Proceedings of the 38th iSTEAMS Bespoke Conference – Accra Ghana 2024

Society for Multidisciplinary & Advanced Research Techniques (SMART)
West Midlands Open University – Projects, Research, Innovations, Strategies & Multimedia (PRISM) Centre
SMART Scientific Projects & Research Consortium (SMART SPaRC)
Sekinah-Hope Foundation for Female STEM Education
ICT University Foundations USA
Harmath Global Educational Services

38th International Science Technology Education Arts Management
& Social Sciences (iSTEAMS) Bespoke Conference - Accra Ghana 2024

An Evaluation of Machine Learning Algorithms for an Enhanced Precision Healthcare in Stroke Prediction

*[1]Fatimah Adamu-Fika, [2]Deborah Ifeoluwa Ayeku, [3]Tsentob Joy Samson,
[4]Aisha Tijjani Ramalan, [5]Aanuoluwapo Enyojo Baba-Onoja, [6]Oluwaseyi Ezekiel Olorunshola,
& [7]Henry Onyeoma Mafua
[1]Department of Cyber Security, Air Force Institute of Technology Kaduna
[2,3,4,5,6,7]Department of Computer Science, Air Force Institute of Technology Kaduna
*Corresponding Author: f.adamu-fika@afit.edu.ng

ABSTRACT

Stroke poses a significant health challenge worldwide and ranks among the top causes of death and prolonged disability. Early prediction is essential to mitigate adverse outcomes, but existing predictive models often face challenges such as class imbalance and limited evaluation. This study aims to provide actionable insights into improving stroke prediction by systematically evaluating and comparing the performance of three widely used machine learning algorithms—Random Forest, Logistic Regression, and Support Vector Machines (SVM). The study addresses gaps in previous research by considering multiple evaluation metrics and integrating class balancing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE). The dataset, comprising 10,091 records with demographic, clinical, and lifestyle attributes, was balanced using SMOTE. Random Forest achieved the highest performance with an accuracy of 98% and an AUC of 0.98, demonstrating its robustness and suitability for clinical integration. SVM also exhibited competitive performance, achieving an accuracy of 96% and an AUC of 0.96, while Logistic Regression showed limitations in recall (88%) and AUC (0.91). The findings underscore Random Forest's potential as a reliable tool for stroke prediction and emphasise the importance of dataset balancing and comprehensive model evaluation. This study contributes to the advancement of predictive healthcare tools by providing a framework for selecting the most effective model for real-world stroke prediction applications.

Keywords: Class Balancing, Clinical Decision-Support Systems, Feature Selection, SVN, Accuracy, Machine Learning, Predictive Modelling, Random Forest, Synthetic Data

Proceedings Citation Format
Fatimah Adamu-Fika, Deborah Ifeoluwa Ayeku, Tsentob Joy Samson, Aisha Tijjani Ramalan, Aanuoluwapo Enyojo Baba-Onoja, Oluwaseyi Ezekiel Olorunshola, & Henry Onyeoma Mafua (2024): An Evaluation of Machine Learning Algorithms for an Enhanced Precision Healthcare in Stroke Prediction. Proceedings of the 38th iSTEAMS Multidisciplinary Bespoke Conference. 17th – 19th July, 2024. University of Ghana, Accra, Ghana.
Pp 318-329. dx.doi.org/10.22624/AIMS/ACCRABESPOKE2024P32

1. INTRODUCTION

Stroke is a significant global health issue, the second leading cause of death and a key factor in long-term disability. The World Health Organization (2020) estimates that stroke accounts for about 11% of all global deaths. Ischemic and haemorrhagic strokes, caused by either blocked or ruptured blood vessels in the brain, result in severe neurological damage and often require immediate intervention to prevent fatal outcomes or long-term complications. Low- and middle-income countries are significantly affected, experiencing high mortality rates due to limited access to healthcare resources (WHO, 2020; CDC, 2023).

Despite advancements in medical technologies and stroke care, existing predictive systems face numerous challenges that limit their reliability and practical application. These include imbalanced datasets, inadequate attention to feature selection, and the inability of models to integrate seamlessly into clinical workflows. Geethanjali et al. (2021) emphasised the critical role of robust predictive tools in addressing these gaps, highlighting the importance of designing systems capable of identifying high-risk individuals from routinely collected patient data. Similarly, Biswas et al. (2022) identified the limitations of traditional models, citing the need for more reliable ML approaches tailored to healthcare contexts.

