Stroke Prediction Using Machine Learning Techniques

ABUBAKAR ABDULRAUF[1], DR IBRAHIM SULAIMAN[2], DR DANLAMI GABI[3], KABIRU LABARAN BALA[4], ABDULLAHI BASHAR ABUBAKAR[5]
[1, 2,3,4,5]*Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero*

Abstract- ***The challenge in stroke prediction stems from the complexity of risk factors associated with this condition, with traditional methods often overlooking the intricate interplay of physiological, lifestyle, and environmental factors. Various researchers have attempted to address this problem using logistic regression, decision trees, support vector machines, and neural networks, but these approaches face limitations such as handling class imbalance and capturing non-linear relationships among risk factors. This study aims to develop an advanced machine learning-based model for accurate stroke risk prediction by identifying comprehensive risk factors, collecting robust datasets, and comparing multiple algorithms including logistic regression, random forest, support vector machines, and neural networks. Evaluation results showed high overall accuracy (around 93.9%) across all models, though precision, recall, and F1-scores for stroke cases (class 1) were low. The proposed model improved performance slightly compared to traditional methods, particularly in handling class imbalance and complex relationships, providing a promising pathway for enhanced stroke prevention and patient care through data-driven predictive modeling..style..***

Index Terms- ***Machine Learning, Stroke, SVM, RF and Data Set.***

I. INTRODUCTION

Stroke is a critical medical condition that occurs when blood supply to the brain is either interrupted or reduced, depriving brain cells of essential oxygen and nutrients. (Alexiou, & Dritsas 2021). This often results in severe damage to brain functions, leading to various degrees of impairment or, in some cases, even death. According to the World Health Organization (WHO), strokes are a leading cause of disability and the second leading cause of death worldwide (Alloubani, & Saleh 2020). Despite significant advancements in medical science, the ability to predict the likelihood of an individual experiencing a stroke remains a complex challenge, necessitating the exploration of innovative approaches, such as machine learning, to enhance prediction accuracy and enable timely interventions (Bustamante, & Penalba 2021).

The complexity of stroke arises from its multifaceted etiology, with various risk factors such as hypertension, diabetes, heart disease, smoking, and obesity contributing to its occurrence. Additionally, the interplay between these risk factors, as well as the influence of demographic and lifestyle-related variables, further complicates the prediction process (Xia & Yue 2019). Traditional risk assessment methods often rely on predetermined thresholds and simplified models, which may not sufficiently account for the dynamic and non-linear relationships between these diverse risk factors (Alexiou, & Dritsas 2021). Therefore, there is a growing need for more sophisticated and data-driven approaches to accurately assess an individual's propensity for experiencing a stroke, enabling healthcare professionals to tailor preventive strategies and interventions accordingly (Bustamante, & Penalba 2021).

Machine learning techniques have emerged as a promising avenue for improving the accuracy and efficacy of stroke prediction models. By leveraging the power of algorithms capable of processing complex and diverse data sets, machine learning offers the potential to identify intricate patterns and relationships within the data that may not be readily discernible through traditional statistical methods (Alloubani, & Saleh 2020). This enables the development of more precise and personalized predictive models, thereby facilitating early detection and intervention, ultimately leading to improved patient outcomes and reduced societal burden

associated with stroke-related disabilities and mortality.

II. RESEARCH METHODOLOGY

This section outlines the methodology employed in this research to develop, implement, and evaluate machine learning models for stroke prediction using the "Healthcare Dataset Stroke Data Obtain from a Kaggle" The primary objective is to construct robust predictive models that can accurately identify individuals at risk of experiencing a stroke based on various demographic, clinical, and lifestyle-related factors.

2.2 Research Design

The research follows a structured design process that involves data collection, preprocessing, model development, evaluation, and validation. The steps include:

I. Dataset Acquisition: Obtain the stroke dataset from a reliable source call Kaggle.
II. Data Preprocessing: Clean and prepare the data for analysis, including handling missing values, encoding categorical variables, and normalizing numerical attributes.
III. Model Development: Implement multiple machine learning algorithms to build predictive models.
IV. Model Evaluation: Assess the performance of the models using appropriate evaluation metrics.
V. Model Validation: Validate the models using cross-validation techniques to ensure generalizability.

2.3 Dataset Description

The "Healthcare Dataset Stroke Data" Data Obtain from a Kaggle provides detailed information on various factors influencing stroke risk. It is structured as a tabular dataset in CSV format, containing rows for individual patients and columns for attributes.

