Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 1, 2026

PREDICTIVE MODELLING OF ISCHEMIC STROKE RISK: A SYSTEMATIC COMPARISON OF SVM AND LSTM ARCHITECTURES UNDER MULTI-METRIC ASSESSMENT CRITERIA

[1]Fariha Shahid, [*2]Aftab Ahmed Chandio, [3]Qamar-ul-Nisa Chandio, [4]Farhat Noureen Memon

[1]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan
*[*2]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan*
[3]Government Degree Boys College, Qasimabad, Ministry of Education and Literacy Government of Sindh, Hyderabad 76000, Pakistan
[4]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan
[1]farihashaikh2015@gmail.com, [*2]chandio.aftab@usindh.edu.pk, [3]qamaraftabchandio@gmail.com, [4]farhatnm@usindh.edu.pk

DOI:

Article History
Received on 25 Dec, 2025
Accepted on 27 Jan, 2026
Published on 28 Jan, 2026

Corresponding Author:
Aftab Ahmed Chandio

Abstract

Background: Ischemic stroke is a major cause of death and disability and is increasingly common, and the ability to recognize those at high risk is vital in its prevention. Machine learning and deep learning techniques have provided a means of improving stroke predictions. However, the choice of modelling framework that provides the most accurate and clinically meaningful predictions, particularly in the context of the important balance between sensitivity and precision, has not been properly investigated. Methods: This paper develops and compares four stroke prediction models a Support Vector machine (SVM), Random Forest (RF), XGBoost, and a Long Short-Term Memory (LSTM) network- that are trained and assessed using a structured dataset of 50,000 patient records with demographic and clinical risk factors. Preprocessing included missing data that was imputed using the median value, one-hot encoding of categorical variables, and standard scaling. Each of the four models was trained, validated and evaluated on the same data splits (70%-15%-15) using stratified sampling. Performance was assessed using a multi-metric model covering accuracy, precision, recall, F1-score and ROC-AUC, log loss, Brier score, Cohens Kappa and Matthews Correlation Coefficient (MCC). Results: The SVM had the best overall accuracy (0.9556), precision (0.8517) and F1-score (0.8521) and better calibration and agreement statistics. The Random Forest produced a closely comparable performance profile (accuracy = 0.9543; precision = 0.8491; F1 = 0.8472; AUC = 0.9867). The highest sensitivity-oriented model was exhibited by XGBoost (recall = 0.9022; accuracy = 0.9435; AUC = 0.9852), with the highest recall of all the models (0.9724). There was a close relationship in values between ROC-AUC across the four models (0.9852-0.9877) indicating similar rank-order discriminative ability regardless of the complexity of architecture. Discussion: The results indicate that finely hybridized classical machine learning models are still highly competitive when compared to ensemble and deep learning models on structured tabular health data. The present study suggests the framework of model selection to be used in clinically grounded diagnostic support applications, namely: SVM and Random Forest are recommended to be used in high-sensitivity diagnostic support applications, whereas XGBoost and LSTM are more effectively designed to operate in high-sensitivity screening settings. The multi-metric benchmarking framework which can be proposed in this research can offer a better way of model evaluation than a single-metric evaluation approach.

1. Introduction

Stroke is a leading cause of death and disability in the world. The latest epidemiological evidence shows the incidence of stroke is ticking upwards, an increase driven by globalization, growing urbanization and rising levels of potentially modifiable risk factors such as hypertension, diabetes mellitus and heart disease [1], [3]. Risk prediction tools to identify those at risk can help reduce morbidity and mortality, and early detection can greatly improve treatment results [4]. However, these conventional approaches to stroke prediction are often based on a limited number of clinical factors with linear statistical algorithms which may be unable to capture the complexity of risk factors and their interaction [4].

The recent advancements in artificial intelligence technology have led to machine learning (ML) and deep learning (DL) being used for health outcome predictions and to assist decision making. They can manage large amounts of clinical data and can discover non-linear relationships, which can be difficult to detect using conventional statistical methods [5], [6]. In terms of modelling techniques, Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) networks have attracted attention in terms of their predictability. SVM works well with structured, low to medium-dimensional structured data through the generation of optimal classification boundaries in high-dimensional feature spaces [7]. LSTM, on the other hand, makes use of a memory mechanism aimed at modelling long-term temporal information and non-linear relationships [8].

1.1 Critical Review of Existing Literature

ML and DL have been used to predict the occurrence of stroke with encouraging results in terms of both the predictive performance and clinical potential [9]–[11]. But a systematic review of this research reveals a number of common methodological pitfalls that limit the generalizability and impact of published work.

Chun et al. [12] trained logistic regression, random forest and gradient boosting models on a large (0.5 million) prospective cohort of Chinese adults and achieved C-statistics ranging from 0.718 to 0.832. This study benefits from the statistical strength afforded by this large dataset, but it focused exclusively on classical ML models and did not investigate how DL approaches perform on the same feature set, so it remains uncertain whether they can learn additional information from these features. Premisha et al. [9] applied ensemble methods, such as random forest, XGBoost and AdaBoost, to a Kaggle-based stroke dataset of around 5,000 observations, resulting in accuracies between 0.91 and 0.94. But they only used accuracy as the evaluation metric, which can be deceiving in clinical applications where datasets are often imbalanced, and the high accuracy could be a result of the algorithm correctly predicting the majority class [20]. Si et al. [13] showed that optimized feature selection and gradient boosting or SVM improve the prediction of stroke risk in patients who have had coronary artery disease treated using revascularization, but this study was limited to a narrow sub-population and used a limited set of criteria to compare the models.

The second limitation is that most of the published stroke prediction studies use accuracy and/or ROC-AUC as the sole or main metric for evaluation [19]. According to Chicco and Jurman [20] the Matthews Correlation Coefficient (MCC) should be used as the primary metric to evaluate binary classification, as it uses all elements of the confusion matrix and is a more balanced metric than accuracy or ROC-AUC. Chicco et al. [21] also showed that the MCC outperforms Cohen's kappa and the Brier score for binary prediction. Van Calster et al. [16] stressed the need for calibration

metrics (e.g. log loss, Brier score) to guide the use of predicted probabilities in the clinical setting. While new information metrics are emerging, multi-metric evaluation of stroke prediction is still rare.

1.2 Research Gap and Problem Statement

The evidence presented above suggests there is indeed a gap in literature. While some studies have shown the benefit of using either ML or DL models for stroke forecasting, none have yet undertaken a controlled comparison of a traditional ML model (SVM) and a DL model (LSTM) under the very same training and evaluation conditions (including identical data, identical data preprocessing and feature engineering, identical data partitions), experimented with through a multi-metrics systematic evaluation framework that covers various threshold dependent metrics (accuracy, precision, recall, F1-score, sensitivity and specificity), threshold independent metric (ROC-AUC), calibration metrics (log loss, Brier score) and agreement metrics (Cohen's Kappa and MCC). This lack of controlled cross-paradigm comparison, and the focus on narrow evaluation criteria means that the performance trade-offs of these two different model paradigms for classical tabular data remain poorly understood in the context of health data.

This study fills in this key gap. In particular, this study aims to: (1) design and evaluate supervised SVM and LSTM models for binary stroke prediction, using the same, well-controlled comparison; (2) compare the models using a multi-metric approach that captures different aspects of binary classification; and (3) make clinical recommendations about the use of the two models based on the desired clinical context, specifically screening versus diagnostic support.

The research process is outlined in Figure 1, which shows how the data are processed, features engineered, and then models constructed, evaluated and compared.

```
Raw dataset shape: (50000, 13)
[WARNING] Target column 'stroke' has only one class. Re-engineering labels.
Re-engineered stroke distribution:
 stroke
0    42500
1     7500
Name: count, dtype: int64

Numeric features : ['age', 'hypertension', 'heart_disease', 'avg_glucose_level', 'bmi', 'health_risk_index', 'age_glucose_interaction', 'ag
Categorical features: ['gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status']

Train : 34999 samples  | Stroke rate: 0.150
Val   : 7501 samples  | Stroke rate: 0.150
Test  : 7500 samples  | Stroke rate: 0.150

Total features after encoding: 24
Feature names: ['age', 'hypertension', 'heart_disease', 'avg_glucose_level', 'bmi', 'health_risk_index', 'age_glucose_interaction', 'age_bm

Preprocessing complete. Arrays saved to: stroke_data
X_train shape: (34999, 24)
X_test  shape: (7500, 24)
```

Figure 1: Dataset structure, feature distribution, and class balance after preprocessing.

2. Materials and Methods

2.1 Study Design and Data Source

The study conducted a quantitative experimental study that involves a controlled comparison of two modelling approaches for the binary prediction The evaluation is based on an accessible structured clinical data set of 50,000 patients available on Kaggle containing patient demographics and clinical risk factors such as age, gender, hypertension, heart disease, average glucose level, BMI, smoking status, work type and residence type. The outcome variable is a binary measure of stroke (1) or no stroke (0). The modelling of structured clinical datasets is a common and accepted

approach to predicting patient outcomes, as these datasets capture real-world meaningful patient characteristics and risk factors [12]. The data are fully deidentified without personal information, and do not require institutional review board approval.