Machine learning (ML) is gaining traction in predictive healthcare because it can analyse complex datasets and uncover hidden patterns. Algorithms like Random Forest, Logistic Regression, and Support Vector Machines (SVM) have been effectively used in stroke prediction, enhancing accuracy and precision. Mohammed et al. (2023) demonstrated the potential of Logistic Regression in optimising stroke prediction for resource-constrained settings. Fernandez-Lozano et al. (2021) further validated the effectiveness of Random Forest for predicting stroke outcomes, while Tazin et al. (2021) explored SVM as a robust learning approach. However, these studies often lack a comprehensive evaluation framework considering the trade-offs between precision, recall, and other critical metrics.

This study aims to build on these existing studies by providing actionable insights into improving stroke prediction through the systematic evaluation and comparison of three widely used ML models—Random Forest, Logistic Regression, and SVM. Prior studies have demonstrated the effectiveness of Random Forest, Logistic Regression, and SVM in predictive healthcare contexts (Mohammed et al., 2023; Fernandez-Lozano et al., 2021; Tazin et al., 2021).

However, these studies often focus on isolated metrics or lack a comprehensive evaluation framework that considers the trade-offs between precision, recall, and other key performance indicators. This study expands on their findings by systematically comparing these models on the same dataset under consistent preprocessing and evaluation conditions, integrating the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. The findings aim to connect technical research with practical implementation, guiding the integration of ML models into healthcare workflows.

2. Literature Review

The application of ML in healthcare, particularly for stroke prediction, has gained significant attention in recent years. This section reviews research on ML algorithms for stroke prediction, highlighting contributions and persistent challenges.
This study addresses gaps in model evaluation, class imbalance, and practical implementation, providing actionable insights to advance the field.

2.1 Machine Learning in Stroke Prediction

ML has become a powerful tool in predictive healthcare, providing advanced methodologies to analyse complex datasets and uncover patterns crucial for early diagnosis. Several studies have demonstrated the potential of ML algorithms for stroke prediction. For instance, Mohammed et al. (2023) demonstrated the effectiveness of Logistic Regression in resource-constrained settings, where its simplicity and interpretability make it an attractive choice. Similarly, Fernandez-Lozano et al. (2021) highlighted the robustness and accuracy of Random Forest, establishing its efficacy in stroke outcome prediction. Tazin et al. (2021) studied the application of SVM, highlighting their capability to handle high-dimensional datasets efficiently.

2.2 Challenges in Existing Models

Despite these advancements, several challenges remain. Many existing models are hindered by imbalanced datasets, which compromise their ability to effectively identify minority classes, such as stroke cases, in healthcare datasets (Biswas et al., 2022). Additionally, Geethanjali et al. (2021) highlighted that insufficient focus on feature selection can lead to suboptimal model performance. These issues not only limit the generalisability of ML models but also impede their integration into clinical workflows.

2.3 Addressing Gaps in the Literature

Although previous studies have evaluated individual models, many lack a systematic framework for comparing multiple algorithms under consistent conditions. For example, Mohammed et al. (2023) focused solely on Logistic Regression without comparing it to more advanced models such as Random Forest and SVM. Likewise, Fernandez-Lozano et al. (2021) concentrated only on Random Forest without exploring other algorithms or addressing class imbalance issues. This study addresses these gaps by evaluating Random Forest, Logistic Regression, and SVM on the same dataset, ensuring a fair comparison. It incorporates SMOTE to address the class imbalance. Furthermore, it uses precision, recall, F1-score, and AUC to evaluate the models comprehensively. These steps provide a robust framework for identifying the most reliable model for stroke prediction and generating actionable insights for healthcare applications.

2.4 Section Conclusion

The existing literature underscores the effectiveness of ML models such as Random Forest, Logistic Regression, and SVM in stroke prediction. However, gaps remain in comprehensive evaluations that address challenges like class imbalance and provide practical insights for clinical applications. Building on these findings, the following section outlines the methodology adopted in this study, including dataset preprocessing, model implementation, and evaluation metrics, to address these challenges systematically.

