Statistical and Machine Learning Approaches to Predicting Stroke Risk

INTRODUCTION AND MOTIVATION

Stroke is a leading cause of mortality and long-term disability worldwide (1) and represents a significant burden on healthcare systems around the world. Early identification of stroke in high-risk individuals may allow preventative measures to be implemented, possibly reducing the social cost and personal impact of stroke. With the increasing collection of clinical, demographic, and lifestyle data, statistical and machine learning methods offer tools for identifying individuals at risk.

In this project, we investigate if stroke outcome can be predicted using clinical, demographic, and lifestyle variables. We compare statistical methods with machine learning and deep learning, specifically, we examine what impact class imbalance and classification threshold selection has on predictive performance.

Research questions

How does predictive performance compare between:

- Classical statistical models
- Machine-learning models
- Deep-learning models?

How do class imbalance and probability threshold selection affect model evaluation and clinical usefulness?

How accurately can statistical models and machine learning predict stroke risk using regularly collected clinical, demographic, and lifestyle data?

To answer these questions, the project fits four models; logistic regression, LASSO-penalised logistic regression, random forest, and feed-forward neural network to a stroke prediction dataset and evaluates them on a split into training set. Performance is compared using ROC-AUC and PR-AUC as well as threshold-based metrics at both a standard 0.5 cutoff and a lower prevalence-based threshold, to show the impact of class imbalance on performance.

DATA DESCRIPTION AND QUALITY CHECK

The dataset used in this project is the 'Stroke Prediction Dataset' available from Kaggle.com https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data (FEDESORIAN, 2021)

The dataset contains 5,110 observations, 12 variables. This includes a binary response variable (stroke), an identifier variable (id), and 10 predictor variables that can be arranged into:

- Clinical variables: hypertension, heart disease, average glucose level, BMI
- Demographic variables: gender, age, residence type
- Lifestyle variables: smoking status, work type, marital status

Prior to model fitting, the dataset was assessed to determine its suitability for analysis. This included examining variable names, dimensions, and any missing or invalid entries. All categorical predictors (gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status) were converted to factor variables. This ensures that R implements them correctly during model fitting. The response variable stroke was recoded as a factor with levels "No" and "Yes", where "Yes" represents the positive class.

The gender variable contained a single 'Other' entry. This was removed as a single observation risks overfitting. The missing data assessment revealed 201 missing entries from the variable bmi. These missing values were imputed using the median value (28.1 kg/m^2). The missing entries were imputed with the median to maintain sample size while being robust to outliers. The identifier variable id was removed as it provides no statistical information, however, it could cause instability in some models. Table 1 summarises the dataset characteristics and key data-quality checks after preprocessing.

Metric	Value
Observations (before cleaning)	5110
Observations (after cleaning)	5109
Number of predictors	10
Missing BMI values (before imputation)	201
Stroke cases (Yes)	249
Non-stroke cases (No)	4860

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory data analysis was conducted to understand the characteristics of the dataset and the relationships between the predictor variables and the response variable. The focus was on patterns in the data and checking that they were clinical plausible rather than hypothesis testing.

The first exploratory step was to examine the stroke variable itself. A bar plot of the distribution of stroke (Figure 1) shows an extreme class imbalance, with far more "No Stroke" than "Stroke" cases. This imbalance requires careful consideration for model evaluation, implying that accuracy here is likely misleading.