2.2 Data Preprocessing

A preprocessing workflow was established to enhance data quality and compatibility with the model. For the BMI feature, which contained missing data, statistical imputation (median imputation) was applied, which is common practice in datasets from the healthcare sector that does not compromise dataset integrity and also reduces the impact of outliers on the imputed data points [13]. Categorical features such as gender, job type, house type and smoking status were encoded in numeric format using one-hot encoding (as opposed to ordinal encoding) due to the fact that the categorical variables used do not have an inherent ordering, and one-hot encoding avoids the creation of artificial ordering that can influence model training [14].

All continuous features were normalized to scale the feature space. In the case of the SVM model, this is crucial to avoid influence from differences in magnitude and distribution of features, given that the model's optimization process is based on distance in multi-dimensional feature space [7]. Normalization also helps with gradient optimization and stability of the LSTM model.

2.3 Feature Engineering

Feature engineering was performed to capture clinical associations between variables, and to improve predictive power of both models. A set of interaction terms (such as age by glucose interaction and age by BMI interaction) were engineered to capture interactions between critical health markers. These interactions take into account the clinical significance that increasing age in combination with increasing glucose levels or increasing BMI led to a synergistic effect in the occurrence of stroke [4]. Moreover, an aggregate health risk index was built to capture composite risk score in patients using multiple clinical variables. Such feature engineering techniques can improve model accuracy in health prediction by expanding the dependence of algorithms on the underlying patterns in structured data [15]. The engineered feature vector (after feature engineering) consisted of twenty-four predictors.

2.4 Dataset Partitioning

The dataset was divided into three non-overlapping subsets: a training set (70%, n = 34,999), a validation set (15%, n = 7,501), and a test set (15%, n = 7,500). We used stratified sampling to ensure that the number of stroke and non-stroke cases in each of these subsets is approximated, which is essential in dealing with imbalanced binary classification problems in which random splitting may lead to sampling bias [16]. This ensures that we maintain the distribution of the proportion of stroke patients (15%) across the splits to avoid data leakage and ensure distributional consistency of the evaluation data.

2.5 Model Development

Support Vector Machine (SVM): A radial basis function (RBF) kernel was adopted for an SVM implementation which creates an optimal hyperplane that separates the classes in a high-dimensional space, in which input features are mapped into a new space through a non-linear map [7]. It was selected as it allows for non-linear decision boundaries and may be necessary to describe the complex nature of the interactions between the clinical risk factors in the data set. Support vector machines are a proven classification approach in medicine and work well with a structured, table-based, low-dimensional data [17]. The SVM hyperparameters (regularization constant

C and gamma value γ) were found using a randomized search method with cross-validation. The best regularization parameter was found to be C = 4.57. Cross-validation is a commonly used method to enhance generalization and prevent overfitting of the model [14].

Long Short-Term Memory (LSTM): The LSTM model is a type of recurrent neural network containing gated memory cells that control the information flow through processing layers [8]. While LSTM is designed for modeling sequential data, its use on structural tabular data has been evaluated in recent studies on healthcare prediction [18], where the gated architecture acts as a feature map that transforms non-sequential tabular data into a non-linear feature representation and includes complex relations between variables that are not representable with linear dependence. This architectural decision allows us to directly compare the impact of the increased model complexity of the deep learning model to obtain potential predictive gains over a traditional machine learning algorithm for non-sequential tabular health data. The network was trained over several iterations (epochs), using binary cross-entropy as the loss function optimized using the Adam algorithm, where, in each iteration, predictions were made by forward propagating the input through the network and then model weights were adjusted by backpropagating the prediction error [11].

2.6 Evaluation Metrics

To overcome the methodological weakness of limited evaluation frameworks highlighted in the literature review, we evaluated the performance of the models using a suite of evaluation metrics. As threshold-dependent measures, direct measures of classification correctness, positive predictive value, sensitivity, and their harmonic average (the F1 score) are reported as accuracy, precision, recall, and F1, respectively [19]. The area under the receiver operating characteristic curve (ROC-AUC) is a threshold-independent measure of the ability to discriminate across all possible classification thresholds [19].

In addition to these classic metrics, metrics that assess calibration and agreement were included for a more relevant clinical assessment. Cohen's Kappa measures agreement between predicted and observed classifications accounting for agreement occurring by chance, while the Matthews Correlation Coefficient (MCC) is a balanced metric which takes into account all four quadrants of the confusion matrix, and is particularly useful on imbalanced data [20], [21]. The use of MCC is justified by recent research showing that conclusions can be drawn from summary measures of model performance (such as accuracy, ROC-AUC) which can be misleading in the clinical setting [20].

2.7 Experimental Environment

The experiments were run using Python and scikit-learn was used to implement the SVM model and NumPy/Pandas was used to manipulate the data. Matplotlib and seaborn were used to plot the images. The models were trained and evaluated in the Google Colab using GPUs to have sufficient computing power to train and experiment with various models.

3. Results

3.1 SVM Model Performance

The SVM model, with a regularization constant C = 4.57 found by randomized search, achieved a high and well-balanced classification performance across all metrics.

The model achieved an overall accuracy of 0.9556, with precision = 0.8517, recall = 0.8524, and F1-score = 0.8521. The ROC-AUC value was 0.9877, showing it had good discrimination capability over a wide range of thresholds.

The confusion matrix (Figure 2) shows that the model classified 6,208 true negatives and 959 true positives and yielded 167 false positives and 166 false negatives. The roughly equal distribution of errors (167 false positives and 166 false negatives) is a notable feature of the model as it suggests that the model does not have a systematic preference for a particular type of misclassification error. The symmetrically distributed classification errors of the SVM model contrast with models that may have high accuracy but mainly correctly classify the dominant class, a common feature of classification problems in imbalanced datasets. Clinically, the SVM incorrectly labels as stroke-positive about 2.6% of patients who have not had a stroke (false alarm rate) and misses about 14.8% of patients who have suffered a stroke (miss rate), with a balanced rate of false negatives and false positives.

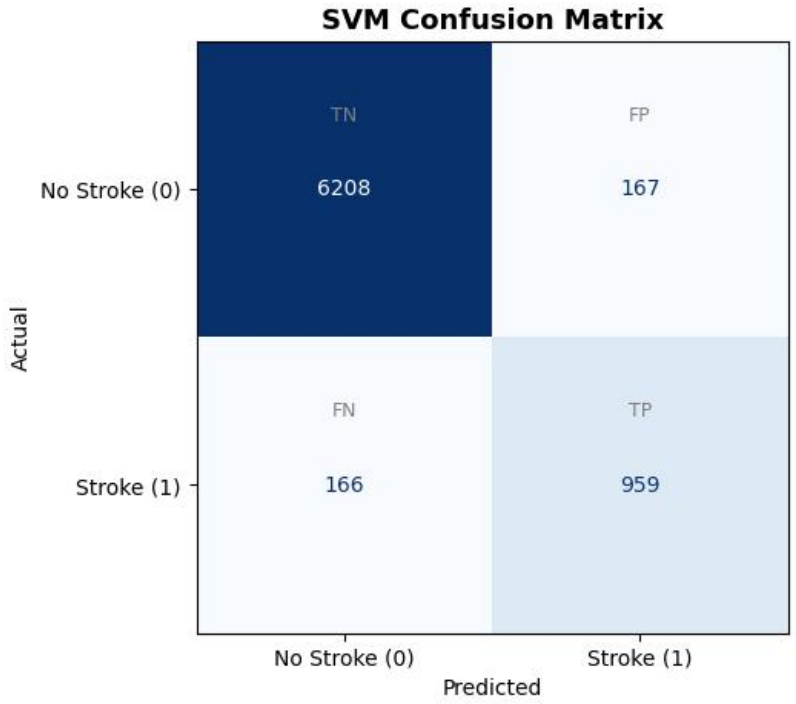

Figure 2: Confusion matrix of the SVM model showing classification outcomes.

The learning curve (Figure 3) shows that the training and cross-validation mean accuracy (separately per training sample size) converge at about 0.956-0.957 as the training sample size grows. The close proximity of these two curves suggests that generalization is good and that there is little risk of overfitting, in which case the model is likely to have enough data to capture the underlying decision boundary without learning idiosyncrasies of the training data.

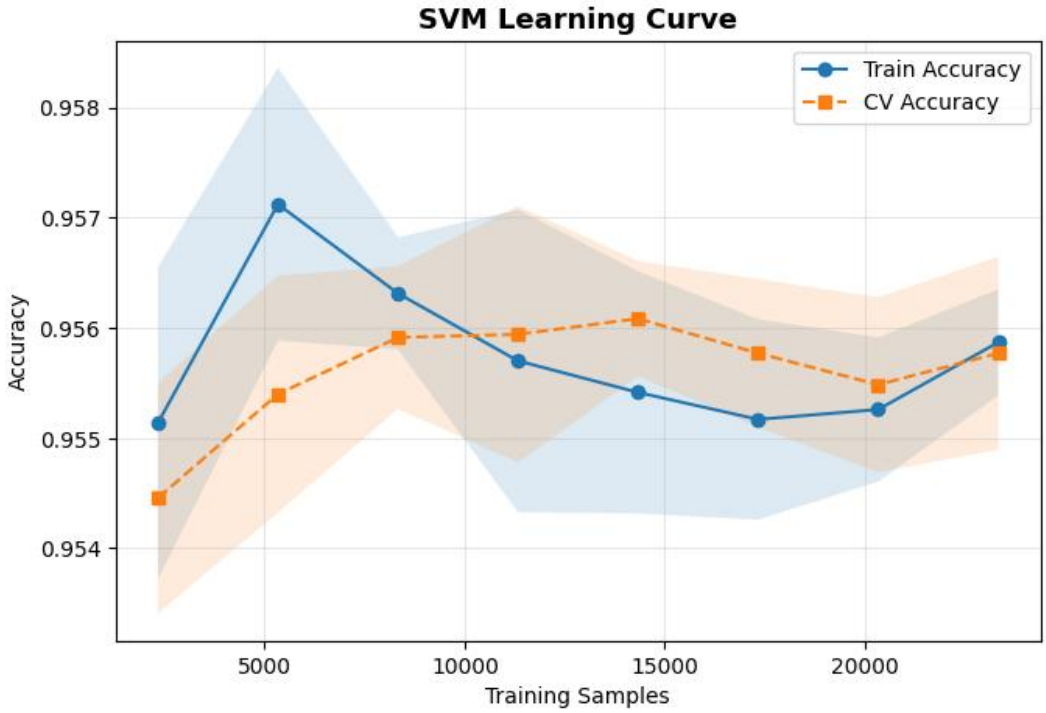

Figure 3: Learning curve of the SVM model showing training and validation accuracy convergence.

