Stroke Risk Prediction Using Supervised Machine Learning Models

Jared Frazier*

Department of Computational Mathematics, Science and Engineering

Michigan State University and

https://github.com/jcf454/cmse492_project

(Dated: December 7, 2025)

Abstract

Stroke represents a leading cause of mortality and long-term disability worldwide, and the early identification of individuals at risk is critical for improving outcomes. Many traditional diagnostic approaches are unable to capture the complex and nonlinear interactions among demographic, lifestyle, and medical characteristics that drive stroke risk. This project develops an end-to-end machine learning system that predicts stroke occurrence from 5,110 health records. The pipeline includes structured preprocessing, exploratory data analysis, feature transformation, supervised model training, and rigorous evaluation on stratified validation and test sets.

Three models of increasing complexity were implemented: a logistic regression baseline, a Random Forest classifier, and a shallow neural network. Performance was evaluated using precision, recall, F1 score, ROC-AUC, and computational cost, with particular emphasis on recall due to the clinical consequences of false negatives. After threshold tuning and class reweighting, the shallow neural network achieved the strongest performance, with a validation recall of 0.82 and a ROC-AUC of approximately 0.83. These findings highlight the importance of nonlinear modeling and validate the feasibility of machine learning methods for medical risk prediction. The proposed framework provides a complete and reproducible modeling pipeline for stroke prediction and emphasizes the central role of preprocessing, model selection, and evaluation in health-related machine learning.

BACKGROUND AND MOTIVATION

Stroke continues to rank among the most prominent causes of death and prolonged neurological impairment globally. Early identification of at-risk individuals is crucial, with timely intervention known to markedly reduce both morbidity and mortality. However, stroke risk is determined by a combination of demographic, behavioral, and clinical factors, including age, hypertension, heart disease, glucose levels, lifestyle habits, and socio-economic conditions. These factors often interact nonlinearly and with subtle effects, making manual risk assessment and traditional scoring systems inadequate.

Many commonly used clinical tools are based on linear or additive assumptions and are not able to account for complex feature interactions. As a result, some high-risk patients may not be identified early enough for effective preventive care. This has severe consequences, as missed stroke risk can lead to avoidable long-term disability, reduced quality of life, and

significant healthcare costs.

Machine learning provides a potent alternative because it can learn nonlinear relationships, model higher-order feature interactions, and reveal patterns that may be challenging for clinicians to recognize. By training models on patient-level data, it becomes possible to generate individualized risk predictions that support clinical decision making.

The purpose of this project is to develop and evaluate a full-fledged predictive system that estimates stroke risk from routinely collected health data. The system is composed of data preprocessing, exploratory analysis, model development, hyperparameter tuning, evaluation, and interpretation. A robust predictive model has potentially profound implications: improved diagnostic sensitivity, earlier intervention, and mitigation of stroke-associated disability over the long term.

DATA DESCRIPTION

Data Origins

The dataset for this project comes from public health surveillance programs designed to monitor the risk of cardiovascular disease in various populations. It comprises de-identified patient records describing demographic information, clinical parameters, lifestyle factors, and comorbidity indicators. The dataset was assembled to support research on cardiovascular and stroke risk in real-world populations rather than being constructed as a toy or synthetic example.

Dataset Characteristics

The dataset contains 5,110 patient records. Numerical variables include age, body mass index (BMI), and mean blood glucose. Categorical variables include gender, work type, residence type, smoking status, and marital status. Binary indicators are provided for hypertension and heart disease. The target variable is a binary label indicating whether the individual has previously experienced a stroke.

This combination of numerical, categorical, and binary features results in a moderately wide tabular dataset that is well-suited for classical machine learning models and shallow neural networks.

Data Quality Analysis

Missing Values

Only a small number of BMI values were missing. Analysis of the missingness pattern suggested that these values were missing at random, possibly due to nonuniform clinical reporting or skipped measurements rather than structural bias. To maintain the numeric BMI distribution without creating artificial differences between groups, mean imputation was applied to fill in missing BMI values.

Class Balance

There is a significant class imbalance: only around five percent of patients are classified as having had a stroke. This imbalance complicates training and evaluation because naive models can achieve deceptively high accuracy by always predicting the majority class, while failing to identify positive cases. To address this imbalance, stratified splitting, class weighting, and appropriate metrics such as recall, F1 score, and ROC-AUC were used instead of accuracy.

