Wadi Alshatti University Journal of Pure and Applied Sciences

مجلة جامعة وادي الشاطئ للعلوم البحتة و التطبيقية

Volume 4, No. 2, July-December 2026 Online ISSN: 3006-0877 المجلد 4، الاصدار 2، يوليو-ديسمبر 2026

RESEARCH PAPER **COMPUTER SCIENCE**

A Comparison of Machine Learning Models for Stroke Risk Prediction

Mona Khalifa A. Aljero [1,*]

[1]Information Technology Department, Faculty of Education, Misurata University, Misurata, Libya

ARTICLE HISTORY

Received 12 March 2026
Revised 15 April 2026
Accepted 20 May 2026
Online 01 June 2026

KEYWORDS

Classification;
GP;
Machine Learning;
Models;
Stroke Prediction;
Unbalanced dataset.

ABSTRACT

Stroke is a major cause of death and permanent disability globally. Effective preventive measures depend on early stroke risk prediction. This, study investigates a form of machine learning classifiers for predicting the risk of stroke and evaluates their performance in comparison to more conventional models, including Artificial Neural Networks (ANN), Logistic Regression (LR), Decision Trees (DT), Support Vector Machines (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), AdaBoost, Gradient Boosting, and Genetic Programming (GP). A public dataset from DataHack comprising 4981 participants was used, with class imbalance addressed by using Synthetic Minority Over-sampling Techniques (SMOTE). The dataset was split into a training set with 80% and a testing set with 20%. According to the experimental results, all the proposed approaches performed well in all the presented categories. Moreover, the GP approach achieved the highest results among the proposed approaches, reaching 95.6% accuracy, a significant improvement over the 90% state-of-the-art accuracy. The GP approach delivers prediction accuracy by creating symbolic expressions that illustrate correlations between the probability of stroke and risk factors. This study highlights the potential of machine learning approaches in predictive healthcare applications while also emphasizing the need for improvements through hybrid approaches, parameter modifications, and practical medical integration.

مقارنة بين نماذج التعلم الالي للتنبأ بخطر الاصابة بالسكتة الدماغية

منى خليفة عبدالسلام الجرو [1,*]

الكلمات المفتاحية

التصنيف
البرمجة الجينية
التعلم الآلي
النماذج الالية
التنبؤ بالسكتة الدماغية
قاعدة البيانات الغير متوازنة

الملخص

تُعدّ السكتة الدماغية سببًا رئيسيًا للوفاة والإعاقة الدائمة على مستوى العالم. وتعتمد التدابير الوقائية الفعّالة على التنبؤ المبكر بخطر الإصابة بها. تبحث هذه الدراسة في نماذج تصنيف التعلّم الآلي للتنبؤ بخطر الإصابة بالسكتة الدماغية، و تقيم أداءها مقارنةً بنماذج التعلم التقليدية، بما في ذلك الشبكات العصبية الاصطناعية، والانحدار اللوجستي، وأشجار القرار، وآلات المتجهات الداعمة، وخوارزمية بايز البسيطة، وخوارزمية أقرب الجيران، وخوارزمية التعزز التكيفي و خوارزمية تعزيز التدرج و البرمجة الجينية. تم استخدام قاعدة بيانات عامة من **DataHack** تضم 4981 مشاركًا. وتم معالجة عدم توازن الفئات باستخدام تقنيات الزيادة الاصطناعية للأقليات .(SMOTE) تم تقسيم مجموعة البيانات إلى مجموعة تدريب بنسبة 80% ومجموعة اختبار بنسبة 20%. وفقًا للنتائج التجريبية، فإن جميع الأساليب المقترحة أدت أداءً جيدًا في جميع جوانب التقييم المستخدمة.-علاوة على ذلك، حققت طريقة البرمجة الجينية أعلى النتائج بين جميع الأساليب المقترحة، حيث بلغت دقتها 95.6%، وهو تحسن كبير مقارنةً بدقة أفضل نتيجة منشورة التي بلغت 90%. تُحقق طريقة البرمجة الجينية دقة تنبؤ عالية من خلال إنشاء تعابير رمزية توضح الارتباطات بين احتمالية الإصابة بالسكتة الدماغية وعوامل الخطر. تُبرز هذه الدراسة إمكانات أساليب التعلم الآلي في تطبيقات الرعاية الصحية التنبؤية، مع التأكيد في الوقت نفسه على الحاجة إلى تحسينات من خلال الأساليب الهجينة، وتعديلات المعلمات، والتكامل الطبي العملي.