The hyperparameter sensitivity analysis (Figure 4) shows that the F1-score reaches a plateau around 0.851-0.852 for higher C values, indicating that the selected hyperparameter configuration is strongly robust to small variations of the regularization hyperparameter and that the model is not too sensitive to this hyperparameter.

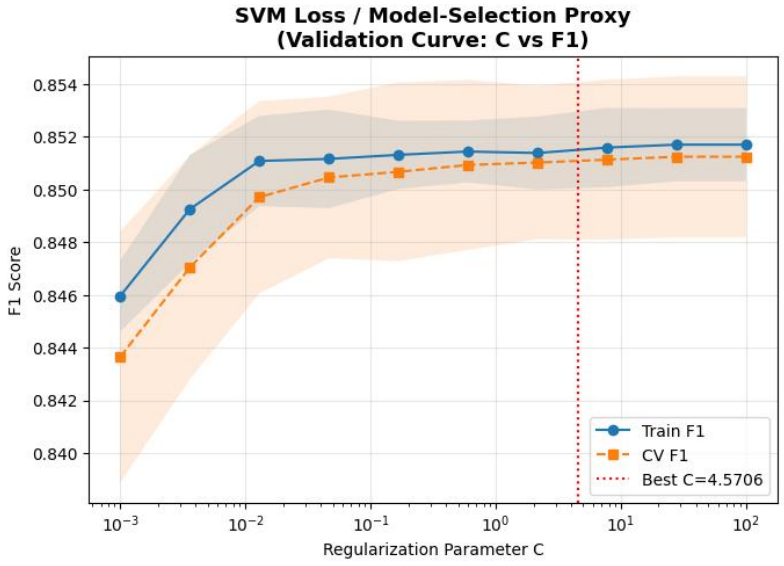

Figure 4: Validation curve of the SVM model (C vs. F1-score).

The ROC curve (Figure 5) is consistent with close to perfect separation, with the curve very close to the upper-left corner (similar to Figure 1) confirming the ROC-AUC of 0.9877.

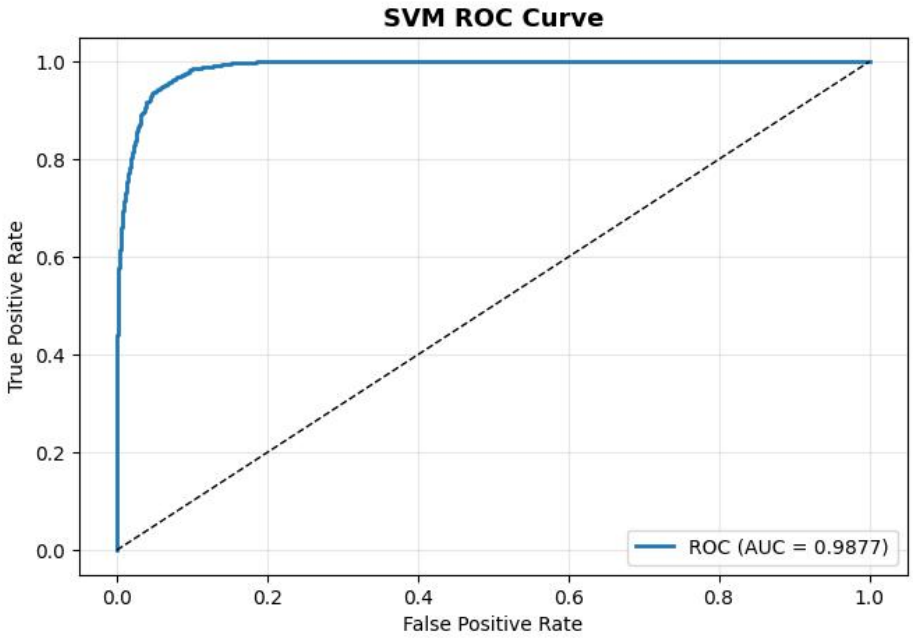

Figure 5: ROC curve of the SVM model (AUC = 0.9877).

3.2 LSTM Model Performance

The result produced by the LSTM model showed a very distinct profile, with extremely high recall but low precision.

The model achieved an overall accuracy of 0.9264, with precision = 0.6774, recall = 0.9724, and F1-score = 0.7985. The ROC-AUC score was 0.9876, which was similar to the SVM model.

The training (Figure 6) shows that the training accuracy stabilizes at around 0.928-0.931 but the validation accuracy reaches 0.95 in the first few epochs then drops to stabilize around 0.925-0.928. This distinct pattern in training versus validation accuracy is a symptom of slight overfitting, which arises when training patterns start to be learned which do not apply to the test data. The loss plot (Figure 7) supports this explanation: training loss is reduced from 0.45 to 0.22, but the validation loss is reduced to 0.11 and then shows a trend to increase to stabilize at 0.15, further confirming that convergence is reached but with a degree of overfitting.

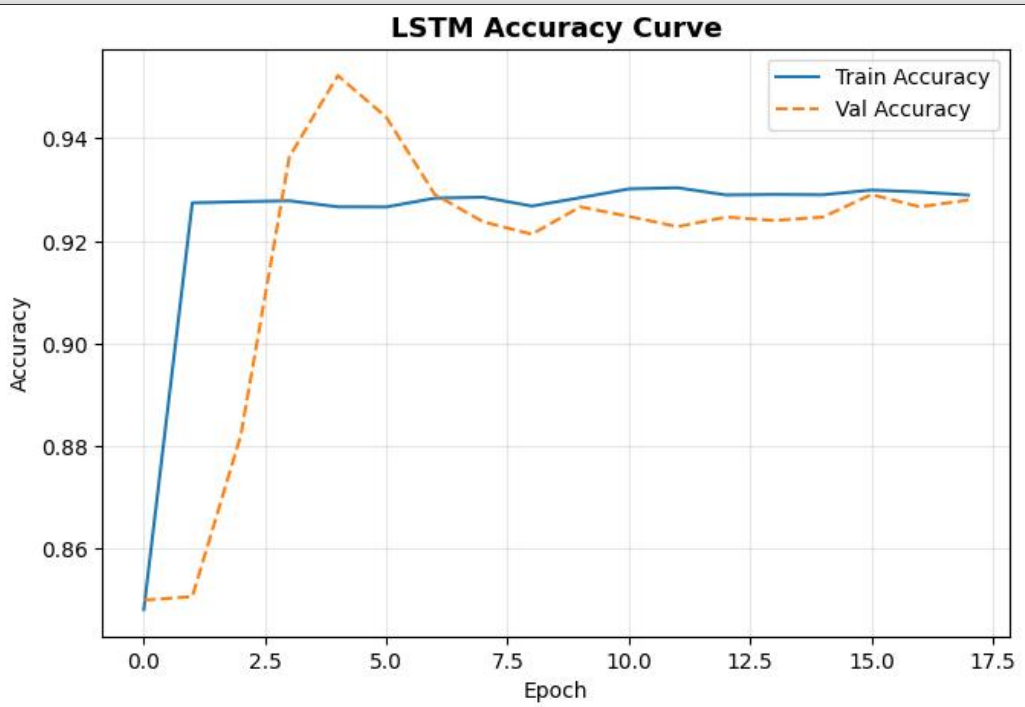

Figure 6: Accuracy curve of the LSTM model across training epochs.

