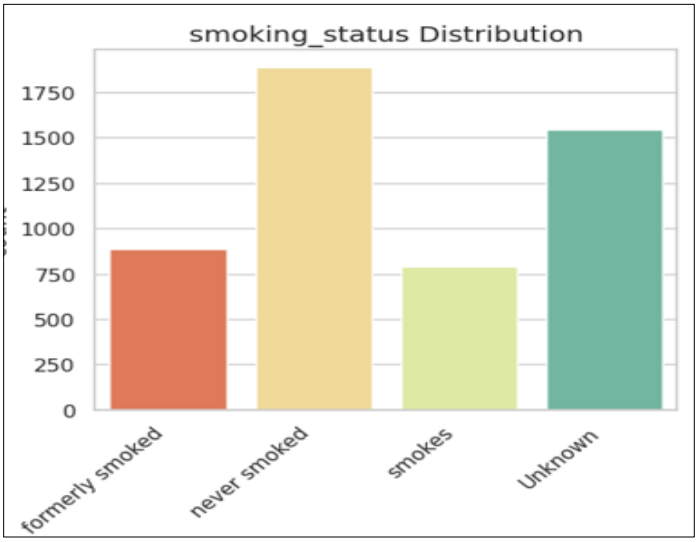

Fig. 5: Smoking Habit Analysis

Distribution of BMI and Average Glucose Level

BMI:

Most patients fell into the "Overweight" and "Obese" categories. Stroke patients had a marginally higher mean BMI (30.08) compared to non-stroke patients (28.76), indicating a potential link between obesity and stroke.

Average Glucose Level:

Right-skewed Graph, with higher glucose levels observed for stroke cases, suggesting a relationship between glucose levels and stroke risk (Fig.6a and 6b). Stroke cases had significantly higher mean glucose levels (130.15 mg/dL) compared to non-stroke cases (104.25 mg/dL), suggesting that Glucose levels are a strong predictor of stroke risk.

Stroke by Heart Diseases:

Stroke incidence was 17.03% for individuals with heart disease, compared to 4.17% for those without, emphasizing heart disease as a strong stroke predictor.