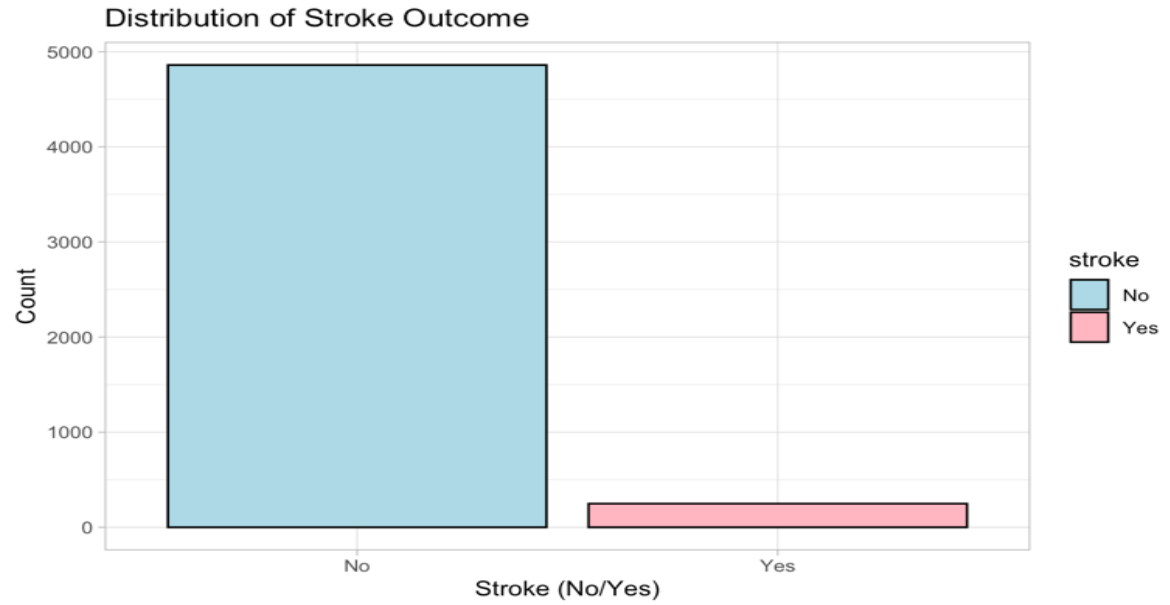

Figure 1

Using contingency tables and stacked proportional bar plots associations between stroke and categorical predictors were examined. For hypertension and heart disease, the bar plots (Figures 2 and 3) show a noticeably higher proportion of stroke among individuals with those conditions compared with those without. These patterns are clinically plausible and consistent with established evidence that hypertension and cardiovascular disease are major risk factors for stroke (2) (3). Similar, but slightly weaker, differences were seen across smoking-status categories, with current and former smokers having higher stroke proportions than never-smokers.