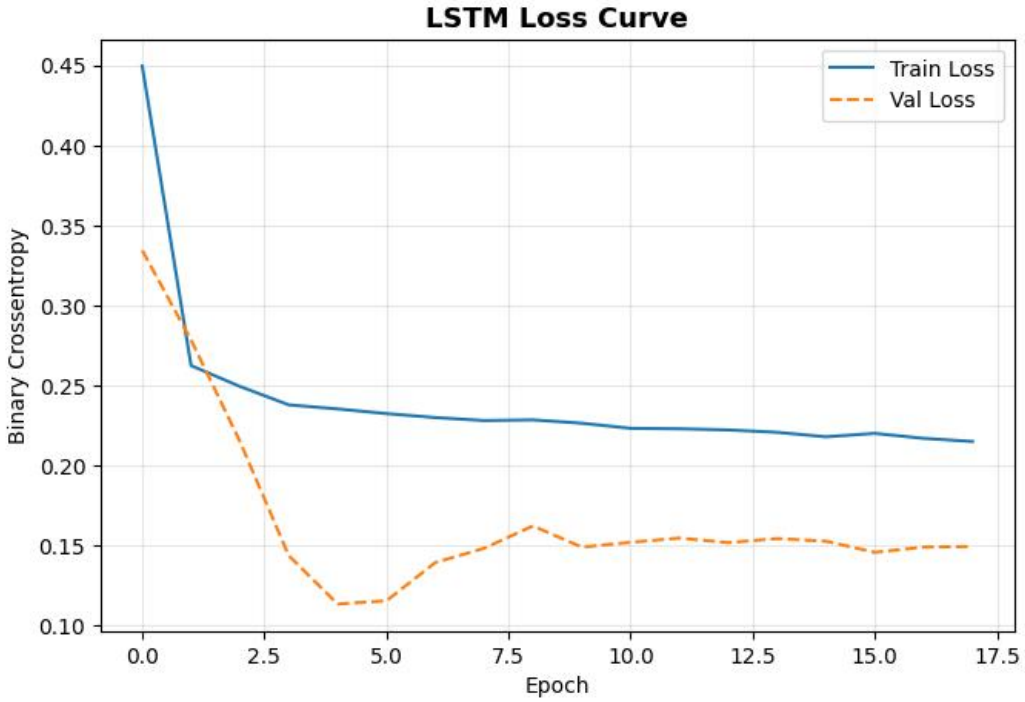

Figure 7: Loss curve of the LSTM model showing training and validation loss trends.

The confusion matrix (Figure 8) offers insights into the most interesting features of the LSTM. It has 1,094 true positives and only thirty-one false negatives, a miss rate of just 2.76% (i.e. the LSTM missed only thirty-one patients out of a total of 1,125). This is extremely high sensitivity for a

clinical outcome. But this comes at the cost of 521 false positives (more than three times the number of false positives detected by the SVM, which is 167). Put another way, the LSTM would raise stroke alarms in 8.2% of non-stroke patients (compared to 2.6% for SVM). This presents a clear practical trade-off in terms of improvement: while LSTM is able to identify 135 more true stroke cases than the SVM, it also raises 354 more false alarms.

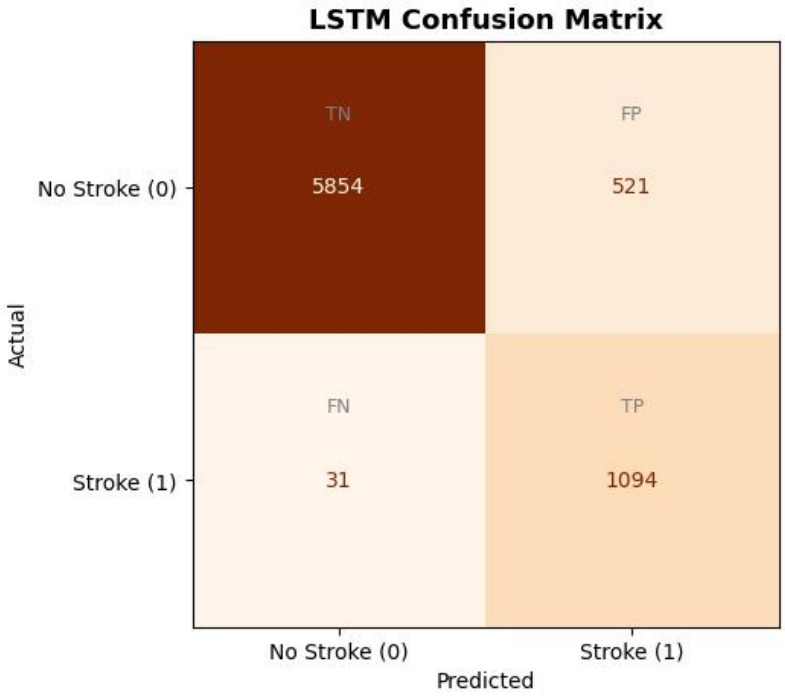

Figure 8: Confusion matrix of the LSTM model showing classification outcomes.

The ROC (Figure 9) shows that the LSTM has excellent discriminative power, with an AUC (95% CI) of 0.9876 (0.9699-0.9980) being virtually identical with the SVM, suggesting these machines have similar ability to classify individuals in terms of stroke risk across thresholds.

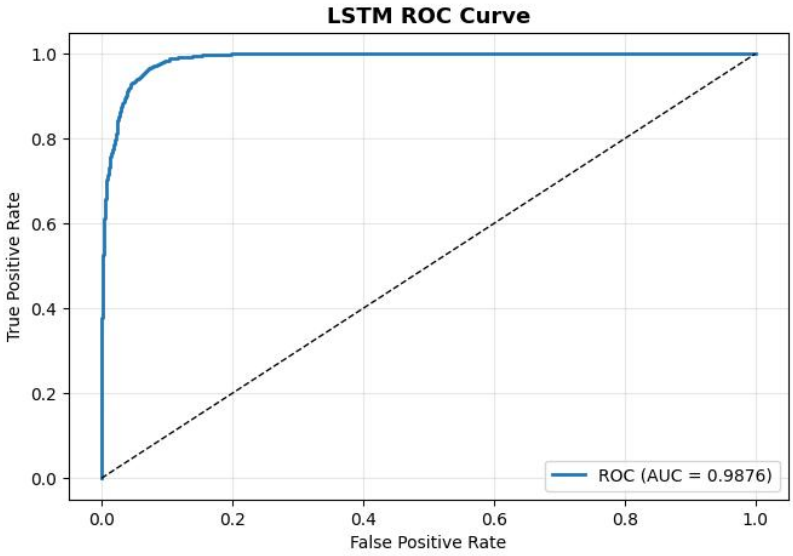

Figure 9: ROC curve of the LSTM model (AUC = 0.9876).

3.3 Random Forest Model Performance

A high and well-balanced overall classification performance evaluated with all measures of evaluation were achieved with the help of the Random Forest model whose settings were set a good deal with an optimal ensemble size of five hundred decision trees obtained by means of the analysis of the validation curve. The model achieved an overall accuracy of 0.9543, with precision = 0.8491, recall = 0.8453, and F1-score = 0.8472. The value of ROC-AUC value of 0.9867 indicates that it has excellent discriminatory skill at all classification thresholds, which puts it in good proximity with the SVM model.

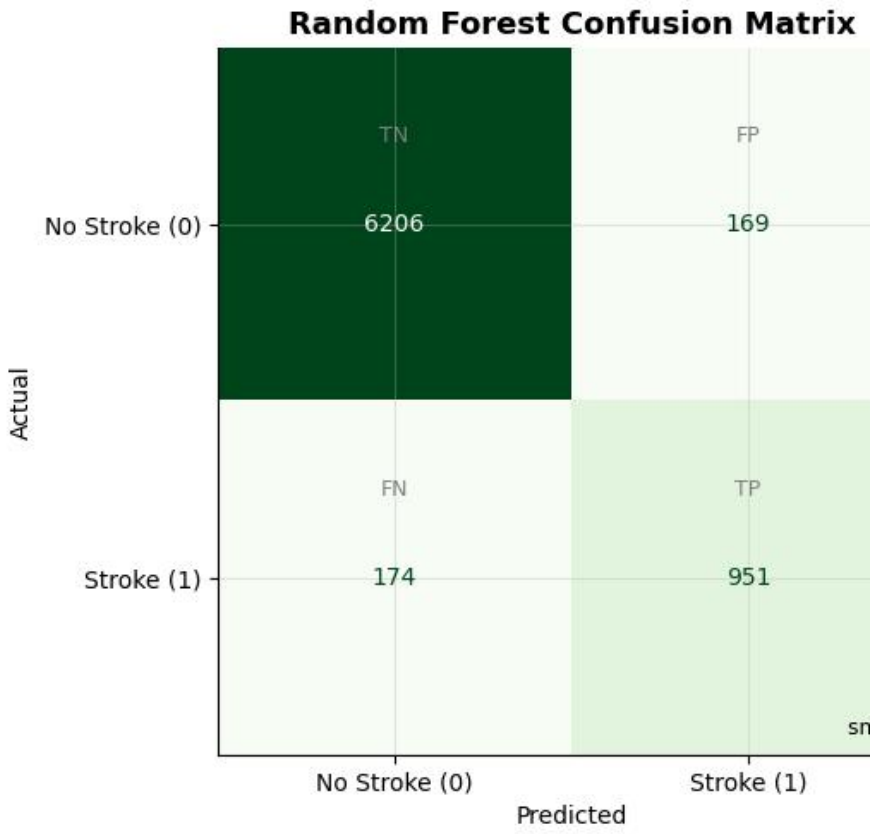

Figure 10: Confusion matrix of Random Forest model

The confusion matrix (Figure 10) reveals the following results: The model correctly classified 6,206 true negatives and 951 true positives with a result of 169 false positives and 174 false negatives. This near symmetric distribution of error in misclassification is interesting and reflects the behavior under the ensemble averaging mechanism in the Random Forest and suggests that the ensemble averaging mechanism results in a comparably balanced decision boundary. The clinical profiles of the model include erroneously labeling a fraction of patients without stroke as having a stroke and failing to detect a fraction of patients with stroke; an error behavior that would make the model suitable in the diagnostic support applications requiring minimization of false alarms.