Introduction

Stroke constitutes a significant cause of mortality and disability globally. Precise prediction of stroke risk can enable early intervention and preventive healthcare strategies. About 1 in 21 people died worldwide from stroke, classified as a major cause of death and disability. A stroke happens every 3 minutes and 17 seconds [1]. 87% of stroke-related disability occurs in low and middle-income countries, ranking stroke the second most common cause of disability worldwide [2]. Researchers have extensively employed traditional statistical models, such as Logistic Regression (LR), to forecast stroke, yet these models may fail to elucidate the interactions among risk factors. More sophisticated machine learning methodologies, including Random Forest (RF), Support Vector Machines (SVM), and

*Corresponding author
https://doi.org/10.63318/waujpasv4i2_01

Artificial Neural Networks (ANN), may offer enhanced accuracy but can be challenging to interpret. Genetic Programming (GP), which formulates mathematical expressions, presents an alternative method for predicting stroke risk factors without preconceived assumptions. Unlike deep learning, GP generates human-readable, interpretable, and extremely precise symbolic representations. While previous research has demonstrated the efficacy of GP in other medical prediction tasks, its utility in forecasting stroke risk has yet to be thoroughly investigated.

This study seeks to address this gap by comparing the performance of GP with classic machine learning models and advanced classifiers such as Gradient Boosting, which achieved an accuracy nearing 96%.

The remaining part of the study's structure is provided below. In section 2, the related work presents the available works on stroke prediction and models used with the evaluation results. In section 3, the detailed of the proposed approaches and the dataset used. In section 4, the evaluation and experimental findings are discussed. In section 5, demonstrated the discussion section. The last section, 6, contains the conclusion of this study.

Related work

Prior studies have concentrated on employing machine learning approaches for stroke prediction. LR is a widely used standard approach with interpretable coefficients; however, it exhibits poor predictive accuracy. Decision Trees (DT) and Random Forest (RF) enhance classification by handling non-linearity and the relationship between risk factors. SVMs have proven to perform well in binary classification, provided that an appropriate kernel is selected ANNs are "black-box" classifiers that limit interpretability, despite their effectiveness.

Evolutionary algorithms for medical prediction have also been discussed. There has also been discussion of medical prediction using evolutionary algorithms. GP has proven its potential to produce efficient and intelligible classifiers through its use in disease classification and biomarker creation. However, there hasn't been enough research done to determine how well it predicts stroke risks.

Authors in [3] developed a machine learning classifier for early stroke detection using a dataset retrieved from the China Health and Retirement Longitudinal Study between 2011 and 2020. They started with the imputation process. The final dataset was made up of 12,978 patients, 1011 classified as having experienced a stroke risk. Two strategies were employed to deal with the imbalanced dataset: Synthetic Minority Over-sampling Techniques (SMOTE) and under-sampling. In their approach, they used eight classifiers: RF, LR, SVM, Elastic Net, Gradient Boost, Lasso, XGBoost, and ANN. The results showed that nine parameters were associated with stroke risk. Moreover, the authors presented a monogram approach associated with the nine parameters for early stroke detection.

In [4], the authors investigated the performance of DT, Naive Bais (NB), Linear Discriminant Analysis (LDA), and K-Nearest Neighbors (KNN) for stroke detection on a dataset of 5110 patients. The KNN achieved the highest accuracy among the used classifiers with 90%. As a preprocessing step, they balanced the dataset used to reach an accurate result. For the evaluation of the four models, five metrics were used to evaluate each model.

[5] provided a machine learning approach for stroke prediction. The authors evaluated their approach on the collected data of 503842 participants (divided into three age groups) from 10 Chinese areas from 2004 until 2008. The dataset is divided into three sets: the training set with 85%, the validation set with 12.75%, and the test set with 2.25%. In order to measure the performance of the proposed approach, they used five different metrics. To ensure the good performance of the proposed approach, the authors compared it with different machine learning models. The top performance was achieved by the proposed ensemble technique.

The authors in [6] presented an ensemble machine learning approach to improve the effectiveness of the prediction. As an ensemble approach, the authors applied an ANN, Extra Trees, XGBoost, genetic algorithm, and RF on two stroke datasets. To balance the two datasets, the SMOTE technique was used. The dataset was split into 80 for the training and 20 for the testing. The highest performance was achieved by the ensemble Extra Trees, followed by the RF approach with 98.24% accuracy, 98.03%, respectively. These results outperformed the state-of-the-art results, indicating the power of this technique and its potential use in the healthcare sector.