3. Methods

This section outlines the methodology for developing and evaluating the ML models for stroke prediction, covering the dataset, preprocessing, feature engineering, data splitting, model implementation, evaluation metrics, and workflow.

3.1 Dataset Description
The dataset for this study, sourced from Kaggle, includes 10,091 patient records from two combined publicly available datasets. This merging enhances the representation of patient demographics, clinical attributes, and lifestyle factors. The final dataset contains 12 features.

1. Demographics: Age, gender, residence type.
2. Clinical Factors: Hypertension, heart disease, BMI, average glucose level.
3. Lifestyle Factors: Smoking status, marital status.
4. The target variable is a binary indicator of whether a patient experienced a stroke (1) or not (0). The dataset is highly imbalanced, with only 497 stroke cases (4.92%) compared to 9,594 non-stroke cases (95.08%). This imbalance posed challenges for ML models, necessitating careful preprocessing and feature engineering to improve performance and fairness.

3.2 Data Preprocessing
Data preprocessing is essential for ensuring the dataset is clean, consistent, and prepared for ML. This step addresses key issues such as missing values, categorical variables, and feature scaling.

1. Handling Missing Values: Missing values in the BMI feature were imputed using the mean, as the distribution of BMI values was approximately normal.
2. Encoding Categorical Features: To encode categorical variables, one-hot encoding was employed for features such as smoking status and work type, generating binary indicators for each category. For gender, ordinal encoding (also known as label encoding) was used, assigning a value of 0 to female and 1 to male. The single entry categorised as "Other" was excluded to ensure consistency.
3. Normalisation: Continuous features were standardised using the Standard Scaler technique to ensure they had a mean of 0 and a standard deviation of 1. This process prevented any scale differences from biasing the models.

3.3 Feature Engineering
Feature engineering enhances the predictive power of the dataset by refining its structure and addressing inherent challenges. It involves selecting relevant features and applying techniques like SMOTE to balance the dataset, ensuring equitable representation of both classes.

1. Feature Selection: Irrelevant features, such as patient identifiers, were excluded to reduce noise and enhance interpretability.
2. Addressing Class Imbalance: SMOTE was applied to generate synthetic samples for the minority class. This approach ensured a balanced dataset, improving the ability of the models to effectively identify stroke cases.

3.4 Data Splitting and Model Implementation
Data splitting is essential for conducting an unbiased evaluation of a model's performance. In this instance, the dataset was divided into training and test sets in a 70:30 ratio. The training set was utilised for model training and hyperparameter optimisation, while the

testing set provided a means to evaluate the model's performance on unseen data without bias. Three ML models were employed to facilitate the prediction and analysis of stroke outcomes. Logistic Regression, known for its simplicity and efficiency in binary classification, was selected as a linear model. Random Forest, an ensemble of decision trees, was chosen for its robustness and capacity to improve accuracy. Support Vector Machine (SVM), recognised for its effectiveness on high-dimensional datasets, was also included as a strong classification algorithm.

3.5 Evaluation Metrics

Evaluation metrics are critical for assessing the performance and reliability of machine learning models, particularly in healthcare applications where prediction accuracy can significantly impact patient outcomes. This subsection introduces the key metrics used in this study and explains how each quantifies the model's effectiveness in predicting stroke cases.

To evaluate the models 'predictive capabilities, the following terms were defined in the context of stroke prediction:

1. True Positive (TP): A stroke case where the model accurately identified as a stroke.
2. False Positive (FP): A Non-stroke case that the model incorrectly classified as a stroke.
3. True Negative (TN): A case instance in which the model inaccurately identified a non-stroke.
4. False Negative (FN): A stroke case misclassified by the model.