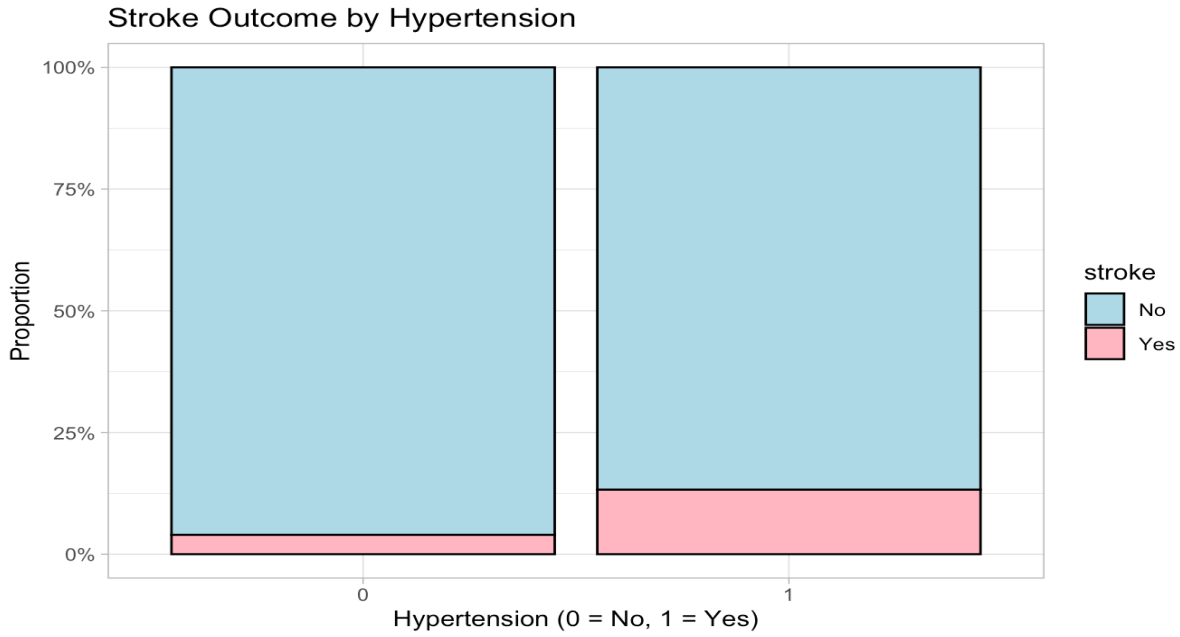

Figure 2

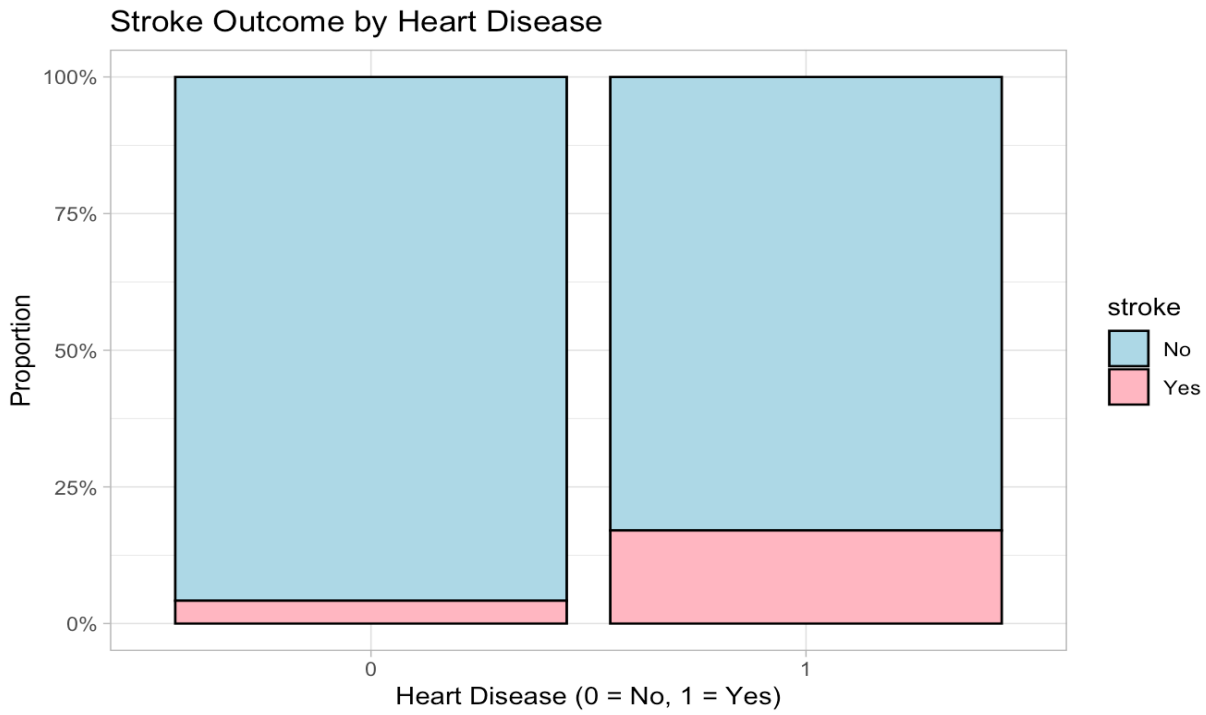

Figure 3

Numerical predictors were examined using boxplots stratified by stroke outcome. Age showed a strong relationship with stroke outcome as seen in figure 4. Individuals with positive stroke outcome were generally older, with higher medians and upper quartiles than those without stroke.