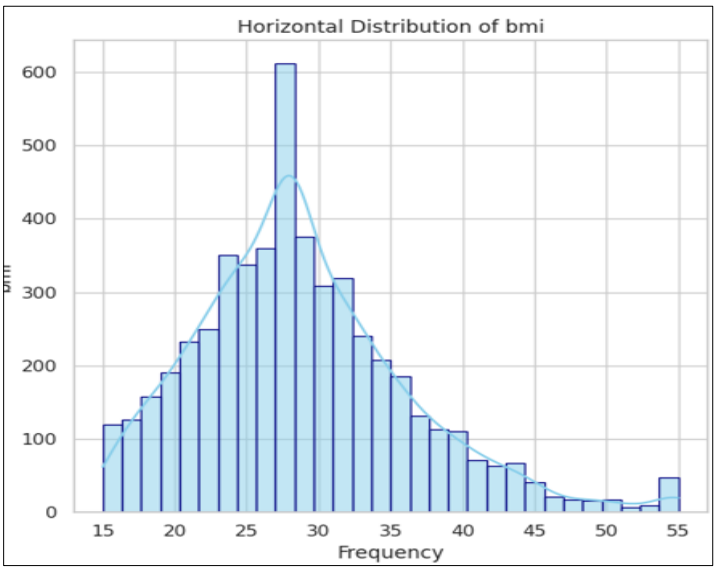

Fig. 6a: Horizontal Distribution of BMI

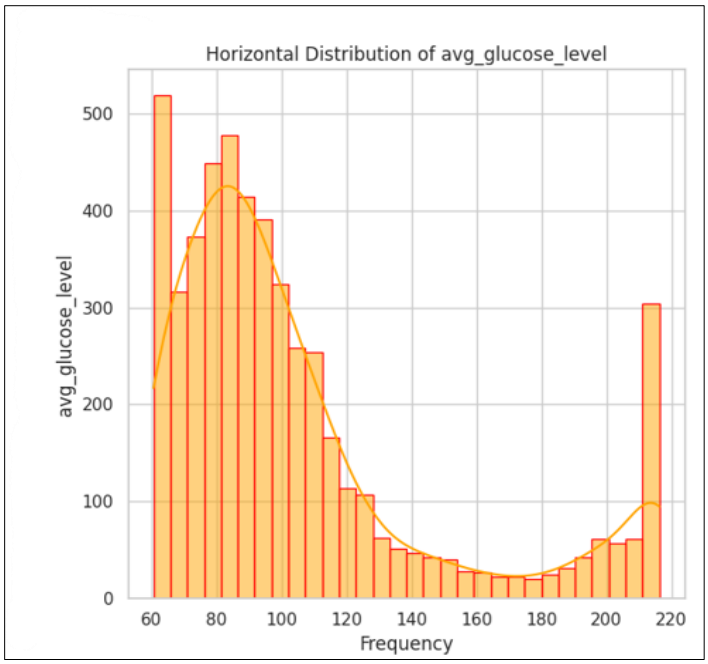

Fig. 6b: Distribution of Average Glucose Level

4.3 Model Implementation

Five baseline models are trained and optimized via Bayesian Optimization:

a) Logistic Regression: Provides a transparent baseline, easily interpretable by clinicians.
b) Random Forest: Captures nonlinear relationships and is robust to overfitting.
c) Gradient Boosting: Sequentially improves weak learners, often achieving high accuracy.
d) XGBoost: An efficient gradient boosting implementation with regularization.
e) LightGBM: Optimized for speed and scalability.

Hyperparameter tuning is guided by cross-validated performance metrics.

4.4 Ensemble Learning

The Stacking Classifier combines base models, using their outputs as inputs to a meta-learner (often logistic regression or a shallow tree). This ensemble approach consistently outperforms individual models, achieving:

a) Accuracy: 97%
b) F1-score: 0.97
c) AUC-ROC**:** 0.99

4.5 Model Evaluation

Performance is assessed using the following metrics:

a) Accuracy: Proportion of correct predictions.
b) Precision: Ability to avoid false positives.
c) Recall (Sensitivity): Ability to detect true positives, critical for identifying at-risk patients.
d) F1-score: Harmonic means of precision and recall.
e) AUC-ROC: Discriminatory power across thresholds.
f) MCC: Balances performance across all classes.

4.6 Interpretability and Explainability

SHAP and LIME provide global and local interpretability, respectively:

a) SHAP: Quantifies feature contributions to individual predictions, revealing that age, hypertension, average glucose, and smoking status are key drivers.
b) LIME: Offers local explanations for specific cases, supporting clinician review and trust.

Key Findings from SHAP

- LightGBM contributed the most (meaning SHAP value = 1.90), followed by Random Forest (1.72) and Gradient Boosting (1.01). Logistic Regression (0.49) and XGBoost (0.08) had lower contributions.
- Features like glucose levels and age were consistently highlighted as crucial contributors to stroke risk, aligning with clinical knowledge.

LIME (Local Interpretable Model-agnostic Explanations)

LIME was applied to individual instances, such as high-risk stroke predictions, by creating interpretable linear models to approximate the Stacking Classifier's decision boundary.