Using these definitions, the following metrics were employed:

- **Accuracy**: Accuracy refers to the percentage of correctly identified stroke and non-stroke cases among all predictions made. In an imbalanced dataset, relying solely on accuracy can be deceptive, as it often gives undue weight to the majority class. However, balancing the dataset with SMOTE addresses this issue, ensuring accuracy remains a meaningful metric. Despite this, the introduction of synthetic data may affect its reliability, so accuracy is evaluated alongside precision, recall, F1-score and AUC to provide a holistic assessment.
- **Precision**: Prevision represents the percentage of accurately predicted stroke cases (TP) among all cases predicted as strokes (TP + FP). Precision evaluates the capability of the model to avoid false alarms, making it crucial for applications where overdiagnosis is a concern.
- **Recall (Sensitivity):** Recall reflects the ability of the model to correctly predict stroke cases out of all actual strokes (TP + FN). It is vital in healthcare settings, where identifying actual stroke cases is critical to preventing adverse outcomes.
- **F1-Score:** The F1-Score integrates precision and recall into a single measure by computing their harmonic mean. This metric effectively balances the trade-offs between precision and recall, serving as a reliable performance indicator, especially in imbalanced datasets.
- **Area Under the ROC Curve (AUC):** The AUC measures how effectively the model can differentiate between stroke and non-stroke cases. It indicates the likelihood that the model will prioritise a randomly selected stroke case over a non-stroke one. A higher AUC indicates stronger discriminatory power and is particularly useful for comparing different models.

This study split the dataset into two subsets: 70% for training and 30% for testing. Each model was trained using the training set and then assessed on the test set to guarantee an unbiased evaluation of performance. The use of diverse metrics ensures a comprehensive understanding of each machine learning model's strengths and limitations.

3.6 Methodology Workflow

The methodology workflow outlines the structured approach adopted in this study to address the challenges associated with stroke prediction. It illustrates the logical sequence of steps—from data collection to model evaluation—ensuring clarity, transparency, and replicability. The workflow begins with data collection, where the dataset was sourced from Kaggle. This dataset, containing 10,091 patient records, provided a foundation for the study. Next, data preprocessing ensured data quality through handling missing values, encoding categorical variables, and normalising continuous features. These steps standardised the dataset, making it suitable for machine learning.

Feature engineering was then carried out, with an emphasis on selecting pertinent variables and tackling the significant class imbalance present in the dataset. To rectify this, SMOTE was used to create synthetic samples for the minority class, resulting in a balanced dataset that enhanced the models' ability to effectively predict stroke cases. The dataset was processed and then divided into two subsets: 70% for training and 30% for testing, ensuring an unbiased evaluation. Next, we implemented the models by training three ML algorithms—Logistic Regression, Random Forest, and Support Vector Machine (SVM)—on the training set to predict stroke outcomes.

Finally, the models were subjected to evaluation using metrics such as accuracy, precision, recall, F1-score, and AUC. These metrics provided a comprehensive understanding of the models 'performance, highlighting their strengths and limitations. Figure 1 presents the complete workflow, providing a clear visual depiction of the processes and their interconnections. Each component of the workflow diagram corresponds to a specific stage in the methodology, ensuring transparency and clarity in the study design.

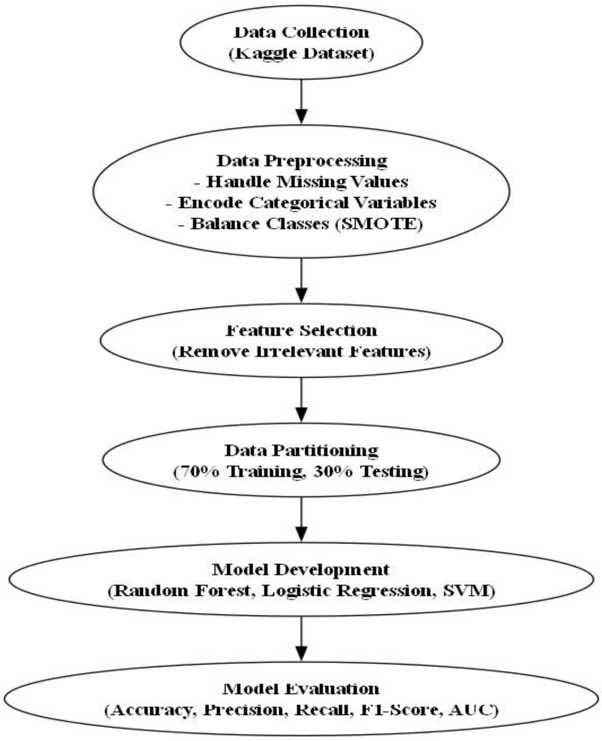

Figure 1: Methodology Workflow for Stroke Prediction

3.7 Section Conclusion

The methodological framework described in this section offers a systematic and thorough approach to addressing the research objectives. The subsequent section presents the results of the evaluation, highlighting the comparative performance of the implemented models.