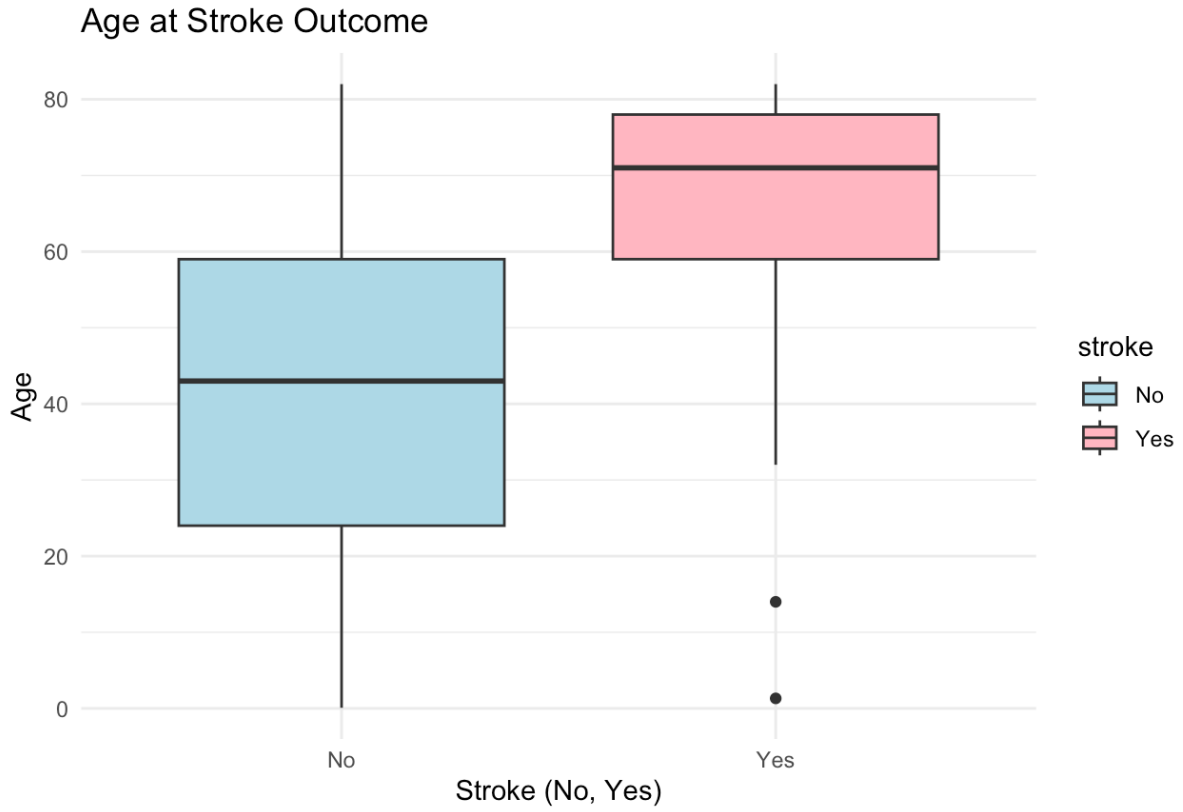

Figure 4

Average glucose level (Figure 5) displayed a similar, though less pronounced, shift towards higher values in the stroke group, whereas BMI distributions overlapped considerably, suggesting a weaker association in this dataset. A small number of high values were visible for glucose and BMI, but these fell within clinically plausible ranges and were therefore retained rather than treated as outliers or removed.