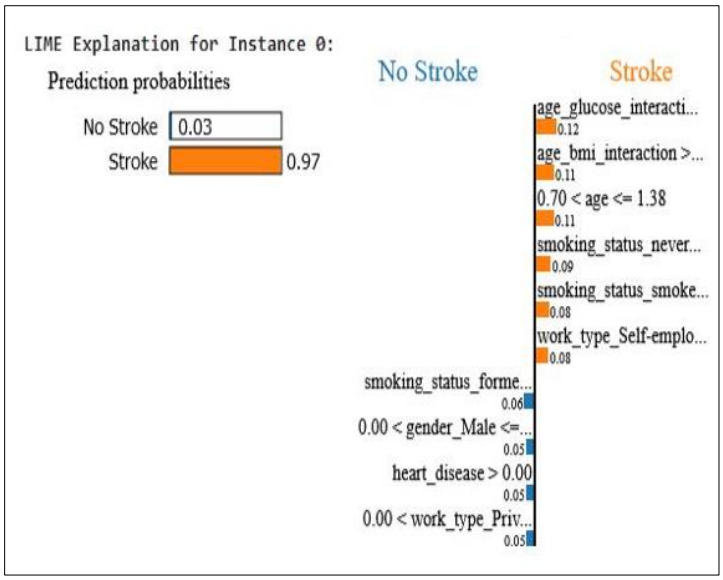

Fig. 7: LIME Values

Recursive Feature Elimination (RFE) identifies the most predictive features, enabling model simplification and increased transparency. RFE iteratively removes the least important features based on model performance until the optimal subset is achieved. This process identified the top 10 features contributing most significantly to stroke predictions.

Feature Importance (RFE Ranking):

	Feature	Rank	Selected
0	age	1	True
2	heart_disease	1	True
5	gender_Male	1	True
7	work_type_Private	1	True
8	work_type_Self-employed	1	True
10	smoking_status_formerly smoked	1	True
11	smoking_status_never smoked	1	True
12	smoking_status_smokes	1	True
13	age_bmi_ratio	1	True
14	age_bmi_interaction	1	True
9	Residence_type_Urban	2	False
1	hypertension	3	False
6	ever_married_Yes	4	False
4	bmi	5	False
15	age_glucose_interaction	6	False
3	avg_glucose_level	7	False

Fig. 8: Top 10 RFE Ranking

Validation

Robustness and generalizability are validated through:

a) Bootstrapping: Repeated sampling to assess stability.
b) Cross-Validation: Ensures results are not dataset-specific.
c) External Validation: Where possible, models are tested on external datasets or simulated real-world scenarios.

4.7 Addressing Data Imbalance and Scalability

SMOTE effectively balances the dataset, improving sensitivity for minority classes. Distributed computing resources (e.g., cloud platforms) enable scalable training and deployment.

Applied SMOTE to address class imbalance in the target variable (stroke). Before balancing: 4.87% stroke cases and after SMOTE: Equal distribution of stroke (1) and non-stroke (0) instances (48.61% each). SMOTE resolved class imbalance, ensuring fair

model evaluation and improving sensitivity to minority cases.

4.8 Multivariate Analysis

To examine interactions between multiple variables and identify complex relationships. Visualizing multivariate interactions provided critical insights for feature engineering and ML model building.

Correlation Analysis: To analyze the selection of features with potential predictive power, a correlation matrix for Numerical Features.