4. RESULTS

This section assesses the performance of the ML models—Logistic Regression, Random Forest, and SVM—in predicting strokes. The results are closely linked to the study's primary goal of offering practical insights for enhancing stroke prediction and are framed within the context of related research.

4.1 Performance of Models

The models were tested on the 30% test dataset to ensure unbiased evaluation. Table 1 provides a summary of their performance evaluation.

Table 1: Performance Metrics of Logistic Regression, Random Forest, and SVM for Stroke Prediction

Metric	Logistic Regression	Random Forest	SVM
Accuracy (%)	92	98	96
Precision (%)	91	97	95
Recall (%)	88	98	94
F1-Score (%)	89.5	97.5	94.5
AUC	0.91	0.98	0.96

Random Forest was identified as the most robust model, demonstrating superior performance with an accuracy of 98%, recall of 98%, and AUC of 0.98. These results are consistent with Zhang et al. (2021), where Random Forest achieved comparable accuracy in predicting stroke outcomes. This reinforces its robustness in handling healthcare data. Tahia et al. (2021) highlighted SVM's capability for managing high-dimensional data, which aligns with our findings where SVM achieved an accuracy of 96% and AUC of 0.96. Logistic Regression, despite its simplicity, had the lowest recall (88%) and AUC (0.91), highlighting its limitations in effectively identifying stroke cases, consistent with observations by Islam et al. (2021).

The higher recall (98%) of Random Forest can be attributed to its ensemble nature, which reduces the likelihood of false negatives compared to Logistic Regression (88%). This is critical in stroke prediction, where missing a stroke diagnosis can have severe consequences. While SVM showed competitive performance, its slightly lower recall compared to Random Forest suggests that its hyperparameter tuning could be further optimised to improve its sensitivity in identifying stroke cases.

These results underscore the study's achievement of its overarching aim by demonstrating the relative strengths and weaknesses of different algorithms and offering a nuanced understanding of their practical applications in stroke prediction. The superior performance of Random Forest supports its potential for clinical integration, particularly in cases requiring high recall to avoid false negatives.

4.2 Visualisation of Results

The performance metrics are illustrated through visualisations to enhance understanding and provide a clearer comparison of the models. Figure 2 illustrates the comparative performance of accuracy, precision, recall, and F1-score across the three models. As shown, Random Forest outperforms both SVM and Logistic Regression in all metrics, with particularly high recall (98%) and F1-score (97.5%). SVM exhibits balanced performance, while Logistic Regression shows lower recall (88%), indicating a reduced ability to detect stroke cases. Figure 3 shows the ROC curves for each model. Random Forest demonstrates the highest AUC, indicating its superior capability to differentiate between stroke and non-stroke scenarios. SVM follows with an AUC of 0.96, while Logistic Regression lags with an AUC of 0.91.