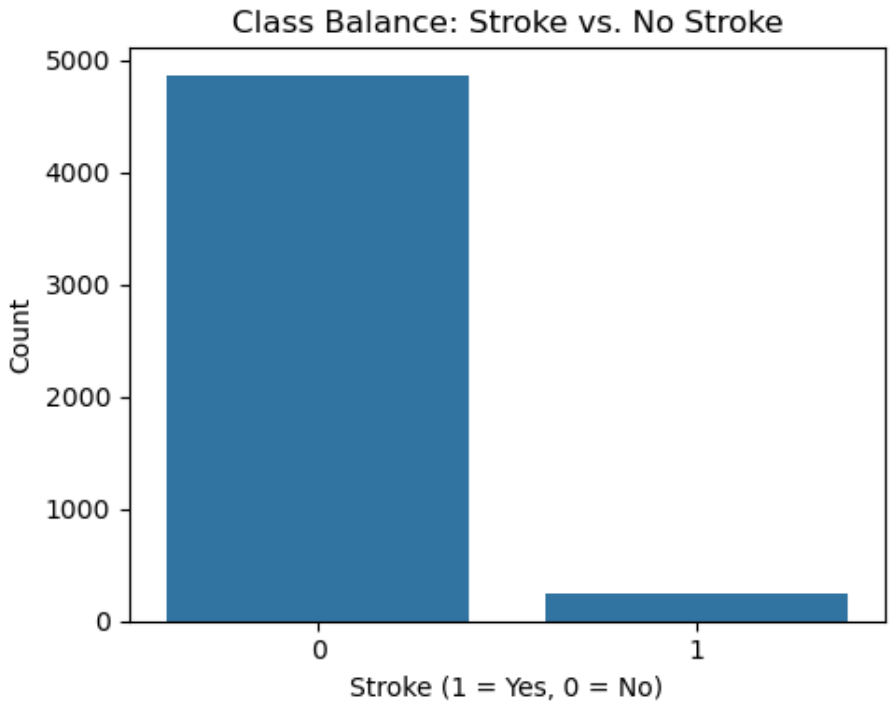

FIG. 1: Class distribution of stroke outcomes. The dataset is highly imbalanced, with stroke cases representing only about five percent of the samples.

Statistical Summary

Exploration of the data showed that the relationship between stroke occurrence and individual features is meaningful. Age showed the greatest correlation with stroke: the prevalence of stroke increased dramatically among older people. Hypertension, heart disease, and increased glucose were also linked to higher risk. These relationships were demonstrated visually using histograms, kernel density plots, and correlation matrices, and they supported the deployment of more expressive machine learning models capable of modeling nonlinear effects.