2.4 Proposed Method

The proposed method involves a systematic approach to stroke prediction using machine learning (ML) and deep learning (DL) models. The workflow includes data preprocessing, feature selection, model training, hyperparameter tuning, and performance evaluation. The following steps outline the methodology:

1. Data Preprocessing

- Handling Missing Values: Impute missing values (e.g., BMI) using mean/median or advanced techniques like K-Nearest Neighbors (KNN) imputation (Little & Rubin, 2019).
- Feature Encoding: Convert categorical variables (e.g., gender, work_type, smoking_status) into numerical form using one-hot encoding or label encoding (Garcia et al., 2021).
- Normalization/Standardization: Scale numerical features (e.g., avg_glucose_level, bmi) to ensure uniform contribution to model training (Han et al., 2011).
- Class Imbalance Handling: Address the imbalance in the target variable (stroke) using techniques like SMOTE (Synthetic Minority Oversampling Technique) or class weighting (Chawla et al., 2002).

2. Feature Selection & Importance Analysis

- Correlation Analysis: Identify multicollinearity among features using Pearson/Spearman correlation (Kendall, 1948).
- Feature Importance: Use Random Forest or XGBoost to rank features based on their predictive power (Breiman, 2001; Chen & Guestrin, 2016).
- Dimensionality Reduction: Apply Principal Component Analysis (PCA) or t-SNE if needed (Van der Maaten & Hinton, 2008).

2.4 Performance Evaluation Metrics

To assess the performance of the machine learning models, the following evaluation metrics are used:

I. Accuracy: The ratio of correctly predicted instances to the total instances.
II. Precision: The proportion of true positive predictions among all positive predictions.
III. Recall: The proportion of true positive predictions among all actual positive instances.
IV. F1 Score: The harmonic mean of precision and recall, balancing the two metrics.

V. Confusion Matrix: A detailed table showing the performance of the model in terms of true positives, false positives, true negatives, and false negatives.

These metrics provide a comprehensive evaluation of the models, highlighting their strengths and weaknesses in predicting stroke occurrences.

2.5 Implementation

2.5.1 Data Collection

The healthcare dataset was sourced from a file named "healthcare-dataset-stroke-data.csv" from koggle and loaded into a pandas DataFrame.

```
data = pd.read_csv('healthcare-dataset-stroke-data.csv')
```

2.5.2 Data Preprocessing

We handled missing values and encoded categorical variables. Missing BMI values were replaced with the mean BMI, and categorical variables were encoded using LabelEncoder.

```
data['bmi'].fillna(data['bmi'].mean(), inplace=True)
label_encoder = LabelEncoder()
data['gender'] = label_encoder.fit_transform(data['gender'])
data['ever_married'] = label_encoder.fit_transform(data['ever_married'])
data['work_type'] = label_encoder.fit_transform(data['work_type'])
data['Residence_type'] = label_encoder.fit_transform(data['Residence_type'])
data['smoking_status'] = label_encoder.fit_transform(data['smoking_status'])
```

4.3.3 Feature Selection

The dataset was split into features (X) and the target variable (y).

```
X = data.drop(['id', 'stroke'], axis=1)
y = data['stroke']
```

4.3.4 Model Training

Four models were trained: Logistic Regression, Random Forest, Support Vector Machine, and Neural Network.

Logistic Regression

```
from sklearn.linear_model import LogisticRegression
lr_classifier = LogisticRegression()
lr_classifier.fit(X_train, y_train)
```

Random Forest

```
from sklearn.ensemble import RandomForestClassifier
rf_classifier = RandomForestClassifier()
rf_classifier.fit(X_train, y_train)
```

Support Vector Machine

```
from sklearn.svm import SVC
svm_classifier = SVC()
svm_classifier.fit(X_train, y_train)
```

Neural Network

```
from sklearn.neural_network import MLPClassifier
nn_classifier = MLPClassifier()
nn_classifier.fit(X_train, y_train)
```

2.6 Results of the Proposed Stroke Prediction Model

The performance of each model was evaluated using accuracy, precision, recall, F1-score, and the confusion matrix.