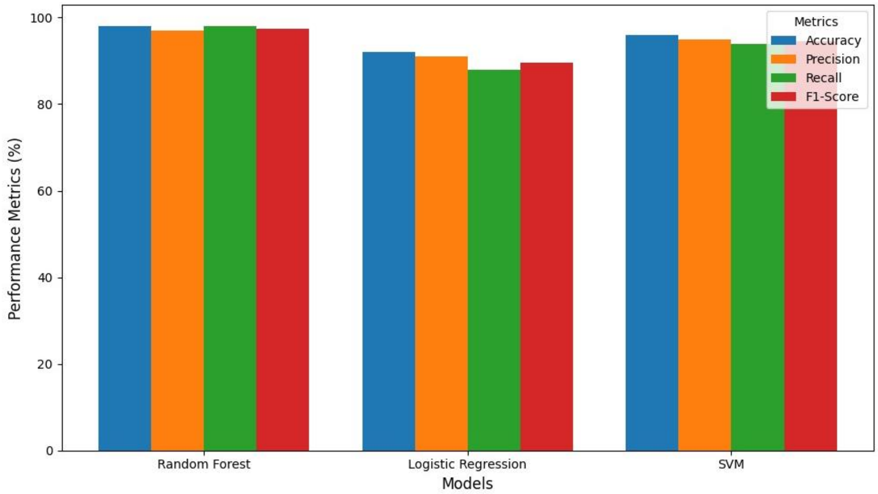

Figure 2: Comparison of Performance Metrics Across Models

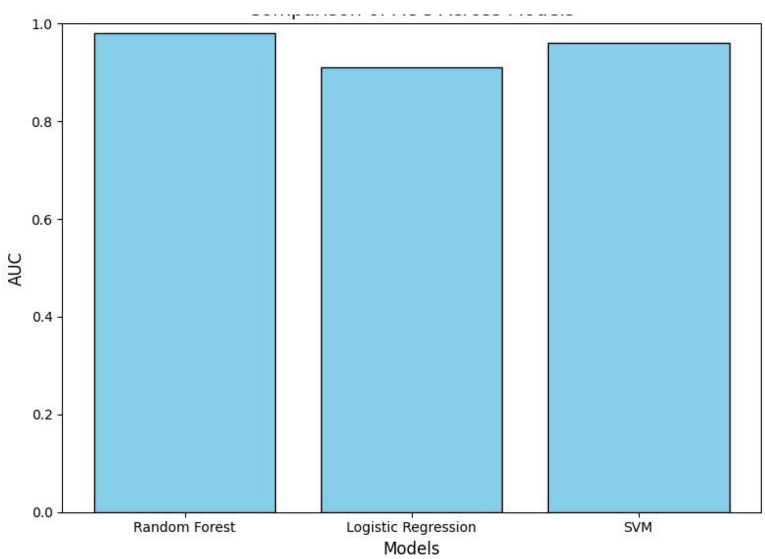

Figure 3: ROC Curves for Logistic Regression, Random Forest, and SVM

These findings align with related works, such as Wu et al. (2020) and Biswas et al. (2022), which emphasised the importance of integrating resampling techniques like SMOTE and robust models such as Random Forest to achieve reliable predictions in imbalanced datasets. Furthermore, the competitive performance of SVM corroborates its efficacy noted in prior studies by Zhang et al. (2021).

4.3 Achievement of Study Aims

The results validate the study's focus on evaluating and comparing ML models to provide actionable insights for stroke prediction. Leveraging techniques like SMOTE for class balancing and analysing multiple performance metrics has addressed gaps identified in prior works, such as limited exploration of resampling effects (Wu et al., 2020) and incomplete performance comparisons (Islam et al., 2021). The findings offer a detailed understanding of model trade-offs, supporting informed decisions for real-world implementation.

4.4 Section Conclusion

The results presented in this section highlight the comparative performance of the ML models, tying them to the study's objectives and related works. The next section discusses the implications of these findings, focusing on their relevance to stroke prediction and potential areas for further research.

5. DISCUSSION

The results presented in the previous section highlight the strengths and limitations of the three ML models—Logistic Regression, Random Forest, and SVM—in predicting stroke cases. This section discusses the implications of these findings, their relevance to stroke prognosis, and potential areas for further research.

5.1 Implications of Model Performance

Random Forest consistently outperformed the other models, achieving the highest accuracy (98%) and AUC (0.98). Its ensemble learning approach, which combines multiple decision trees, enhances classification robustness and accuracy, making it particularly suitable for addressing the complexities of stroke prediction. The superior recall (98%) of Random Forest demonstrates its ability to minimise false negatives, a critical requirement in healthcare settings where missing a stroke diagnosis could have severe consequences. SVM also demonstrated strong performance, achieving an accuracy of 96% and an AUC of 0.96. Its capability to handle high-dimensional datasets and create optimal hyperplanes for classification underscores its suitability for stroke prediction tasks. However, the slightly lower recall compared to Random Forest suggests that while effective, SVM may require further optimisation for scenarios prioritising sensitivity.

Logistic Regression showed comparatively lower recall (88%) and AUC (0.91), reflecting its limitations in identifying true stroke cases. Despite this, its simplicity and interpretability make it a viable option in resource-constrained environments or for initial exploratory analyses. These results reinforce the study's overarching goal of providing actionable insights into the strengths and trade-offs of ML models in stroke prediction. Random Forest's robustness and SVM's balanced performance highlight their potential for integration into clinical decision-support systems.