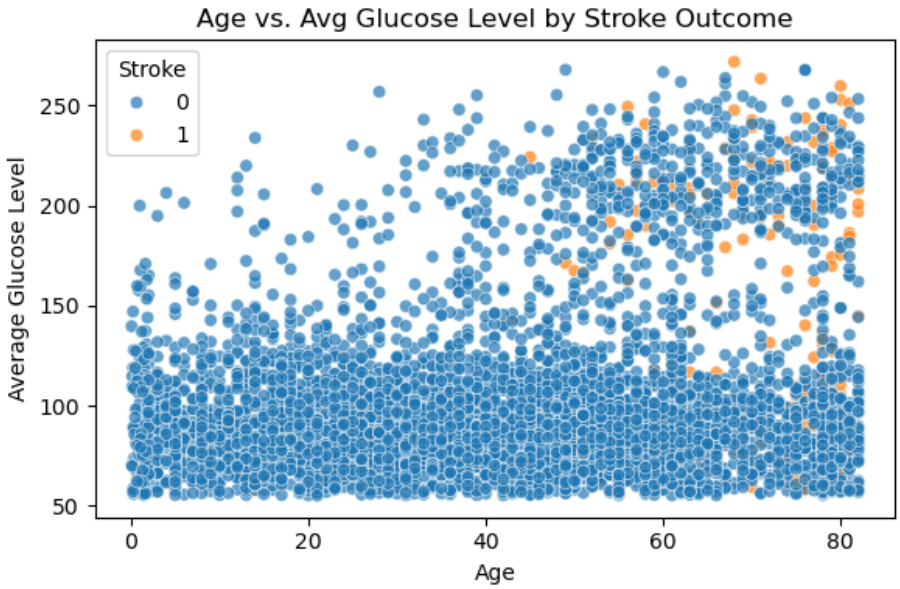

FIG. 2: Relationship between age, glucose level, and stroke occurrence. Stroke cases tend to cluster at higher ages and elevated glucose levels, supporting clinical expectations and motivating the need for nonlinear models.

PREPROCESSING

The preprocessing pipeline was set up to ready the dataset for powerful supervised learning while maintaining significant clinical trends.

Data Splitting

Due to the extremely imbalanced nature of the dataset, a 60/20/20 stratified split was performed to ensure consistent proportions of stroke cases across the training, validation, and test sets. Stratification prevented the minority class from being underrepresented in any single split and made evaluation more reliable.

Feature Engineering

Feature engineering was performed to investigate potential nonlinear associations. Age and BMI were binned to compare risk across strata and to study how stroke prevalence

changes across different age and BMI ranges. Although these engineered features were not ultimately used in the final models, they provided useful insight into the structure of the data and informed model selection decisions.

Scaling, Transformation, and Encoding

Categorical variables were represented using one-hot encoding, creating numeric feature vectors for all models. Missing BMI values were filled using mean imputation, as described above. Numerical features were standardized for the neural network model to stabilize gradient-based optimization. Standardization was not required for the Random Forest or logistic regression baselines, since tree-based methods and linear models with appropriate regularization are relatively robust to differences in feature scale.

Overall, preprocessing ensured that all models benefited from clean, numeric inputs and that imbalanced class distributions did not bias the assessment.

MACHINE LEARNING TASK AND OBJECTIVE

The problem addressed in this project is supervised binary classification: predicting whether someone has had a stroke based on a set of demographic, clinical, and lifestyle characteristics. Machine learning is well suited to this task because conventional clinical scoring methods are not able to easily capture the complex, nonlinear feature interactions that underlie stroke risk. In contrast, machine learning models can learn these relationships directly from data.

The main goal is to develop accurate models for clinically relevant measures, with particular emphasis on recall. False negatives in medical prediction problems are much more costly than false positives, because missing a high-risk patient may lead to catastrophic health consequences. As a result, model evaluation in this project focuses on recall, F1 score, and ROC-AUC instead of accuracy, which is misleading in highly unbalanced settings.

MODELS

Model Selection

Models of increasing complexity were developed to test how predictive performance improves with representational power.

Model 1: Logistic Regression

The first model, logistic regression, serves as a simple and easily interpretable linear baseline. It provides information about the extent to which linear relationships are sufficient for predictive performance and serves as a reference for more complex models.

Model 2: Random Forest Classifier

The second model is a Random Forest classifier, an ensemble classification method that combines multiple decision trees to model nonlinear structure and feature interactions. Random Forests perform well on mixed data types and provide feature importance estimates, which are particularly useful for tabular medical data.

Model 3: Shallow Neural Network

The final model is a shallow neural network with two fully connected hidden layers (32 and 16 units, respectively), ReLU activation functions, dropout regularization, and a sigmoid output layer. This architecture strikes a balance between expressive power and avoiding overfitting on a limited dataset. Early stopping and dropout layers were employed to ensure stability in training and generalization.

These three models span a broad range of simple, moderately complex, and highly flexible approaches, enabling a comprehensive comparison across algorithmic families.

Regularization and Hyperparameter Tuning

For each model, training procedures and regularization strategies were tailored to their structure. Logistic regression used L2 regularization through the inverse regularization strength hyperparameter. Random Forest training controlled model complexity through limits on tree depth, number of trees, and minimum samples per split, with these hyperparameters tuned via a constrained randomized search. The neural network employed dropout layers, class weighting, and early stopping, and its architecture and learning rate were tuned to achieve stable convergence without overfitting.

TRAINING METHODOLOGY

Loss Functions

Logistic regression and the neural network were trained using the binary cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right], \tag{1}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability of stroke.