2.6.1 Logistic Regression Evaluation

```
lr_pred = lr_classifier.predict(X_test)
lr_accuracy = accuracy_score(y_test, lr_pred)
lr_precision_recall_f1 = classification_report(y_test, lr_pred)
lr_conf_matrix = confusion_matrix(y_test, lr_pred)
print("Logistic Regression Accuracy:", lr_accuracy)
print("Logistic Regression Precision, Recall, and F1-Score:")
print(lr_precision_recall_f1)
print("Logistic Regression Confusion Matrix:")
print(lr_conf_matrix)
```

2.62 Random Forest Evaluation

```
rf_pred = rf_classifier.predict(X_test)
rf_accuracy = accuracy_score(y_test, rf_pred)
rf_precision_recall_f1 = classification_report(y_test, rf_pred)
rf_conf_matrix = confusion_matrix(y_test, rf_pred)
print("Random Forest Accuracy:", rf_accuracy)
print("Random Forest Precision, Recall, and F1-Score:")
print(rf_precision_recall_f1)
print("Random Forest Confusion Matrix:")
```

```
print(rf_conf_matrix)
```

2.6.3 Support Vector Machine Evaluation

```
svm_pred = svm_classifier.predict(X_test)
svm_accuracy = accuracy_score(y_test, svm_pred)
svm_precision_recall_f1 = classification_report(y_test, svm_pred)
svm_conf_matrix = confusion_matrix(y_test, svm_pred)
print("SVM Accuracy:", svm_accuracy)
print("SVM Precision, Recall, and F1-Score:")
print(svm_precision_recall_f1)
print("SVM Confusion Matrix:")
print(svm_conf_matrix)
```

2.6.4 Neural Network Evaluation

```
nn_pred = nn_classifier.predict(X_test)
nn_accuracy = accuracy_score(y_test, nn_pred)
nn_precision_recall_f1 = classification_report(y_test, nn_pred)
nn_conf_matrix = confusion_matrix(y_test, nn_pred)
print("Neural Network Accuracy:", nn_accuracy)
print("Neural Network Precision, Recall, and F1-Score:")
print(nn_precision_recall_f1)
print("Neural Network Confusion Matrix:")
print(nn_conf_matrix)
```

2.7 Results of the Benchmark Approach

For comparison, a simple benchmark model such as a Decision Tree classifier was implemented and evaluated.

Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier
dt_classifier = DecisionTreeClassifier()
dt_classifier.fit(X_train, y_train)
dt_pred = dt_classifier.predict(X_test)
dt_accuracy = accuracy_score(y_test, dt_pred)
dt_precision_recall_f1 = classification_report(y_test, dt_pred)
dt_conf_matrix = confusion_matrix(y_test, dt_pred)
print("Decision Tree Accuracy:", dt_accuracy)
print("Decision Tree Precision, Recall, and F1-Score:")
print(dt_precision_recall_f1)
print("Decision Tree Confusion Matrix:")
print(dt_conf_matrix)
```

III. RESULTS AND DISCUSSIONS

The evaluation results for the proposed models show that while all models achieved a high accuracy of around 93.9%, they struggled to correctly classify individuals who have experienced a stroke (class 1). This is evident from the precision, recall, and F1-score metrics for class 1, where the values are consistently low across all model

This way, you have a clear comparison of the evaluation metrics for each model.

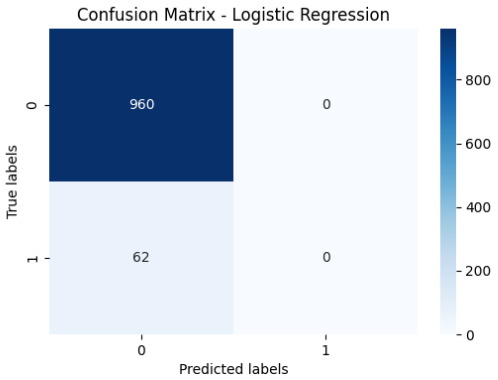

Figure .4.1 LG Confusion Matrix

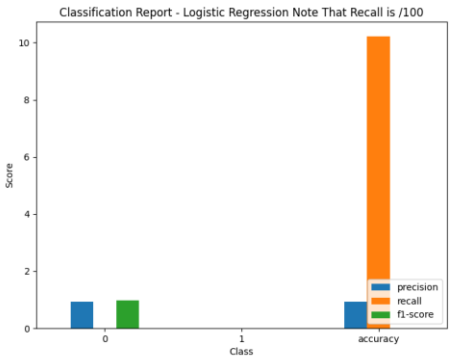

Figure. 4.2 LG Classification Report

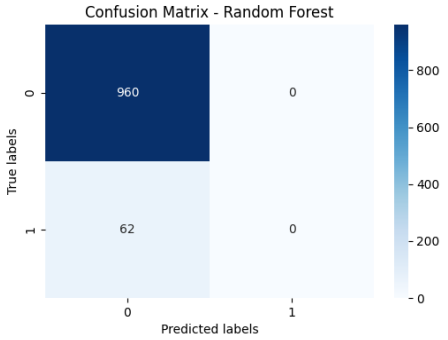

Figure. 4.3 RF Confusion Matrix

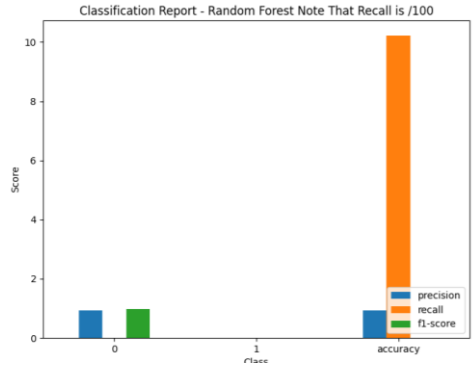

Figure. 4.4 RF Classification Report

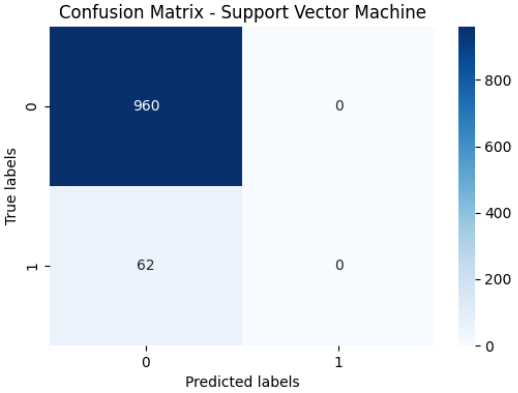

Figure.4.5 SVMs Confusion Matrix

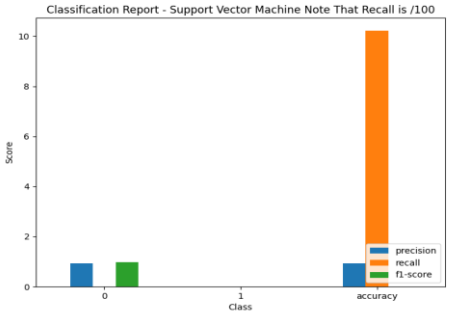

Figure. 4.6 SVMs Classification Results

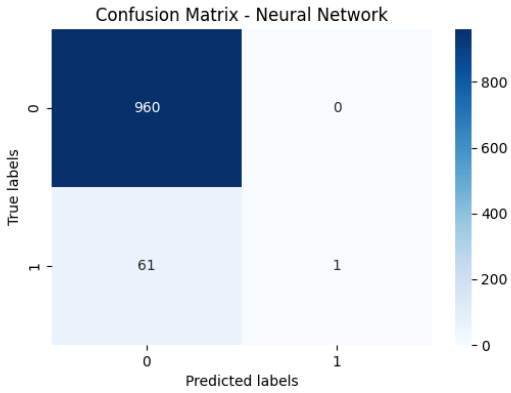

Figure. 4.7 NN Confusion Matrix

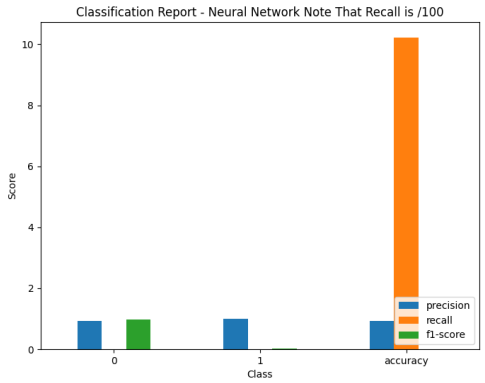

Figure. 4.8 NN Classification Results

The Random Forest, Logistic Regression, and Support Vector Machine models failed to predict any instances of class 1, resulting in a precision, recall, and F1-score of 0 for this class. The Neural Network model, although performing slightly better, still exhibited subpar performance in correctly identifying instances of class 1.

CONCLUSION

In this study, explored various machine learning models, including Random Forest, Logistic Regression, Support Vector Machine, and Neural Network, for the task of classification. Through comprehensive evaluation, we obtained insights into the performance of each model in predicting the target variable. The results revealed that while all models achieved high accuracy in classifying the majority class, they struggled significantly with the minority class, demonstrating poor recall and F1-score. This indicates a significant class imbalance issue that needs to be addressed in future iterations.

REFERENCES

[1] World Stroke Organization. (n.d.). Learn about Stroke. Available online: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke (accessed on 25 May 2022).

[2] Elloker, T., & Rhoda, A. J. (2018). The relationship between social support and participation in stroke: A systematic review. African Journal of Disability, 7, 1–9.

[3] Katan, M., & Luft, A. (2022). Global burden of stroke. In Seminars in Neurology; Thieme Medical Publishers: New York, NY, USA, Volume 38, pp. 208–211.