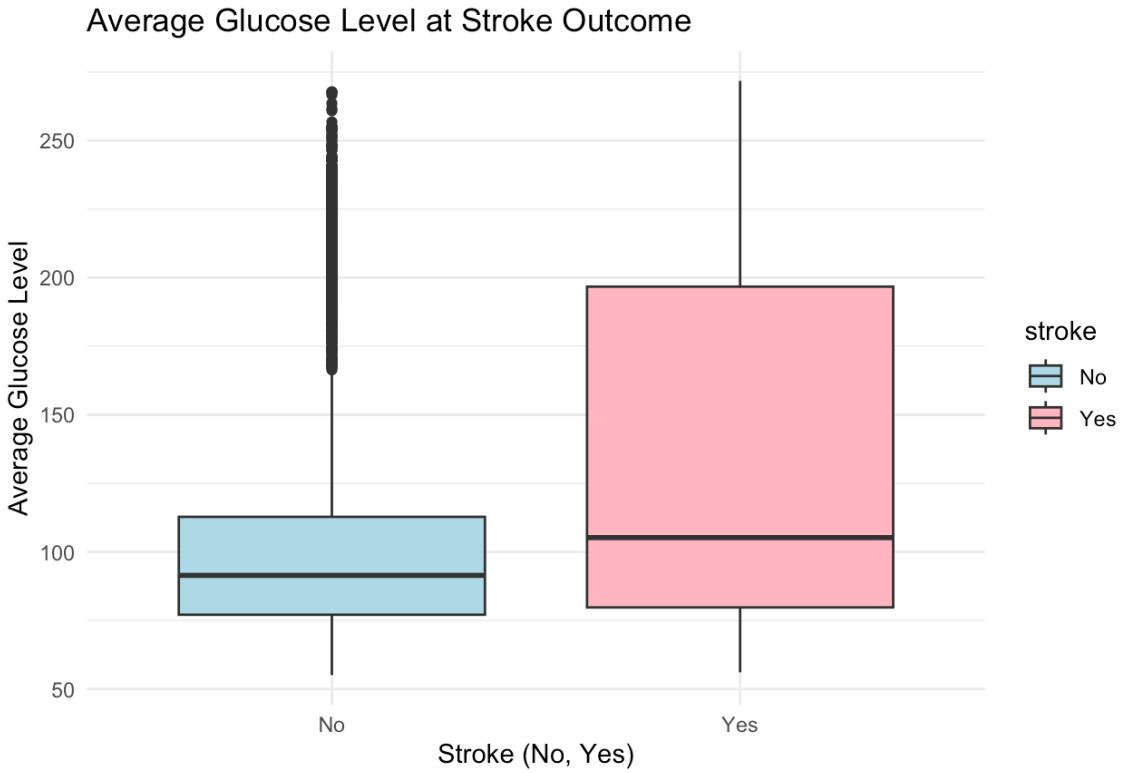

Figure 5

The exploratory analysis suggests that stroke is more common among older individuals, those with hypertension or heart disease, and those with higher glucose levels, with smoking showing a weaker but similar pattern; these clinically plausible trends are patterns that the subsequent models are designed to learn and reflect.

METHODOLOGY

After cleaning, the data was split into (80%) training and (20%) test sets. Since the outcome is heavily imbalanced, the split used stratified sampling on the stroke outcome to preserve the class distribution in both sets. The resulting training set had 4,088 observations and the test set had 1,022. Stratified sampling is extremely important with class imbalance because simple splits can allocate few positive cases to one divided. A fixed random seed (15) was set to ensure reproducibility .

To allow for fair comparison, all models were trained using the same predictor set.

Logistic Regression

A standard logistic regression model was used as a baseline for predicting stroke outcome. It assumes that a predictor has linear effect on the log odds of a stroke, as well as observations being independent. The coefficients of the model can be exponentiated to give odds ratio. This shows how much the odds of a stroke change when a predictor increases by one unit. Logistic regression is widely used in medical research due to it being easy to interpret and effective for binary outcomes, hence why it was used in this project.

LASSO Logistic Regression

A LASSO-penalised logistic regression model was also fitted to predict stroke outcome. Like standard logistic regression, it models the log-odds of stroke as a linear function of the predictors, but it adds an L1 penalty that shrinks some coefficients towards zero. Ten-fold cross-validation (cv.glmnet) was used to choose the strength of this penalty. The model was refitted on different training folds, and performance was evaluated across a grid of λ values. Two key values are returned, lambda. Min, which gives the lowest cross-validated deviance, and lambda.1se a slightly larger penalty that is within one standard error of the minimum and produces a simpler model. In this project the model at lambda.min was selected so that predictive performance was prioritised while still shrinking weak predictors, making LASSO a useful complement to the baseline logistic regression when several correlated variables are present.

Random Forest

A random forest was fitted using 500 decision trees, with 3 predictors randomly chosen at each split. Adding more trees would usually give very small increases in accuracy but also increasing computation time. The value $mtry = 3$ was given as it is standard to set $mtry$ to approximately the square root of the number of predictors. With 10 predictors in the dataset, $\sqrt{10} \approx 3$. Each tree was trained on a bootstrap sample of the training data, and the final prediction for each individual was obtained by taking the majority vote across all trees. This reduces variance and allows the model to capture non-linear relationships and interactions.