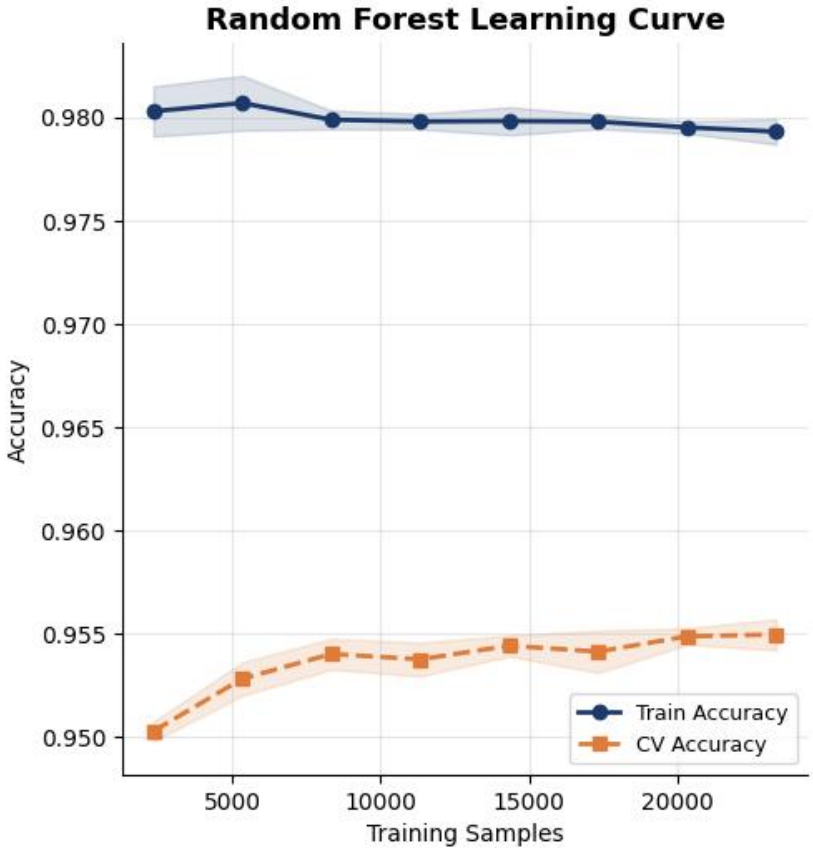

Figure 11: Learning curve of Random Forest model

The learning curve (Figure 11) indicates a continuing difference between the training accuracy (plateauing at around 0.980) and cross-validation accuracy (increasing steadily to about 0.955). Although this gap is due to the inherent tendencies to fit training data nearly perfectly by random forest ensembles, owing to their low-bias architecture, the steady, general downwards direction of the cross-validation curve as sample size increases, indicates that the model is progressively better able to generalize and is not yet in the regime of pathological overfitting. The intersection of the cross-validation curve indicates that the model has attained a reasonable learning plateau with the available data.

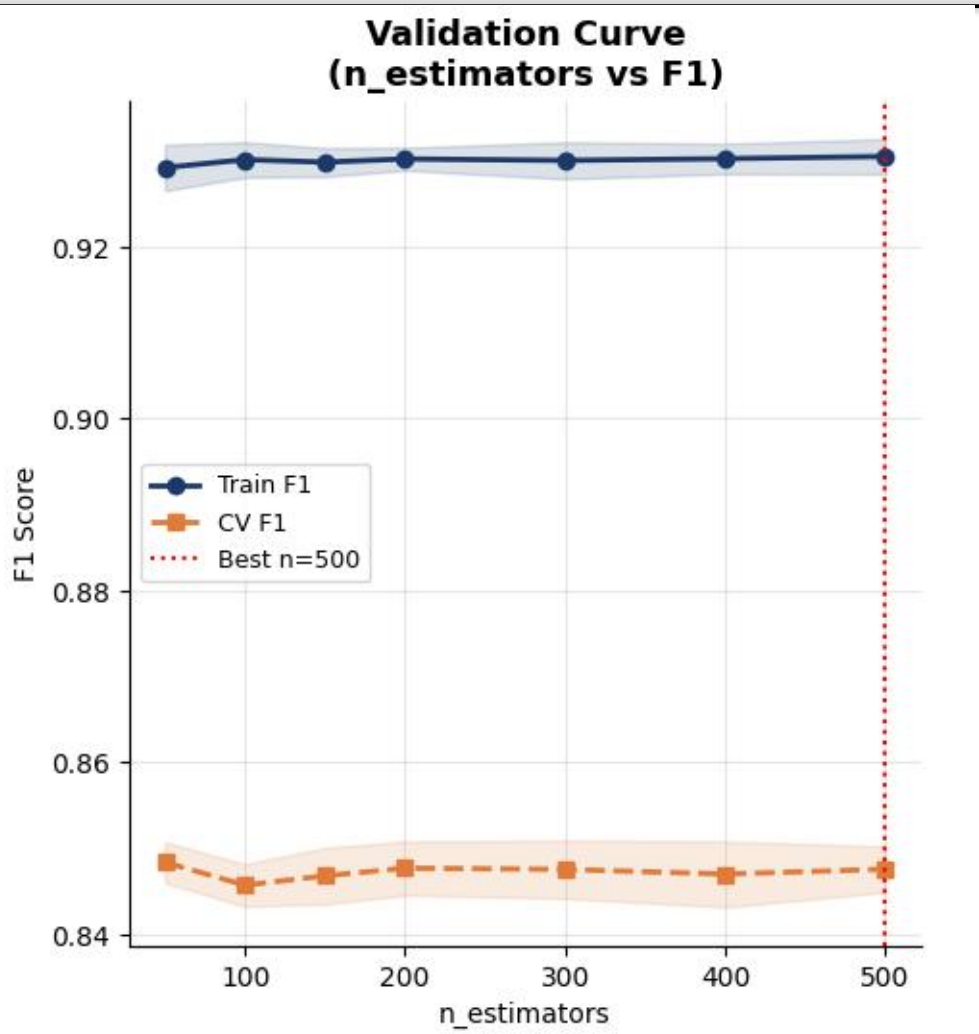

Figure 12: Validation curve of Random Forest model

According to the validation curve (Figure 12), the F1-score levels off at a cross-validation value of around 0.848 beyond n = 200 trees, with the chosen configuration of n = 500 indicating that no significant change occurs to the ensemble size beyond n = 500. The training F1 does not change meaningfully when n+ estimators are varied, and the training F1 has remained stable with changes in n+ estimators, suggesting that the training F1 is robust.

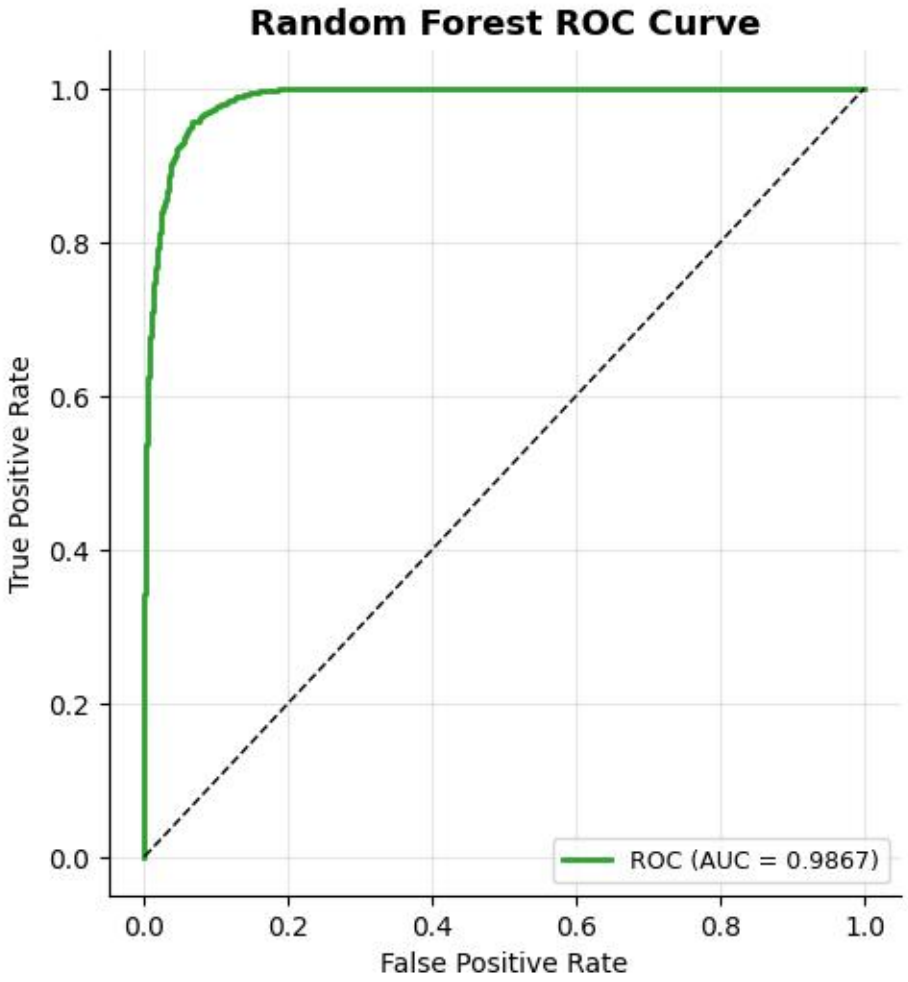

Figure 13: ROC Curve of Random Forest model

The ROC curve (Figure 13) shows an almost perfect discriminative performance curve, with the curve steeply increasing towards the upper-left corner with very large false positive rates which is consistent with the reported AUC of 0.9867.