In [7], the stroke prediction using a machine learning model was developed using a stroke dataset obtained from Kaggle. The authors applied some steps for the preprocessing, starting with dealing with the missing data, then they handled the imbalanced dataset (using the under-sampling technique), ending with performing label encoding. They used six machine learning techniques (LR, DT, NB, RF, SVM, and KNN). In their approach, they started with the preprocessing and balanced the dataset by applying the under-sampling technique. The best performance was achieved by the NB model with 82.2% accuracy. The model was trained using DT, RF, and Multi-layer Perceptron for stroke prediction, according to a research report [8], the three approaches' obtained accuracies were fairly similar, with a few minor variations. DT, RF, and Multi-layer Perceptron all had computed accuracy of 74.31%, 74.53%, and 75.02%, respectively.

It is still crucial to use suitable sampling techniques to handle the imbalanced issue in order to enhance the current studies on stroke prediction models. The performance of machine learning is affected by studies that detect the ratio of an imbalanced dataset. It is crucial to use SMOTE rather than other sampling techniques to obtain a balanced dataset because other techniques, such as down-sampling, could end up in the loss of important information needed to develop the prediction model. Machine learning experts working in healthcare must use appropriate techniques to address the imbalance issue. As a result, a higher F1-score will result from this without any data loss. Despite GP being proven to be efficient for other medical prediction tasks, its effectiveness in predicting stroke risk has not yet been fully investigated. This study aims to close this gap by evaluating the GP's performance with different models.

The primary contributions of this study are as follows: (a) applying different machine learning models on the stroke dataset; (b) the imbalanced issue was solved by using the SMOTE technique; and (c) the experimental results of the classification approaches were presented and compared.

Methodology

Dataset

DataHack provides a dataset named ‘data_stroke dataset’. This dataset is available on: https://github.com/nmelnyko/

data_stroke/blob/main/data_stroke.csv [9]. The features of the dataset and the description is shown in Table 1. This specific dataset comprises of 4981 (patients) rows and 11 (features) columns, the distribution of this dataset presented in Table 2. '1' in the output column ('stroke') is more likely than '0' in the same column (as shown in Figure 1, 248 have a stroke: 4733 do not have a stroke), indicating that this binary dataset is extremely unbalanced. The ratio of patients with a stroke to those without is approximately 1:9.

Table 1: Exploratory Data_stroke analysis.

No	Attribute Name	Description
1	id	Unique identifier
2	Gender	Gender of the patient
3	Age	Age of the patient
4	Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5	Heart-disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6	Ever-married	"No" or "Yes"
7	Work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8	Residence_type	"Rural" or "Urban"
9	Avg_glucose	average glucose level in blood
10	Bmi	body mass index
11	Smoking_status	"Formerly smoked", "never smoked", "smokes" or "Unknown"
12	Stroke	1 if the patient had a stroke or 0 if not

Table 2: Data_stroke dataset distribution of No-stroke and Stroke.

	No-stroke	Stroke	Total
Stroke -dataset	4733	248	4981

It is noticeable from the dataset that there are more women (58%) than men (42%), as shown in Figure 2. There are approximately 4.5 times as many patients without a history of heart disease as there are those with one, as shown in Figure 3. There are around four times as many patients without hypertension as those with it, as shown in Figure 4. The figures from 5 to 11 demonstrated the distribution of the rest of the features of the used dataset. The numeric attributes are: hypertension, heart disease, and stroke, which take only 0 or 1. The “Id” feature has been excluded from this study.