Feed-forward neural network

A feed-forward neural network was trained using the h2o framework after converting the training and test sets into h2o objects. The network had two hidden layers with 16 and 8 providing enough capacity to model non-linear relationships while avoiding an overly deep architecture that could overfit given the size of the dataset, and used Rectifier (ReLU) activation functions, and was trained on 50 epochs and a fixed seed. This model represents deep learning, and the units and epochs were chosen to provide some non-linear modelling while still being simple enough to train reliably on the sample.

Model evaluation and threshold selection

Each model outputs a predicted probability that an individual has experienced a stroke. These probabilities were first evaluated using threshold-independent performance measures**,** specifically the Receiver Operating Characteristic (ROC) curve and its area under the curve (ROC-AUC), and the Precision–Recall (PR) curve and its area under the curve (PR-AUC). ROC-AUC measures how well a model can rank individuals from lower to higher stroke risk across all possible thresholds, while PR-AUC focuses on performance for the positive stroke class. PR-AUC is particularly informative in this setting due to the low prevalence of stroke. This allows models to be compared without fixing a single cutoff and therefore provide an assessment of overall discrimination of threshold choice.

To convert predicted probabilities into binary stroke / no-stroke classifications, two probability thresholds were considered. First, the conventional cut-off of 0.5 was applied, and standard confusion-matrix metrics (accuracy, sensitivity, specificity, precision, and F1 score) were computed. Because stroke is rare in the dataset, this threshold caused all models to classify almost all individuals as “no stroke”, resulting in high overall accuracy but near-zero sensitivity, which is clearly unsuitable for screening applications.

To address this limitation, a second, lower threshold equal to the observed stroke prevalence in the training data (approximately 4.9%) was used. At this prevalence-based threshold, models classified more individuals as high risk, substantially increasing sensitivity and F1 score at the expense of reduced specificity and accuracy. This highlights a fundamental trade-off in medical prediction: lowering the classification threshold reduces false negatives and improves case detection but increases the number of false positives, which may place additional burden on healthcare resources. Reporting results at both thresholds, alongside ROC and PR curves, provides a more complete and clinically relevant assessment of model performance and illustrates why sensitivity-oriented thresholds are often preferred in risk-screening contexts

RESULTS

Model fitting and interpretation (Logistic regression)

Predictor	OR	95% CI
Hypertension: Yes (vs No)	1.591	1.113 - 2.274
Heart disease: Yes (vs No)	1.192	0.774 - 1.836
Age (per year)	1.08	1.066 - 1.094
Gender: Male (vs Female)	1.071	0.786 - 1.46

Residence: Urban (vs Rural)	1.06	0.784 - 1.435
BMI (per unit)	1.01	0.986 - 1.035
Average glucose level (per unit)	1.003	1 - 1.005
Smoking: Smokes (vs Unknown)	0.965	0.592 - 1.574
Smoking: Never (vs Unknown)	0.877	0.601 - 1.279
Smoking: Unknown (reference)	0.875	0.552 - 1.388
Ever married: Yes (vs No)	0.765	0.471 - 1.24
Work: Private (reference)	0.308	0.059 - 1.617
Work: Govt job (vs Private)	0.254	0.047 - 1.387
Work: Self-employed (vs Private)	0.222	0.04 - 1.22

The logistic regression results (Table 2) indicate that age and hypertension are the strongest predictors of stroke. Each additional year of age increases the odds of stroke by approximately 8%, while individuals with hypertension have substantially higher odds compared to those without. Average glucose level shows a small positive association, whereas BMI and most categorical variables do not exhibit clear effects after adjustment. Overall, age, hypertension, and glucose level emerge as the most influential predictors in the model.

Discrimination using ROC and PR curves

All four models were first compared using ROC and precision–recall (PR) curves on the test data. The logistic-regression model achieved the highest ROC AUC (0.850), with LASSO very close (0.849), followed by the random forest (0.838) and the neural network (0.812), showing that all models can rank individuals by stroke risk reasonably well but that the simple GLM performs at least as well as the others. The PR-AUC values were similar. GLM (0.178) was slightly better than LASSO (0.159) and random forest (0.155), with the feed-forward lower (0.124), suggesting that logistic regression gives the best overall discrimination for the rare stroke class in this dataset.