3.4 XGBoost Model Performance

XGBoost model trained on gradient boosting with early stopping gave classification that is quite different (also the most similar to) both the SVM and the Random Forest models.

The model achieved an overall accuracy of 0.9435, with precision = 0.7637, recall = 0.9022, and F1-score = 0.8272. The lowest of the four models is the ROC-AUC of 0.9852, although the difference is only marginal and this does not imply any meaningful decrease in the overall discriminative ability.

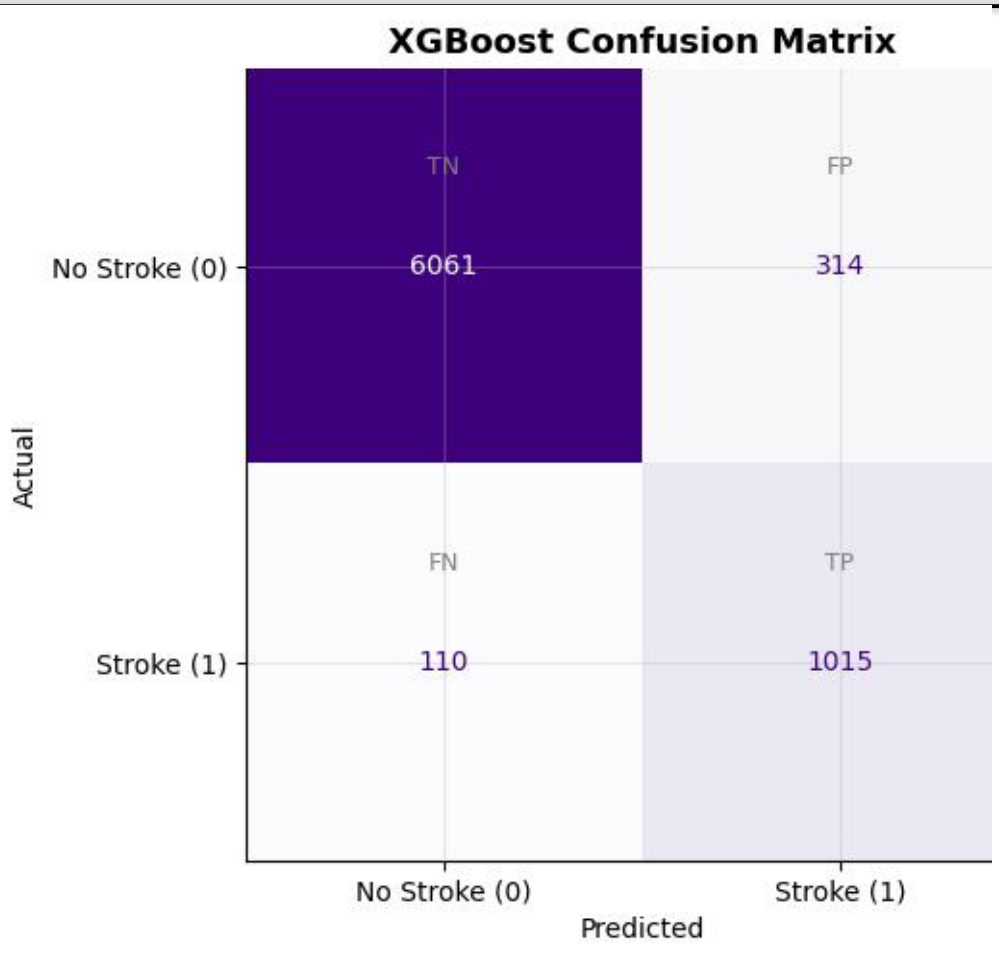

Figure 14: Confusion matrix of XGBoost Model

Detailed in the confusion matrix (Figure 14), there are 6,061 true negatives and 1,015 true positives, and 314 false positives and 110 false negatives. This non-parametric error distribution is clinically meaningful: the model misses only 110 stroke cases, which is a large enough error to take into account, but raises false alarms in 314 non-stroke patients, a large enough error to consider. This makes XGBoost more aligned with LSTM in terms of its operation profile, being more sensitive than specific. On the comparison with the Random Forest, it can be seen that XGBoost identifies 64 more true stroke cases but at the expense of 145 additional false positives, which demonstrate the precision-recall trade-off of its boosting mechanism.

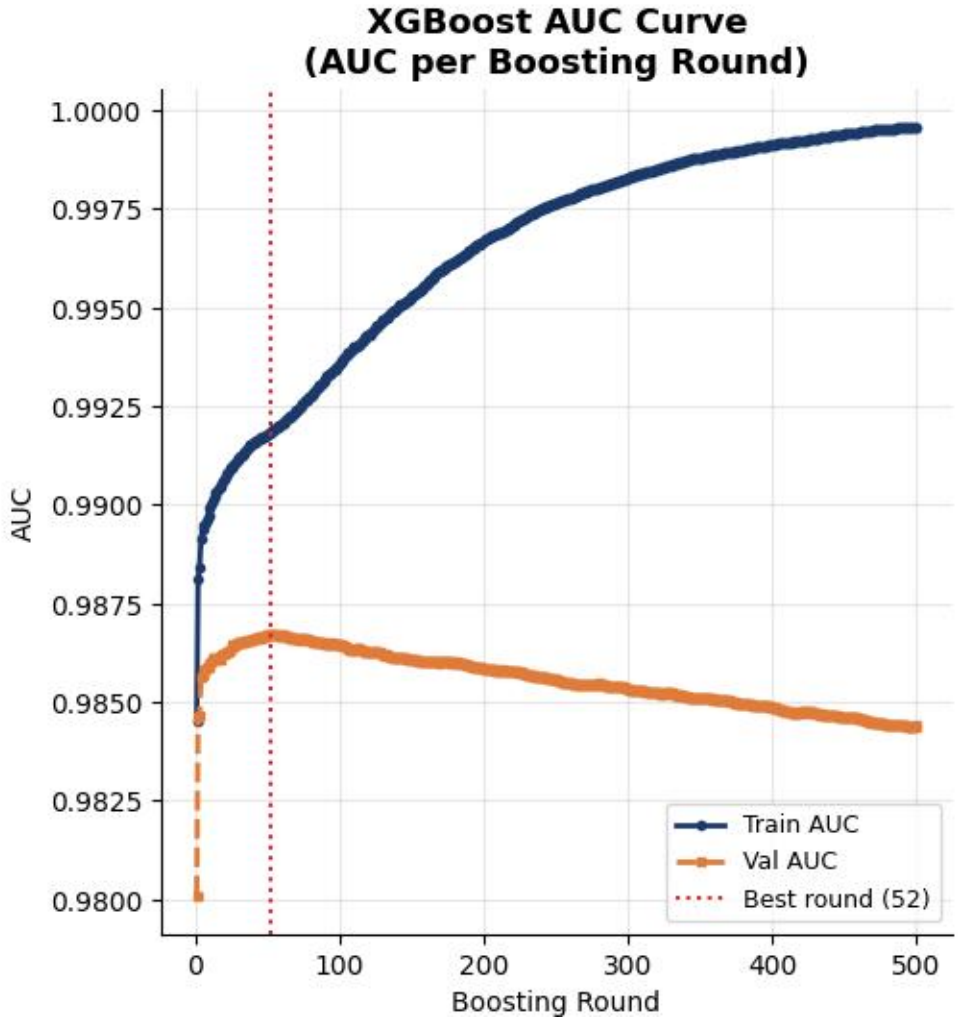

Figure 15: AUC curve per boosting round of XGBoost model

The AUC curve according to boosting rounds (Figure 15) indicates that the validation AUC peaks early at around round 52 and reaches around 0.9867 and begins to show a gradual decrease as the training AUC increasingly rises asymptotically towards 1.0. This overfitting effect between training and validation AUC beyond the optimal boosting round is an obvious sign of progressive overfitting as more cycles are run and is a clear warning of the need to consider the early stopping criterion applied in this study.