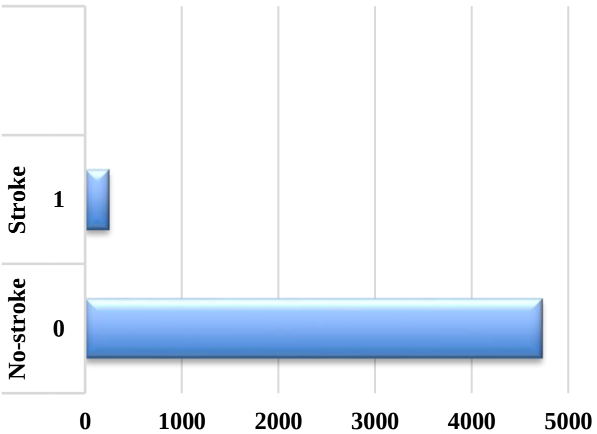

Figure 1: Distribution of the stroke dataset

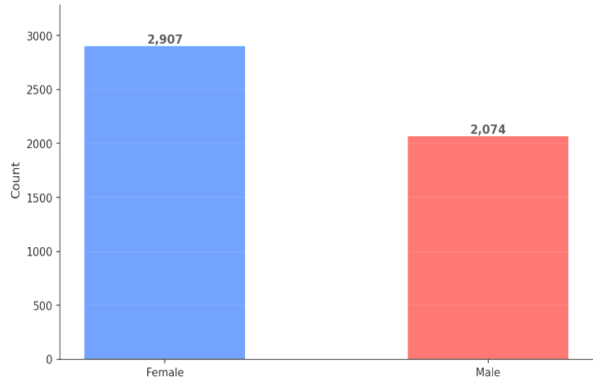

Figure 2: Distribution of gender in the dataset

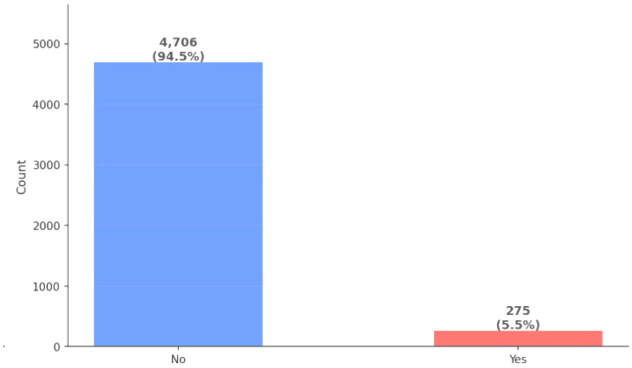

Figure 3: Distribution of patients with heart disease in the dataset

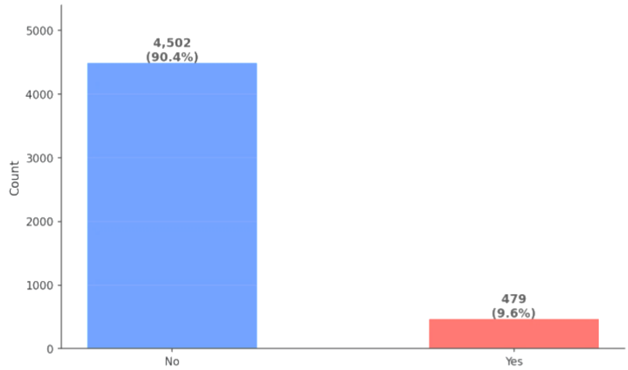

Figure 4: Distribution of patients with and without hypertension in the dataset

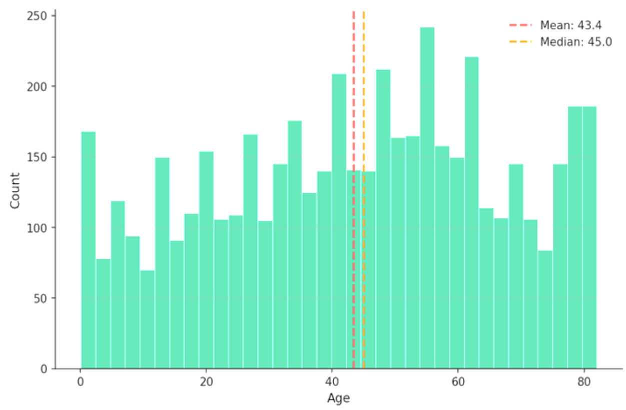

Figure 5: Distribution of age in the dataset

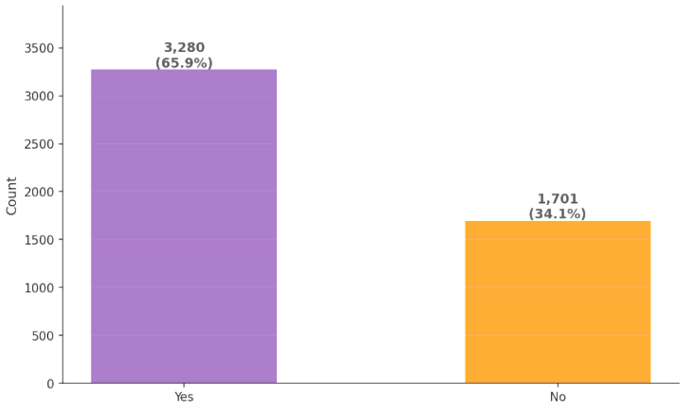

Figure 6: Distribution of marriage in the dataset

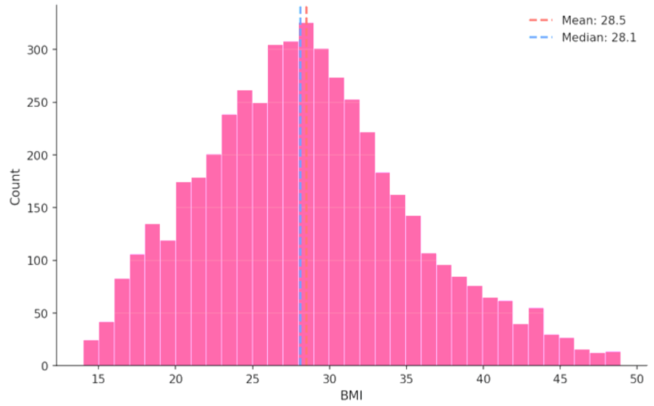

Figure 10: Distribution of BMI in the dataset

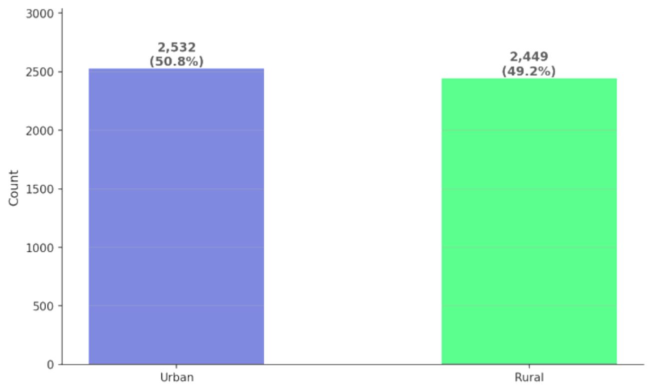

Figure 8: Distribution of residence type in the dataset

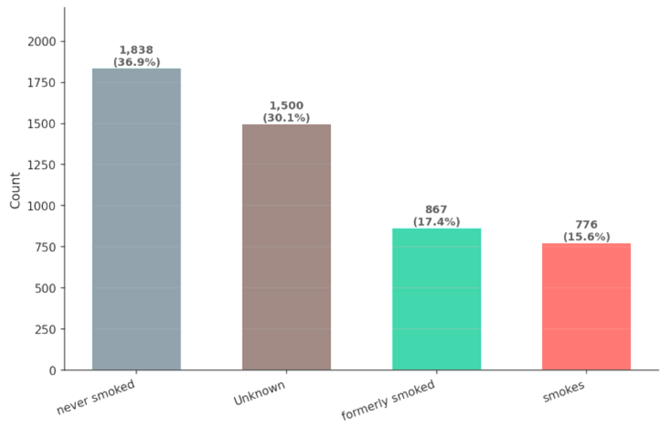

Figure 11: Distribution of smoking status in the dataset

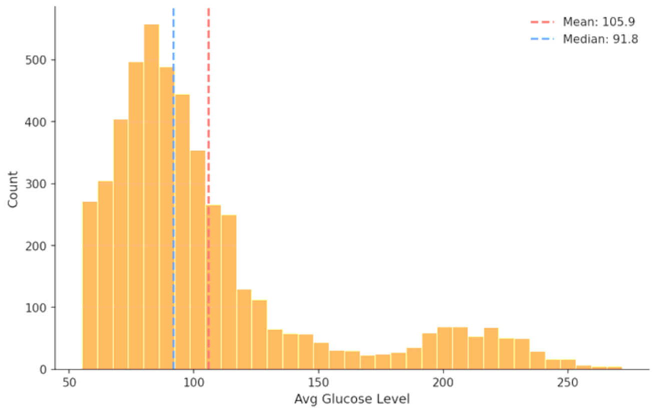

Figure 9: Distribution of avg glucose level in the dataset

Preprocessing

Data preprocessing is performed to balance the dataset for higher results. To avoid divergence from appropriate training, preprocessing is necessary before model development to eliminate undesired noise and outliers in the dataset. This step addresses any issues that prevent the model from operating more effectively. Cleaning the data and ensuring that it is prepared for model creation comes next after gathering the relevant dataset. The steps of the preprocessing are: normalization, imputation of missing values, and one-hot encoding.

The selected dataset for the stroke prediction task is incredibly unbalanced (4733 without stroke, 248 with stroke). If such an unbalanced dataset is not managed, the prediction is ineffective, and the results are inaccurate. To achieve a robust result, the dataset must be balanced. Therefore, it is necessary to deal with this unbalanced dataset first to obtain an effective model. In this study, SMOTE is employed for this reason, as it is a powerful technique [10]. It is a technique used to address an unbalanced dataset by generating new samples to overcome the unbalanced issue. After applying the preprocessing step on the dataset, we randomly split it into training and testing sets with an 80:20 ratio (80 for the training set and 20 for the testing set), as shown in Figure 12. Then the SMOTE technique will be applied to the training set.