Random Forest training minimized Gini impurity at each decision node:

$$G = \sum_{k=1}^{K} p_k(1-p_k), \tag{2}$$

where p_k is the proportion of samples in class k in a node.

Training Process

For each model, training procedures were adapted to its structure and needs. Logistic regression was trained with binary cross-entropy loss and appropriate regularization. Random Forest training reduced Gini impurity at every decision node, and important hyperparameters such as tree count, depth, and split thresholds were optimized via constrained randomized search.

The neural network was trained using the Adam optimizer and binary cross-entropy loss. Since the dataset is heavily imbalanced, class weights were used to increase the penalty on

misclassifying stroke cases. Regularization was applied through dropout layers, and validation loss was monitored with early stopping to avoid overfitting. The neural network's training curves, including loss and recall over epochs, were plotted to evaluate model stability.

Random Forest and neural network models both underwent threshold tuning. The default probability threshold of 0.50 is not appropriate for imbalanced data, so a sweep over candidate thresholds on the validation set was carried out. The optimal threshold parameters for the Random Forest and the neural network were found to be approximately 0.10 and 0.50, respectively.

Model Summary Table

TABLE I: Summary of models, parameters, and training methodology.

Model	**Parameters**	**Hyperparameters**	**Loss Function**
Logistic Regression	Weights per feature	C, penalty	Binary cross-entropy (L
Random Forest	Tree structures	n_estimators, max_depth, min_samples_split	Gini impurity
Shallow Neural Network	Dense layer weights	Learning rate, batch size, epochs	Binary cross-entropy (L

METRICS

Primary Metric

Evaluation metrics were chosen to reflect the data imbalance and the clinical significance of correctly identifying positive stroke cases. The primary metric is recall, which measures the proportion of actual stroke cases that are correctly identified. This reflects the importance of minimizing false negatives in a medical context.

Secondary Metrics

Secondary metrics include precision, which counts the proportion of predicted stroke cases that are truly positive; the F1 score, which balances precision and recall; and ROC-AUC,

which measures the model's ability to rank positive cases ahead of negative cases across thresholds. Accuracy was deliberately avoided as a main outcome metric, as it is misleading in imbalanced conditions—a simple majority-class predictor can exceed 95% accuracy while having no clinical relevance.

Metric Definitions

The F1 score is defined as:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}. \tag{3}$$

ROC-AUC evaluates the area under the receiver operating characteristic curve, summarizing the trade-off between true positive and false positive rates across probability thresholds.

RESULTS AND MODEL COMPARISON

Performance Comparison

Baseline models provided a reference point for evaluation. The majority-class baseline was unable to detect any stroke cases and had zero recall and F1 score. Logistic regression showed slight improvement but was restricted by its linear structure and inability to capture nonlinear relationships among features.

The Random Forest classifier showed significant gains, reaching recall of approximately 0.50 and ROC-AUC of about 0.80 after threshold tuning. These observations indicate that the model successfully captured nonlinear interactions but still struggled with the very low prevalence of stroke cases.

The shallow neural network exhibited the best overall performance. With a probability threshold of 0.50, it achieved a validation recall of 0.82, an F1 score of 0.24, and a ROC-AUC greater than 0.83. These measures show that the neural network is superior at learning complex feature interactions compared with the other models. Training time (around five seconds) was low, confirming that the implementation of such neural networks is computationally practical.

TABLE II: Model performance metrics on the validation set.

Model	**Precision**	**Recall**	**F1**	**ROC-AUC**
Majority Baseline	0.000000	0.00	0.000000	N/A
Logistic Regression Baseline	1.000000	0.02	0.039216	0.842160
Random Forest	0.137143	0.48	0.213333	0.810864
Shallow Neural Network	0.140893	0.82	0.240469	0.834156

Computational Efficiency

TABLE III: Training and inference time for each model.

Model	**Training Time (s)**	**Inference Time (s)**	**Hardware Used**
Majority Baseline	N/A	N/A	CPU
Logistic Regression Baseline	N/A	N/A	CPU
Random Forest	0.4103	0.0571	CPU
Shallow Neural Network	4.7203	0.2859	CPU

Analysis and Discussion

In general, the shallow neural network offered the best compromise between sensitivity, discrimination capability, and computational efficiency. Its performance confirms that non-linear models are better suited to this problem than purely linear methods. The Random Forest improved substantially over the baseline but did not match the neural network in recall or ROC-AUC. Logistic regression underperformed due to its limited expressive capacity.