At the standard 0.5 threshold

When the default probability cutoff of 0.5 was used, all models predicted almost everyone in the test set as “no stroke”. This gave very high accuracy (around 95%) and almost perfect specificity, but sensitivity was 0, meaning that no stroke cases were correctly detected, and precision was effectively zero. These results show that accuracy at a 0.5 threshold is misleading and that, if used as a screening tool, such a classifier would fail to identify patients who have had a stroke.

At a prevalence-based threshold

At the lower threshold equal to the stroke prevalence in the training data around (0.049) performance changed in a more clinically useful way. Logistic regression reached sensitivity 0.816 and specificity (0.737), LASSO sensitivity (0.857) and specificity (0.713), and random forest sensitivity (0.755) and specificity (0.752), so these models started to detect a large proportion of stroke cases while still correctly classifying most non-stroke patients. The neural network remained more conservative, with sensitivity (0.326) but higher specificity (0.901) and accuracy (0.874), meaning it missed many stroke cases compared with the other models.

Across all models, precision at the prevalence-based threshold stayed around 0.13–0.14, because strokes are rare and many flagged individuals do not actually have a stroke. F1 scores summarised this trade-off. The GLM achieved the highest F1 (0.232), with LASSO and random forest slightly lower and the neural network lowest (0.199), indicating that logistic regression gave the best balance between to catching strokes while reducing the false positives under this lower cutoff.

Discussion

The comparison of models shows that a simple logistic regression performs at least as well as more complex machine learning methods. The ROC-AUC and PR-AUC values for logistic regression are slightly higher than those for LASSO and random forest and clearly higher than those for the neural network, so the added flexibility of more complex models does not translate into noticeably better discrimination of stroke cases.

The results also highlight how critical class imbalance and threshold selection are for medical prediction. At the default 0.5 cutoff, all models achieved high accuracy but failed to identify any stroke cases, showing that accuracy and even ROC-AUC can be misleading when the positive class is rare. The prevalence-based threshold dramatically increased sensitivity and F1 scores for the GLM, LASSO, and random forest models, but reduced specificity and accuracy, displaying the fundamental trade-off between avoiding false negatives and reducing false positives that could overload services.

Limitations and Conclusion

This project had several limitations that should be considered when interpreting the results. Firstly, the dataset has an extreme imbalance, with less than 5% of observations being a positive outcome for stroke. This makes evaluation difficult. While

threshold adjustments partly handle this, more advanced imbalance handling were not investigated in this project.

Missing BMI values were addressed using simple median imputation, although robust, it does offer uncertainty about the missing values and relationship between BMI and the other predictors. The dataset is only observational with limited information about sampling quality with lack of other important risk factors.

This project evaluated the ability of statistical and machine learning algorithms to predict stroke outcome using commonly collected clinical, demographic, and lifestyle variables. Although the limitations exist, the results suggest that logistic regression evaluated at a suitably low threshold achieved the best balance between the discrimination (ROC-AUC and PR-AUC) and the threshold-based metrics (sensitivity, specificity, precision, F1). LASSO produced very similar discrimination, whereas random forest and feed-forward very little improvement. For screening settings where missing a stroke case carries a high cost, using a GLM or LASSO model with a calculated low, sensitivity-focused threshold is a reasonable strategy, so that individuals flagged as high risk can be referred for further clinical assessment rather than treated solely on the model output. This could be extended by adding more advanced techniques for class imbalance and missing-data handling, carrying out more structured tuning and calibration of complex models, and testing all models on independent datasets to check how well they generalise in the real world.

References

(1)https://pmc.ncbi.nlm.nih.gov/articles/PMC11786524/pdf/10.1177_17474930241308142.pdf
(2) https://www.nhs.uk/conditions/stroke/causes/
(3) https://pmc.ncbi.nlm.nih.gov/articles/PMC4419105/pdf/WJCC-3-418.pdf