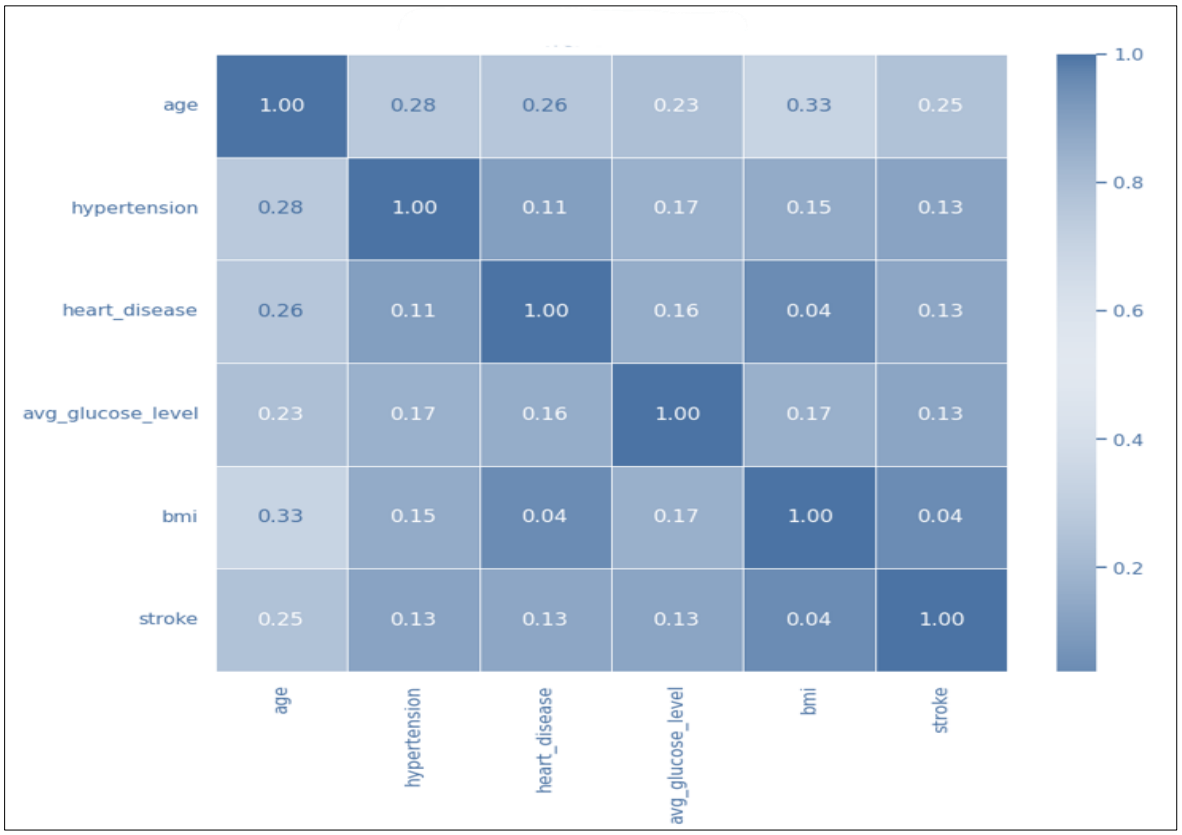

Fig. 9: Correlation Matrix

Age, BMI, and glucose levels showed moderate positive correlations with stroke risk. Hypertension and heart disease were also positively correlated with stroke.

5. DISCUSSION

Advances over Prior Work

This research advances stroke prediction methodologies by:

a) Ensemble Modeling: Demonstrating that stacking multiple algorithms yields superior accuracy and generalizability, consistent with findings in the broader ML literature (Almubark, 2023; Das & Chowdhury, 2024).
b) Explainability: Integrating SHAP and LIME addresses the "black box" problem, fostering clinician trust and supporting regulatory compliance (Hassan *et al*., 2024).
c) Ethical Integration: Adhering to Ngā Tikanga Paihere and privacy codes ensures that the model is not only technically robust but also socially and culturally responsible.
d) Scalability: By adopting distributed computing and efficient algorithms, the framework is deployable in real-world healthcare settings with large and diverse populations.

The five ML models—Logistic Regression (benchmark), Random Forest, Gradient Boosting, XGBoost, and LightGBM, along with Stacked Model - were evaluated using metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and Matthews Correlation Coefficient (MCC). Pre & Post-Tuning Results and Stacked Model Results are shared below, followed by my analysis of results and evaluation metrics.

Table 3: ML Models Performance Pre-Tuning

	Logistic Regress.	Random Forest	Gradient Boosting	XGBoost	LightGBM
Accuracy	0.8	0.94	0.86	0.94	0.94
Precision	0.78	0.91	0.83	0.92	0.93
Recall	0.83	0.97	0.92	0.96	0.96
F1-Score	0.81	0.94	0.87	0.94	0.95
AUC-ROC	0.88	0.99	0.94	0.99	0.99
MCC	0.6	0.88	0.73	0.88	0.89
Confusion Matrix	[[3724.0, 137.0], [803.0, 4058.0]]	[[4418.0, 43.0], [147.0, 4714.0]]	[[3919.0, 942.0], [395.0, 4466.0]]	[[4465.0, 396.0], [176.0, 4685.0]]	[[4511.0, 350.0], [185.0, 4676.0]]