MODEL INTERPRETATION

Although neural networks are less interpretable than tree-based models, understanding feature contributions using the Random Forest's feature importance analysis helped validate the neural network's behavior. Age was the strongest predictive factor, as documented in

existing medical literature. Predictions by the models were also significantly influenced by increases in glucose levels, hypertension, and heart disease.

Permutation importance and SHAP analyses on a subset of the data indicated that the models relied on clinically meaningful features rather than noise. These interpretability results enhance confidence in the predictions of the neural network and support its use as a decision-support component rather than a purely black-box system.

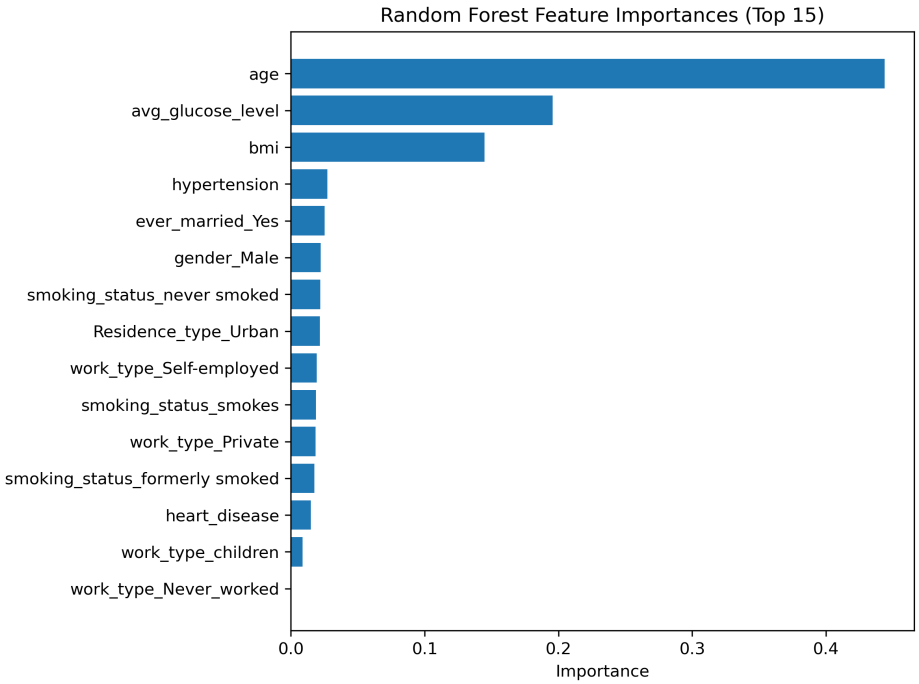

FIG. 3: Top 15 feature importances from the Random Forest classifier.

CONCLUSION

Summary of Findings

In this project, a complete machine learning pipeline for stroke prediction was devised, including data preprocessing, exploratory analysis, model development, threshold tuning, evaluation, and interpretation. Through gradual and incremental experimentation with increasingly complex models, the shallow neural network was found to be the best model, achieving strong recall and ROC-AUC performance on an imbalanced dataset.

Limitations and Future Work

Although the model performed fairly well, precision was modest due to the infrequent occurrence of stroke cases. Future work could exploit more sophisticated oversampling approaches, calibrated probability estimates, richer feature sets from electronic health records, or more complex neural network architectures to improve precision without sacrificing recall.

Final Remarks

This project demonstrates the application of machine learning for medical risk prediction and presents a well-documented reproducible pipeline capable of supporting the early detection of individuals at high risk of stroke. The work also illustrates the importance of preprocessing, model selection, and evaluation choices when dealing with highly imbalanced clinical data.

---

* frazi153@msu.edu

[1] Scikit-learn developers, "Scikit-learn: Machine Learning in Python," https://scikit-learn.org.

[2] World Health Organization, "Global Health Estimates: Stroke Mortality and Burden," (2022).