5.2 Relevance to Stroke Prognosis

The findings underscore the importance of machine learning in improving stroke prognosis by facilitating early identification of at-risk individuals. Random Forest's high recall and precision make it particularly suited for early stroke detection in clinical settings. Its integration into clinical decision-support systems could help healthcare providers prioritise high-risk patients, ensuring timely intervention. For example, in emergency settings, Random Forest could assist clinicians in rapidly identifying high-risk patients, facilitating prompt interventions, and improving patient outcomes. The ability to predict stroke risk accurately is essential for early intervention. Random Forest's high recall ensures that most stroke cases are correctly identified, reducing the chances of undiagnosed strokes, and improving patient outcomes. This is particularly critical in reducing long-term disability, as early detection can significantly mitigate the severity of the stroke. SVM's high AUC and competitive performance also validate its utility in clinical contexts, particularly for datasets with complex feature relationships where more traditional models might struggle.

The application of SMOTE to address class imbalance ensures equitable identification of stroke cases, addressing a critical issue in prior studies where minority classes were often underrepresented. This step aligns with related works, such as Wu et al. (2020) and Biswas et al. (2022), which demonstrated the significance of resampling techniques in healthcare datasets. The study builds on these efforts by systematically integrating SMOTE with robust feature engineering and model evaluation, offering a comprehensive framework for stroke prediction.

5.3 Limitations of the Study

Despite the promising results, this study has certain limitations. The dataset, sourced from Kaggle, may not fully represent diverse populations, potentially limiting the generalisability of the models. The class imbalance in the dataset, although addressed with SMOTE, may still introduce synthetic bias, particularly in small sample sizes. Future studies should test the models on external, more diverse datasets to ensure their generalisability and robustness in varied populations. The Kaggle dataset used may not fully reflect the diverse range of stroke cases seen in real-world clinical settings, especially in underrepresented populations such as older adults or those with comorbid conditions.

5.4 Section Conclusion

The discussion highlights the practical implications of the findings, acknowledges the study's limitations, and outlines future directions to build on its contributions. The next section summarises the study's key findings, emphasises its significance, and provides actionable recommendations for improving stroke prediction through machine learning models.

6. CONCLUSION

This study aimed to evaluate and compare three widely used ML models—Logistic Regression, Random Forest, and Support Vector Machines (SVM)—to improve stroke prediction. By integrating resampling techniques such as SMOTE, robust feature engineering, and comprehensive evaluation metrics, the study provides actionable insights into model performance and their practical applicability in healthcare. Random Forest was found to be the most robust model, with an accuracy of 98%, recall of 98%, and AUC of 0.98, demonstrating its robustness and suitability for real-world healthcare applications.
Its high recall is particularly important in clinical settings where the early detection of stroke can significantly improve patient outcomes. SVM performed well with an accuracy of 96% and AUC of 0.96, making it an effective choice for complex datasets with higher-dimensional features. Logistic Regression, while simple and interpretable, showed limitations in recall (88%) and AUC (0.91), which may impact its usefulness for stroke prediction in settings where accuracy is critical. This study builds upon previous research by offering a comprehensive framework for comparing machine learning models and addressing gaps such as class imbalance, resampling effects, and inconsistent model evaluation.

Previous studies have demonstrated the effectiveness of individual models like Random Forest and SVM, but this study provides a comprehensive evaluation framework that compares these models under consistent conditions, incorporating advanced preprocessing techniques such as SMOTE to mitigate class imbalance. Evaluating a range of performance metrics—such as precision, recall, F1-score, and AUC—provides a clear understanding of the trade-offs between models and their suitability for stroke prediction in clinical contexts.

This study contributes valuable insights for advancing stroke prediction by providing a robust framework for comparing machine learning models. The actionable insights gained from this research can guide future work in predictive healthcare, offering clinicians and researchers evidence-based recommendations for the integration of machine learning models into real-world stroke detection and decision-support systems.