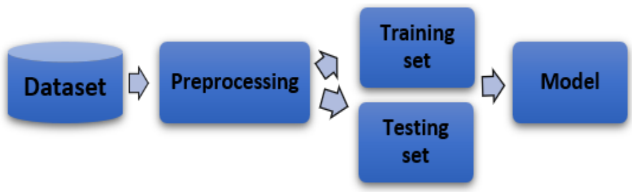

Figure 12: The proposed approach methodology

Machine Learning Approach

Nine machine learning approaches were employed in this study for stroke risk classification on one public health dataset. KNN, DT, LR, NB, SVM, AdaBoost, Gradient Boosting, ANN, and GP are the machine learning approaches that were applied to the chosen dataset for stroke risk prediction. The next section explains these nine models in more detail.

1. **KNN model:** It is an effective and simple model that is implemented for text classification. It is a supervised machine learning approach that relies on the minimal

distance between a train set and the test set [11]. The KNN shows the points of the data as vectors, using the Euclidean technique, and the distance between points is measured. K shows how many points are needed for the test dataset classification. In this study, the value of K is one, and the weighted parameters set to "distance".

2. **DT model:** It is a supervised machine learning technique that separates data into small subsets. DT is considered one of the most efficient and popular classification models [12]. The reason behind the popularity is the simplicity of use and explainability. DT forms a hierarchical structure, starting from the root node, then the branch node, and ending with the leaf node. The final classification is shown by the leaf node. One of the famous advantages of DT is that it is easy to interpret for humans. However, one of the disadvantages of the DT is a high overfitting probability. In this study, the Gini Index was used in the DT model, and the max_depth was set to 6.
3. **LR model:** It is a powerful supervised machine learning model for classification. It is a statistical model commonly used for binary classification. In LR, prediction variables can be both qualitative and quantitative, and response variables can be either qualitative or categorical. The variable Y exhibits the Bernoulli distribution, accompanied by the probability density function, regardless of whether a variable is binary or dichotomous that the response variable is divided into two categories: "success" ($Y = 1$) or "fail" ($Y = 0$) [13]. The max iteration is set to 1000 in this study, and the solver is 'lbfgs'.
4. **NB model:** It is a simple and efficient model. NB is also known as a probabilistic model, which uses the Bayes theorem. It is a popular model due to its ability to deal with high-dimensional data [14]. It follows the assumption that each predictor is independent of the others. NB produces the conditional probability of a particular objective after calculating the previous probabilities and likelihood probability for a particular objective. The type 'GaussianNB' was used here.
5. **SVM model:** It is a supervised machine learning approach used for regression and classification. By using the 'kernel trick' with high-dimensional data, this model will be an effective approach for binary classification. The 'rbf' was used as a kernel in this study with C=1.0. Finding the optimal hyperplane in the input frame that serves as the boundary between two categories is the basic concept behind SVM, one of the machine learning-based classification techniques. A hyperplane with the largest margin derived from discriminating boundaries is the most optimal hyperplane [15].
6. **AdaBoost model**: It is an ensemble machine learning approach. It combines many learners (mostly weak ones) to build a strong learner. Eventually, this model led to enhancing the performance of these multiple weak classifiers [16]. For this model, the maximum depth of the tree is set to 1, which helps in preventing overfitting. Moreover, the n-estimator is set to 100.
7. **Gradient Boosting model**: It is a famous model in machine learning regarding its ability to reach high performance scores and scalability to different issues [17]. The Gradient Boosting model is also an ensemble machine learning approach that uses decision trees for building a strong classification model by combining multiple weak classification models. In this study, the n_estimators was set to 100, and the learning_rate was 0.1 with 3 as max_depth.
8. **ANN model:** It is a powerful approach that consists of neurons aligned in layers. ANN attempts to mimic the neural network of the human brain. In an ANN, neurons are fundamental components that take in inputs and apply a specified function of transfer to generate outputs. Every neuron's input-output configuration follows a pattern [11]. This model has the ability to learn difficult patterns from a dataset and predict non-linearly. Two hidden layers with 'ReLU' activation and Adam optimizer used, which adapts learning rates automatically.
9. **GP model:** There are several stages in this model, as shown in Figure 13. It can be presented by different representations. In this study, the tree-based technique was used, with the same parameters that were used in [18] study. The strengths of the GP model are that it automatically discovers feature interactions without manual selection, well performed with limited datasets, is embedded for real-time decision-making, and provides interpretable rules, which makes it useful for healthcare specialists.