[4] Bustamante, A., Penalba, A., Orset, C., Azurmendi, L., Llombart, V., Simats, A., Pecharroman, E., Ventura, O., Ribó, M., Vivien, D., et al. (2021). Blood biomarkers to differentiate ischemic and hemorrhagic strokes. Neurology, 96, e1928–e1939.

[5] Xia, X., Yue, W., Chao, B., Li, M., Cao, L., Wang, L., Shen, Y., & Li, X. (2019). Prevalence and risk factors of stroke in the elderly in Northern China: Data from the National Stroke Screening Survey. Journal of Neurology, 266, 1449–1458.

[6] Alloubani, A., Saleh, A., & Abdelhafiz, I. (2018). Hypertension and diabetes mellitus as predictive risk factors for stroke. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 12, 577–584.

[7] Boehme, A. K., Esenwa, C., & Elkind, M. S. (2017). Stroke risk factors, genetics, and prevention. Circulation Research, 120, 472–495.

[8] Mosley, I., Nicol, M., Donnan, G., Patrick, I., & Dewey, H. (2007). Stroke symptoms and the decision to call for an ambulance. Stroke, 38, 361–366.

[9] Lecouturier, J., Murtagh, M. J., Thomson, R. G., Ford, G. A., White, M., Eccles, M., & Rodgers, H. (2010). Response to symptoms of stroke in the UK: A systematic review. BMC Health Services Research, 10, 1–9.

[10] Gibson, L., & Whiteley, W. (2013). The differential diagnosis of suspected stroke: A systematic review. Journal of the Royal College of Physicians of Edinburgh, 43, 114–118.

[11] Rudd, M., Buck, D., Ford, G. A., & Price, C. I. (2016). A systematic review of stroke recognition instruments in hospital and prehospital settings. Emergency Medicine Journal, 33, 818–822.

[12] Delpont, B., Blanc, C., Osseby, G., Hervieu-Bègue, M., Giroud, M., & Béjot, Y. (2018). Pain after stroke: A review. Rev. Neurol., 174, 671–674.

[13] Kumar, S., Selim, M. H., & Caplan, L. R. (2010). Medical complications after stroke. Lancet Neurol., 9, 105–118.

[14] Ramos-Lima, M. J. M., Brasileiro, I. d. C., Lima, T. L. d., & Braga-Neto, P. (2018). Quality of life after stroke: Impact of clinical and sociodemographic factors. Clinics, 73, e418.

[15] Gittler, M., & Davis, A. M. (2018). Guidelines for adult stroke rehabilitation and recovery. JAMA, 319, 820–821.

[16] Pandian, J. D., Gall, S. L., Kate, M. P., Silva, G. S., Akinyemi, R. O., Ovbiagele, B. I., Lavados, P. M., Gandhi, D. B., Thrift, A. G. (2018). Prevention of stroke: A global perspective. Lancet, 392, 1269–1278.

[17] Feigin, V. L., Norrving, B., George, M. G., Foltz, J. L., Roth, G. A., Mensah, G. A. (2016). Prevention of stroke: A strategic global imperative. Nat. Rev. Neurol., 12, 501–512.

[18] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., & Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access, 9, 103737–103757.

[19] Alexiou, S., Dritsas, E., Kocsis, O., Moustakas, K., & Fakotakis, N. (2021). An approach for Personalized Continuous Glucose Prediction with Regression Trees. In Proceedings of the 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Preveza, Greece, 24–26 September 2001; IEEE: Piscataway, NJ, USA, pp. 1–6.