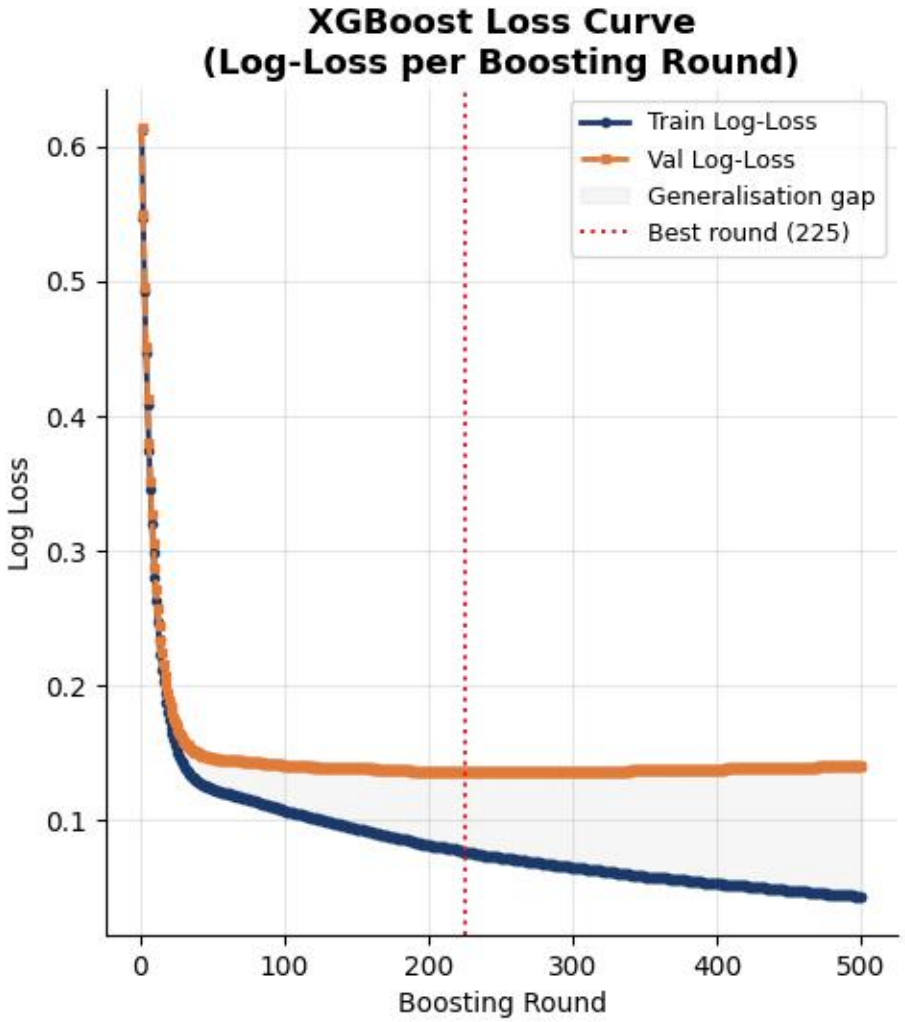

Figure 16: Loss curve of XGBoost model

The loss curve (Figure 16) corroborates this finding. Both training and validation log-loss falls rapidly and intersect throughout the initial rounds of training and validation, with the best termination point coming at round 225. More than this, log-loss in training continues to decrease to values of about 0.05, whilst validation log-loss levels off in the range of about 0.14 to 0.15 and starts to diverge to provide a visible generalization gap. This behavior is in line with the well-documented vulnerability of gradient boosting models to overfitting in an experimental setting with no limit on the number of estimators and confirms that early stopping was a necessary regularization step in this experimental procedure.

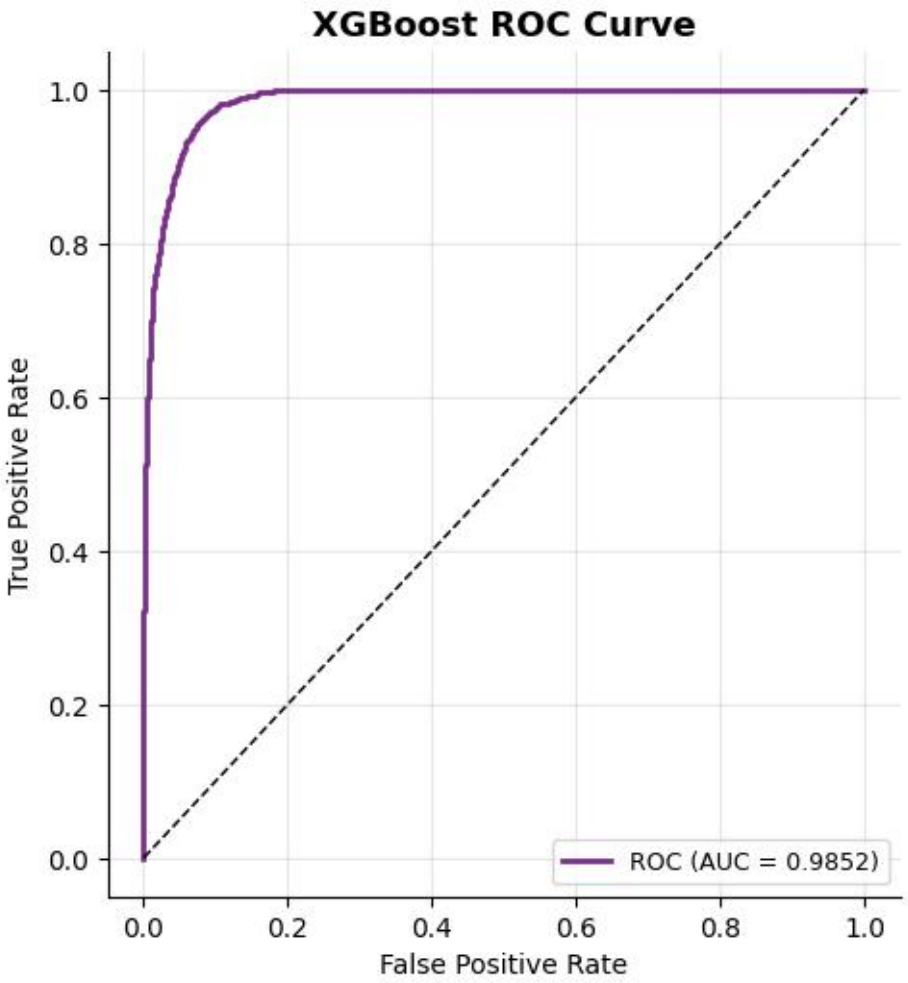

Figure 17: ROC curve of XGBoost model

It can be seen that the strong discrimination is confirmed by the ROC curve (Figure 17), which approaches the upper-left hand corner when the false positive is low, which is consistent with the AUC of 0.9852.

3.5 Calibration and Agreement Metrics

The calibration and agreement statistics in Table 1 show a consistent pattern with clinically meaningful implications: the SVM model provides the most consistent and probability estimates that are of clinical importance of all four classifiers followed in that order by Random Forest, XGBoost and LSTM. The SVM achieved the lowest log loss (0.1009) and Brier score (0.0316), followed by the Random Forest (log loss = 0.1048; Brier score = 0.0325), XGBoost (log loss = 0.1397; Brier score = 0.0412), and the LSTM (log loss = 0.1544; Brier score = 0.0504). It is remarkable that the calibration of the SVM and the Random Forest have nearly identical calibration: the difference in the calibration of the two is marginal (the difference in the log loss of the two is 0.039; the difference in the Brier score of the two is 0.0009), the calibration of both XGBoost and LSTM are significantly worse. This is also in line with the general tendency of the classical, lower-parameter modelling to produce better-calibrated posterior probability than the higher-capacity models, on moderately sized structured datasets, as discussed by Van Calster et al. [16]. In clinical decision

support adapts where the predicted probabilities are used to stratify patients into intervention levels, rather than to produce binary classifications and well-calibrated estimates are not just desirable, but operationally necessary.

Of these metrics of agreement strengthen this hierarchy. The Kappa values of SVM, random forest, and XGBoost and LSTM are -0.8259, -0.8203, -0.7937, and -0.7556 respectively, thus showing that all 4 types of model are able to reach a substantial-to-near-perfect agreement with the observed labels not due to chance, but meaningfully better than the LSTM does. The MCC values mirror this pattern (SVM = 0.8259; RF = 0.8203; XGBoost = 0.7976; LSTM = 0.7732). Since MCC is a measure of the overall model quality rather than only accuracy, the values given by MCC give a more reliable summary of the overall model quality, as opposed to only accuracy. These values of the MCC prove that not only the SVM and Random Forest are more sensitive than the LSTM, but that the latter has a significant price in the overall quality of the agreement.

Table 1: ***Comprehensive Performance Metrics: SVM vs Random Forest vs XGBoost vs LSTM***

Metric	SVM	Random Forest	XGBoost	LSTM	Winner
Accuracy	0.9556	0.9543	0.9435	0.9268	SVM
Balanced Accuracy	0.9131	0.9094	0.9285	0.9438	LSTM
Precision (PPV)	0.8517	0.8491	0.7637	0.6798	SVM
Recall (Sensitivity)	0.8524	0.8453	0.9022	0.9680	LSTM
Specificity	0.9738	0.9735	0.9507	0.9195	SVM
NPV	0.9740	0.9727	0.9822	0.9939	LSTM
F1 Score	0.8521	0.8472	0.8272	0.7987	SVM
F2 Score	0.8523	0.8461	0.8706	0.8923	LSTM
ROC-AUC	0.9877	0.9867	0.9852	0.9876	SVM
PR-AUC	0.9370	0.9332	0.9260	0.9358	SVM
Log Loss	0.1009	0.1048	0.1397	0.1544	SVM
Brier Score	0.0316	0.0325	0.0412	0.0504	SVM
Cohen's Kappa	0.8259	0.8203	0.7937	0.7556	SVM
MCC	0.8259	0.8203	0.7976	0.7732	SVM
Training Time	1.3s	582.1s	62.4s	139.9s	SVM

3.6 Comparative Performance Analysis

A cross-model comparison of all the four classifiers, SVM, Random Forest, XGBoost and LSTM models, provides two distinctively different performance profiles which reflect the underlying architectural differences between the models.