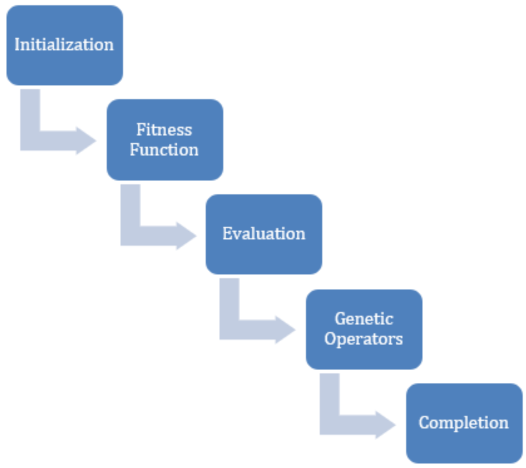

Figure 13: The stages of the GP model

The number of generations used in this study was 500, with a population of 100 individuals. The Ramped-half-and-half used as a creation method of the GP model. The termination criteria is set to be the generation number. The value of the crossover operator was 80%, while 20% for the mutation operator.

Evaluation

As an evaluation metric Accuracy was used to assess the performance of the proposed approach. Moreover, the other metrics were applied as shown in Table 3.

Discussion

To classify whether the patient can face a stroke or not, we proposed nine machine learning approaches employing SVM, LR, DT, ANN, KNN, NB, AdaBoost, Gradient Boosting, and GP**.** In order to prevent overfitting, we applied a combination of 10-fold cross-validation. This process uses various subsets of the training set for training the classifier and validating it on the subset that remained unused. This process will decrease the ability of overfitting. An overfitted model becomes unreliable in practical applications.

Table 3: The formula of the evaluation metrics.

The Metric	The Formula
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TN + FN}$
F1-score	$2\frac{Precision * Recall}{Precision + Recall}$

Where: *TP*: True Positive, reflects that both values (actual and predicted) are true. *FN*: False Negative, reflects that both values are false. A True Negative (*TN*) reveals that the actual value is True, and the predicted value is false. A false Positive (*FP*) reflects that the actual value is false, while the predicted one is true [19-22].

The majority of models achieved a high and notably consistent accuracy. This clustering of results suggests that after SMOTE balancing, most classifiers converge toward a similar decision boundary on this dataset, indicating that the feature space dominated by age, glucose level, and BMI provides relatively clear separability once class imbalance is addressed. In comparison with the proposed approaches' performance and the state-of-the-art, it shows that the GP approach is the best performing approach, achieving 95.6% Accuracy, 93.0% F1-score, and 95.6% Recall.

As demonstrated in Table 4, the GP model outperformed other classifiers in terms of accuracy, F1-score, and recall. We observed an increase in performance in all proposed approaches compared to the state-of-the-art in all aspects of the provided measurements. To achieve high results, we employed preprocessing methods and a feature selector with the GP model. The commonly utilized one-point mutation method, which replaces the root of subtree at another position randomly in an offspring with a randomly generated subtree, was used in this GP approach. Additionally, we used the conventional single-point crossover operator. The strength of the GP model is its ability to automatically construct non-linear combinations of features such as age, glucose and log (BMI) through an evolutionary symbolic search process. GP has a small but consistent advantage because these engineered interactions capture complex physiological relationships that fixed-structure models cannot leverage.

Table 4: Performance of the nine employed approaches and the state-of-the-art.

Model	Accuracy; %	F1-score %	Recall%
SVM	91.6	91.0	91.6
LR	95.1	92.8	95.1
DT	95.1	92.8	95.1
KNN	95.1	92.8	95.1
ANN	95.1	92.8	95.1
NB	95.1	92.8	95.1
AdaBoost	95.1	92.8	95.1
Gradient Boosting	95.1	92.8	95.1
GP	**95.6**	**93.0**	**95.6**
state-of-the-art [23]	91.0	90.0	94.0

In LR, DT, ANN, KNN, NB, AdaBoost, and Gradient Boosting, there were 4.1% increase in Accuracy, 2.8% increase in F1-score, and 1.1% increase in Recall. While with the SVM, it was 0.6%, 1.0% and 3.6% increase in Accuracy, F1-score, and Recall, respectively. SVM was the last classifier in the improvement list among the other eight presented models. However, all proposed approaches presented an increment in performance of the stroke risk prediction.

In this study, we used Python 3.7 for the experiment executions. While the DEAP package (distributed evolutionary algorithms in Python) was used to make the GP model.

With a 3% improvement over the state-of-the-art, the F1-score now stands at 93%, demonstrating a high level of efficiency with the GP model.