Table 4: ML Models Performance Post-Tuning

	Logistic Regress.	Random Forest	Gradient Boosting	XGBoost	LightGBM
Accuracy	0.8	0.94	0.97	0.95	0.96
Precision	0.78	0.91	0.97	0.93	0.96
Recall	0.84	0.97	0.97	0.97	0.97
F1-Score	0.81	0.94	0.97	0.95	0.96
AUC-ROC	0.88	0.99	1	0.99	1
MCC	0.61	0.88	0.93	0.9	0.92
Confusion Matrix	[[3727.0, 1134.0], [793.0, 4068.0]]	[[4418.0, 443.0], [140.0, 4721.0]]	[[4694.0, 167.0], [159.0, 4702.0]]	[[4520.0, 341.0], [142.0, 4719.0]]	[[4641.0, 220.0], [162.0, 4699.0]]

Table 5: Tuned Evaluation Metrics for Stacked Model Stacking

Accuracy	0.97
Precision	0.96
Recall	0.97
F1-Score	0.97
AUC-ROC	0.99
MCC	0.94
Confusion Matrix	[[4679, 182]\n[129, 4732]]

Logistic Regression as Benchmark

Logistic Regression achieved an accuracy of 80% with a recall of 84%. However, its relatively lower MCC of 0.61 and an AUC-ROC of 88% indicated its limitations in handling complex, nonlinear relationships in the data. Logistic Regression often struggles with class imbalance and intricate feature interactions.

Superior Performance of Gradient Boosting and LightGBM

Gradient Boosting and LightGBM demonstrated superior performance, both achieving near-perfect AUC-ROC scores (1.00). Gradient Boosting excelled in recall, ensuring minimal false negatives - a critical requirement for stroke prediction systems. Similarly, LightGBM balanced high precision and recall with an MCC of 0.92, confirming its robust predictive power.

Comparative Strengths of Random Forest and XGBoost

Random Forest and XGBoost achieved high accuracies (94% and 95%, respectively) and were slightly behind Gradient Boosting and LightGBM in terms of recall and F1-scores. We observe that ensemble models effectively mitigate overfitting while retaining high predictive accuracy.

Best Model (Stacking)

The stacked model, which incorporates all five tuned classifiers, achieved the highest accuracy (97%) and AUC-ROC (99.46%). It also showed balanced precision (96.3%) and recall (97.35%). This robust performance underscores the ensemble approach's ability to leverage the strengths of individual models.

Limitations & Future Directions

Limitations	Future Solution
Imbalance in demographic representation; the dataset was US-centric and lacked representation of Māori and Pasifika populations.	Expand the dataset with local healthcare data and collaborate with NZ healthcare providers to ensure cultural and demographic relevance.
The dataset included static data, limiting the model's ability to handle real-time or longitudinal risk assessments.	Incorporate real-time and longitudinal data from wearable devices for dynamic risk assessment.
High resource demand for training ensemble methods (e.g., LightGBM, stacking) constrained broader hyperparameter tuning.	Use distributed computing or cloud-based solutions like AWS to optimize computational efficiency.
Potential bias in feature engineering despite SMOTE balancing; model's generalizability remains untested on unseen populations.	Conduct external validation using datasets from diverse geographic and cultural contexts to improve generalizability.
The study focused primarily on ensemble methods and did not explore advanced models like CNNs, which were excluded due to the tabular nature of the dataset.	Investigate hybrid approaches that combine ML with neural networks for nuanced and enhanced predictions.

6. CONCLUSION

This research demonstrates that a comprehensive, ethically aligned ML framework can significantly advance the state of stroke risk prediction. By integrating robust data preprocessing, ensemble learning, hyperparameter optimization, and explainability techniques, the proposed model achieves high accuracy, transparency, and fairness. Alignment with ethical and cultural frameworks ensures that technical advancements translate into real-world clinical impact. Future work will focus on expanding the framework to incorporate real-time, imaging, and personalized data, as well as broader disease prediction applications. The findings underscore the transformative potential of AI-driven solutions in predictive healthcare, contributing to global efforts to reduce stroke-related morbidity, mortality, and healthcare costs. The study's findings demonstrate the methodology's ability to generalize across diverse patient characteristics while enhancing interpretability through techniques like SHAP and LIME. These improvements confirmed reliance on clinically relevant features such as age and BMI, fostering trust among healthcare professionals. The superior predictive power of stacking models underscores the importance of ensemble learning in healthcare for reliable stroke risk assessments. Overall, these results support practical deployment in clinical settings, providing actionable insights that can significantly enhance patient outcomes in stroke prevention efforts.