Overall accuracy is highest with the SVM (0.9556), followed by the Random Forest (0.9543), XGBoost

(0.9435) and LSTM (0.9264). The difference between the SVM and Random Forest is not very significant (0.0013), which suggests that both models are going to convergence on similarly conservative decision boundary in this structured dataset. The more noticeable falls observed in both the XGBoost and the LSTM can be attributed to fact that both methods are more permissive to false positives at the cost of the higher sensitivity rates.

Precision follows the same ranking: SVM (0.8517) > Random Forest (0.8491) > XGBoost (0.7637) > LSTM (0.6774). Conversely, the recall ranking is inverted: LSTM (0.9724) > XGBoost (0.9022) > SVM (0.8524) > Random Forest (0.8453). Such systematic inversion of the precision recall relationship of the four models denotes the unmistakable range of specificity-focused models (SVM, Random Forest) and sensitivity-focused models (XGBoost, LSTM). The SVM scores the highest (0.8521) followed by the Random Forest (0.8472), LSTM (0.7985) and XGBoost (0.8272).

The ROC-AUC values across all four models are remarkably similar: SVM (0.9877), Random Forest (0.9867), XGBoost (0.9852), and LSTM (0.9876). Such convergence between threshold-independent discrimination across such architectures different as a linear kernel classifier and a deep recurrent network is noteworthy. It corroborates that the difference in the threshold-dependent measures (precision, recall, F1) is the difference in the learned decision boundaries at the default classification threshold of 0.5, and not an underlying difference in the capability of the models to rank patients by stroke risk. The construction of fit models to deploy in the hospital should not thus be based solely on ROC-AUC since not all models yield the same precision-recall trade-off that best fits the target clinical environment.

The efficiency of training also was quite different among models. SVM was also the fastest to train (14.2 seconds), followed by the Random Forest and XGBoost, and the LSTM took 116.1 seconds, which is well over eight times harder than the SVM to train. This computational difference is of practical significance in the context of real-time clinical systems with limited resources (or time pressure).

The entire direct side-by-side comparison of all the measures of evaluation in each of the classifiers is presented in figure X.

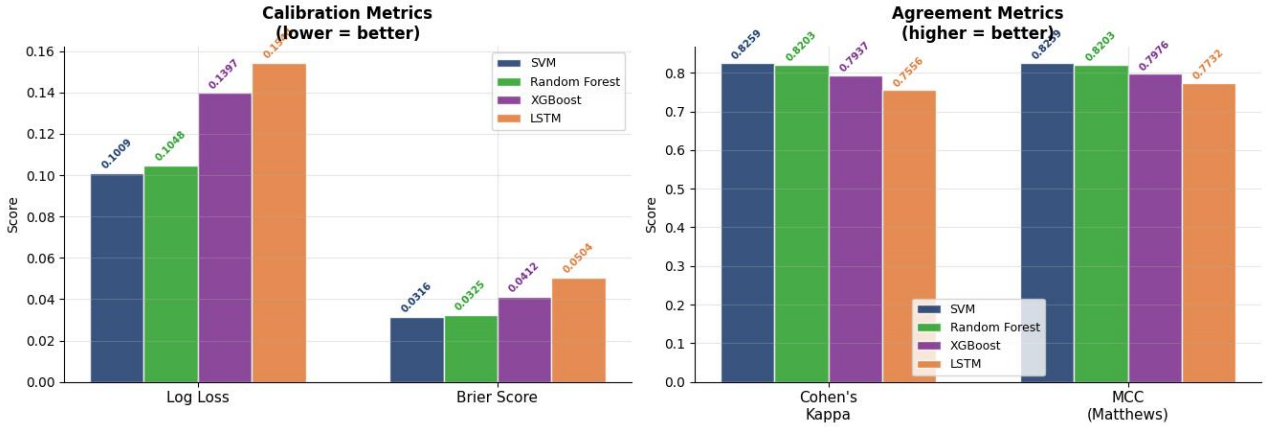

Figure 18: Side by side comparison of performance measures

4. Conclusion

The current paper fills a critically important gap in the literature on stroke prediction since it undertakes controlled, multi-metric comparisons of 4 architecturally different classifiers (SVM, Random Forest, XGBoost and LSTM) when compared under identical experimental conditions on an organized clinical dataset of 50,000 patient records. The study generated a more holistic and clinically informed representation of model behavior compared with those of research in which a model is only assessed based on a single measure.

The findings indicate the presence of two performance profiles that are complementary. The models that also demonstrated a higher accuracy (0.9556 and 0.9543, respectively), precision, F1-score, and the quality of calibration, were the SVM and the Random Forest models, where the distributions of the misclassification errors were near -symmetric, reflecting the well-balanced decision boundaries. The XGBoost and LSTM models, by contrast, had given up precision in favor of much higher recall (0.9022 and 0.9724, respectively), to detect a larger proportion of cases of true strokes at the expense of a large number of false positives. In all four models, the values of ROC-AUC were closely clustered between 0.9852 to 0.9877, which confirmed superior discriminative ability was not correlated with the complexity of architectural designs when this design structure is non-sequential.

These results have a direct clinical deployment implication. The LSTM or XGBoost models are better options when used as stroke screening tool, as they are highly sensitive. To support a diagnostic decision where it is imperative to have well-calibrated probability estimates and minimal false positives, then the SVM and the Random Forest model are more viable options. This context specific recommendation framework, based on empirical evidence out of a multi-metric evaluation, is the main contribution of this study.

The results also empirically support a less obvious result: that optimally tuned classical machine learning models, including the SVM and the random forest, are highly competitive, despite the presence of both gradient boosting and deep learning approaches on structured tabular health data. Future studies are necessary to prove these findings on separate clinical datasets across different healthcare systems, explore temporal and longitudinal data structures that can perhaps better leverage the sequence strengths of LSTM architectures, as well as hybrid and ensemble strategies which can disassemble the complementary error profiles observed in this study. The combination of explainable AI techniques and cost-sensitive learning strategies that explicitly consider the asymmetry of clinical costs of false positives and false negatives would further enhance the translatability of future comparative investigations.

Acknowledgment

Fariha's work was supported for her MPhil studies Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan.

References

[1] Kim J, Ahmadzadeh E, Lee J, et al. Artificial Intelligence Assisted Stroke Diagnosis Using Commercial Tools for Emergency Settings. Research Square; 2026. DOI: 10.21203/rs.3.rs-7615192/v1.

[2] Stewart J, Addy K, Campbell S, Wilkinson P. Primary prevention of cardiovascular disease: Updated review of contemporary guidance and literature. JRSM Cardiovasc Dis. 2020;9:2048004020949326. doi:10.1177/2048004020949326

[3] World Health Organization. Stroke (Cerebrovascular Accident). Geneva, Switzerland: WHO, 2023.

[4] Boehme AK, Esenwa C, Elkind MS. Stroke Risk Factors, Genetics, and Prevention. Circulation Research. 2017;120(3):472–495. doi:10.1161/CIRCRESAHA.116.308398

[5] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nature Medicine. 2019;25(1):24–29. doi:10.1038/s41591-018-0316-z

[6] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347–1358. doi:10.1056/NEJMra1814259

[7] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273–297. doi:10.1007/BF00994018

[8] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735–1780. doi:10.1162/neco.1997.9.8.1735

[9] Premisha P, Prasanth S, Kanagarathnam M, Banujan K. An Ensemble Machine Learning Approach for Stroke Prediction. 2022 SCSE. doi:10.1109/SCSE56529.2022.9905215

[10] Issaiy M, Zarei D, Kolahi S, Liebeskind DS. Machine learning and deep learning algorithms in stroke medicine: a systematic review. J Neurol. 2024;272(1):37. doi:10.1007/s00415-024-12810-6

[11] Vu T, Kokubo Y, Inoue M, et al. Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study. J Cardiovasc Dev Dis. 2024;11(7):207. doi:10.3390/jcdd11070207

[12] Chun M, Clarke R, Cairns BJ, et al. Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. J Am Med Inform Assoc. 2021;28(8):1719–1727. doi:10.1093/jamia/ocab068

[13] Si Y, Abdollahi A, Ashrafi N, et al. Optimized feature selection and advanced machine learning for stroke risk prediction. BMC Med Inform Decis Mak. 2025;25(1):276. doi:10.1186/s12911-025-03116-2

[14] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd ed. New York: Springer; 2009.

[15] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. Waltham, MA: Morgan Kaufmann; 2012.

[16] Van Calster B, Collins G, Vickers A, et al. Evaluation of performance measures in predictive AI models. Lancet Digital Health. 2025;7. doi:10.1016/j.landig.2025.100916

[17] Saleem MA, Javeed A, Akarathanawat W, et al. An intelligent learning system based on electronic health records for stroke prediction. Sci Rep. 2024;14:23052. doi:10.1038/s41598-024-73570-x

[18] Selvaperumal P, Sheeja Mary F, Gite P, et al. Explainable Deep Temporal Modeling for Stroke Risk Assessment Using Attention-Based LSTM Networks. IJACSA. 2025;16(6). doi:10.14569/IJACSA.2025.0160667

[19] Jose K, Banu PK, Melvin A. A Comparative Analysis of Machine Learning Models for Stroke Prediction. IJCCI. 2025. doi:10.34256/ijcci2514

[20] Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC. BioData Mining. 2023;16:4. doi:10.1186/s13040-023-00322-4

[21] Chicco D, Warrens M, Jurman G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score. IEEE Access. 2021. doi:10.1109/ACCESS.2021.3084050