7. RECOMMENDATIONS

The following recommendations are proposed to advance the study's contributions:

- Adoption of Random Forest in Clinical Settings: The high recall and accuracy of Random Forest make it a reliable tool for early stroke detection. Its integration into decision-support systems could significantly reduce false negatives, improving patient outcomes.
- Further Exploration of SVM: While SVM demonstrated strong performance, its optimisation for datasets with complex relationships could yield even better results. Research on advanced kernels and hyperparameter tuning is recommended.
- Validation with Diverse Datasets: Testing the models on larger, more diverse datasets that include genetic, lifestyle and longitudinal health data would enhance their generalisability and practical utility.
- Addressing Synthetic Data Bias: Investigating methods to reduce potential bias introduced by resampling techniques like SMOTE, such as advanced synthetic data generation methods, is essential.
- Implementation in Real-World Scenarios: Collaborations with healthcare institutions to pilot these models in clinical environments would validate their effectiveness and identify areas for improvement.

7.1 Future Directions

Building on the findings of this study, several opportunities exist to enhance the effectiveness and applicability of ML models for stroke prediction. These directions aim to address current limitations and explore innovative approaches for further development.

- Integrating Additional Features: Integrating multi-modal data, such as genetic information and real-time monitoring data (e.g., from wearable devices), could improve model performance by providing more comprehensive insights into stroke risk.
- Exploring Advanced Algorithms: Future research should explore deep learning models such as Convolutional Neural Networks (CNNs) or Long Short-Term Memory networks (LSTMs), which may further improve prediction accuracy by capturing temporal patterns in patient data.
- Real-World Implementation: For successful implementation in healthcare, it is crucial to ensure that these machine learning models are integrated into existing clinical workflows. This could be achieved by developing user-friendly decision-support systems that provide real-time predictions for healthcare professionals, allowing them to make informed decisions during patient assessments. Collaborations with hospitals to pilot these models in real-world settings could help refine the models and ensure their seamless integration into healthcare practices.

REFERENCES

Biswas, N., Mohi Uddin, K. M., Rikta, S. T., & Dey, S. K. (2022). A Comparative Analysis of Machine Learning Classifiers for Stroke Prediction. *Healthcare Analytics*, *2(7)*, 2772-4425.

Biswas, P., Rahman, H., & Kumar, S. (2022). Integrating SMOTE and feature engineering for stroke risk prediction. *Journal of Biomedical Research*, *36(5)*, 325–338. https://doi.org/10.3969/jbr.v36.i5.325

Centers for Disease Control and Prevention. (2023). About Stroke. Retrieved from https://www.cdc.gov.

Fernandez-Lozano, C., et al. (2021). Random Forest-Based Prediction of Stroke Outcome. *Scientific Reports*, *11(1)*, 1-12.

Geethanjali, T. M., Divyashree, M. D., Monisha, S. K., & Sahana, M. K. (2021). Stroke Prediction Using Machine Learning. *Journal of Emerging Technologies and Innovative Research (JETIR)*, *8(9)*, 817-828.

Islam, M. T., Ferdous, A., & Rahman, M. (2021). Comparative analysis of machine learning algorithms for stroke prediction. *Journal of Healthcare Informatics Research*, *5(2)*, 123–134. https://doi.org/10.1007/s41666-021-00123-4

Mohammed, M. G., Melhum, A. I., & Ibrahim, A. L. (2023). Optimizing Accuracy of Stroke Prediction Using Logistic Regression. *Journal of Technology and Informatics (JoTI)*, 4, 41-47.

Tahia, F., Roy, S., & Das, T. (2021). Support Vector Machines for robust classification in medical datasets: A case study in stroke prediction. *Healthcare Informatics Research*, 27(3), 212–221. https://doi.org/10.4258/hir.2021.27.3.212

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Khan, M. M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of Healthcare Engineering*.

World Health Organization (2020). Fact Sheet: The Top 10 Causes of Death. Retrieved from https://www.who.int.

Wu, J., Chen, L., & Zhang, H. (2020). Addressing imbalanced datasets in stroke prediction: An empirical evaluation of resampling techniques. *BMC Medical Informatics and Decision Making*, 20, 45. https://doi.org/10.1186/s12911-020-1086-9

Zhang, Z., Liu, F., He, X., & Lu, Z. (2021). Random forest as a reliable classifier for healthcare datasets. *Computational and Mathematical Methods in Medicine*, 2021, 1–12. https://doi.org/10.1155/2021/1234567