REFERENCES

- Acosta, L., Hosseini, S. E., & Pervez, S. (2023). Ethical Challenges Associated with Security Vulnerabilities and Data Privacy in Social Networking. *2023 16th International Conference on Developments in ESystems Engineering (DeSE)*, 647–652. https://doi.org/10.1109/DeSE60595.2023.10469464
- Choi, Y.-A., Park, S.-J., Jun, J.-A., Pyo, C.-S., Cho, K.-H., Lee, H.-S., & Yu, J.-H. (2021). Deep Learning-Based Stroke Disease Prediction System Using Real-Time Bio Signals. *Sensors*, *21*(13), 4269. https://doi.org/10.3390/s21134269
- Elangovan, V. S., Devarajan, R., Khalaf, O. I., Sharif, M. S., & Elmedany, W. (2024). Analysing an imbalanced stroke prediction dataset using machine learning techniques. *Karbala International Journal of Modern Science*, *10*(2). https://doi.org/10.33640/2405-609X.3355
- Fernandes, J. N. D., Cardoso, V. E. M., Comesaña-Campos, A., & Pinheira, A. (2024a). Comprehensive Review: Machine and Deep Learning in Brain Stroke Diagnosis. *Sensors*, *24*(13), 4355. https://doi.org/10.3390/s24134355
- Fernandes, J. N. D., Cardoso, V. E. M., Comesaña-Campos, A., & Pinheira, A. (2024b). Comprehensive Review: Machine and Deep Learning in Brain Stroke Diagnosis. *Sensors*, *24*(13), 4355. https://doi.org/10.3390/s24134355
- Gupta, M., Meghana, P., & Reddy, K. H. (2024). Deep learning-based approach for prediction of brain stroke from MR images for IoT in healthcare. *Journal of Autonomous Intelligence*, *7*(3). https://doi.org/10.32629/jai.v7i3.1101
- Mansur, S., Sattar, K., Hosseini, S. E., Pervez, S., Ahmad, I., Saleem, K., & Zohier Elhendi, A. (2025). Sales forecasting for retail stores using hybrid neural networks and sales-affecting variables. *PeerJ Computer Science*, *11*, e3058. https://doi.org/10.7717/peerj-cs.3058
- Mostafa, S. A., Elzanfaly, D. S., & Yakoub, A. E. (2022). A Machine Learning Ensemble Classifier for Prediction of Brain Strokes. *International*

Journal of Advanced Computer Science and Applications, *13*(12). https://doi.org/10.14569/IJACSA.2022.0131232

- Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, Md. M., & Mridha, M. F. (2023). Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention. *IEEE Access*, *11*, 52288–52308. https://doi.org/10.1109/ACCESS.2023.3278273
- Rawat, M., Hosseini, S. E., & Pervez, S. (2023). Sentiment Analysis for Assessing Customer Satisfaction in Chatbot Service Encounters. *2023 16th International Conference on Developments in ESystems Engineering (DeSE)*, 105–109. https://doi.org/10.1109/DeSE60595.2023.10469554
- Rehman, A., Alam, T., Mujahid, M., Alamri, F. S., Ghofaily, B. Al, & Saba, T. (2023). RDET stacking classifier: a novel machine learning based approach for stroke prediction using imbalance data. *PeerJ Computer Science*, *9*, e1684. https://doi.org/10.7717/peerj-cs.1684
- Sushma Reddy, R, Dr. M. G., & G, N. (2023). Deep Transfer Learning Based Stroke Risk Prediction. *International Journal for Research in Applied Science and Engineering Technology*, *11*(9), 684–689. https://doi.org/10.22214/ijraset.2023.54857
- Wilson, J., Hosseini, S. E., & Pervez, S. (2023). Identification of Fake News in Social Media Using Sentimental Analysis. *2023 IEEE Industrial Electronics and Applications Conference (IEACon)*, 220–224. https://doi.org/10.1109/IEACon57683.2023.10370300