From Figure 14, it is notable that all the proposed approaches surpassed the state-of-the-art in all terms except the SVM with recall. Overall, the nine proposed approaches reached a high performance in different terms, which reflect the effectiveness of these approaches in the prediction. The rest of the proposed models perform approximately the same, and that is something worth noticing. There are reasons why the proposed models perform similarly. SMOTE is useful because it creates samples of the minority class, which makes it easier for the proposed models to make good classifications. SMOTE looks at the existing minority class samples that are closest to each other, and then it creates new samples that are somewhere in between these samples. This makes the feature space more even. Which means the complex approaches do not do much better than the simpler approaches. Some features, like age and glucose and BMI, have an effect on what the proposed models predict, Figure 15 shows the average rank of all the features among the proposed models. When these features are modified to be fair and balanced, the proposed models will understand the data well. The number of features in the dataset and the *size* of the dataset also matter. All these things together make the proposed models perform in a way. The proposed models perform similarly because of these reasons. That is what is important to remember about the performance of the proposed models.

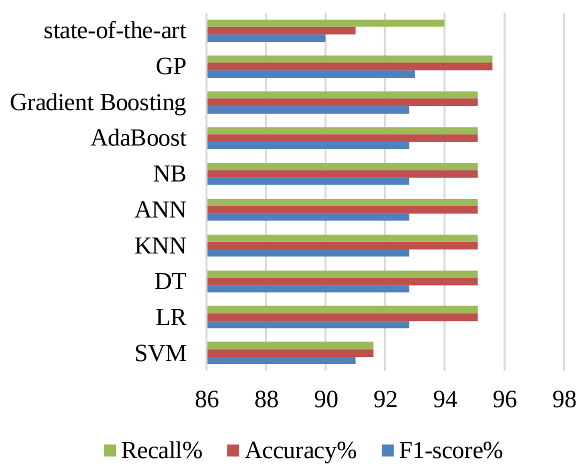

Figure 14: Recall, Accuracy, and F1-score representation

SVM recorded the lowest accuracy (91.6%), F1-score (91.0%), and recall (91.6%), falling notably below the cluster. While SVM with an RBF kernel is theoretically well-suited for non-linear classification, its underperformance here may be attributed to sensitivity to the synthetic SMOTE samples; the RBF kernel's decision boundary can be disrupted by interpolated points that do not reflect true clinical distributions. Additionally, SVM's lack of a native probabilistic framework means its threshold decisions are less flexible than probabilistic models in imbalanced adjacent scenarios.

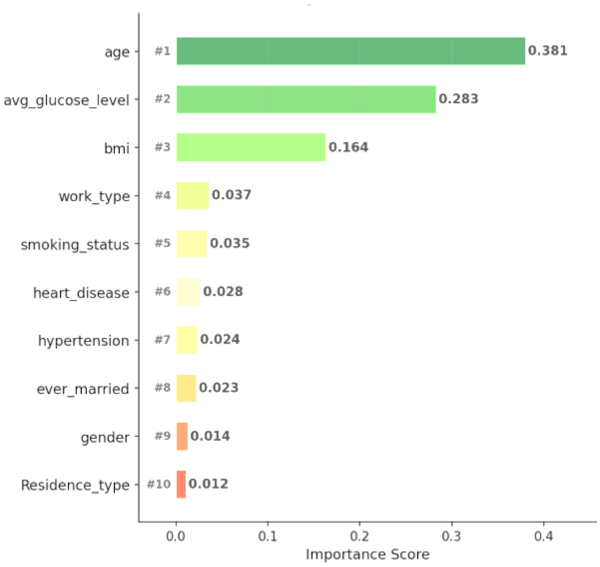

Figure 15: Average ranking of the features importance

In stroke prediction, recall (sensitivity) is arguably the most critical metric, as a missed stroke case (false negative) carries far greater clinical cost than a false alarm (false positive). From this perspective, GP and the majority cluster (both achieving ~95.1–95.6% Recall) are clinically preferable to SVM's 91.6%, which would miss roughly 1 in 12 true stroke patients.

To sum up, the experimental results confirm that the proposed GP is the most effective model for this stroke risk prediction task, leveraging symbolic feature construction to outperform all other proposed approaches. The good and steady performance shows that the SMOTE technique is really strong. The fact that SVM does not perform as well as the other models shows how important it is to make sure the model and the dataset work well together when using SMOTE. These results tell us that for predicting the risk of stroke, we can use either a GP or one of the models that work well with SMOTE to get a useful way for classification. SMOTE is a part of this, and it helps a lot with stroke risk prediction.

Conclusion

This study aimed to predict stroke risk at early stage, which helps raise awareness of the effectiveness of machine learning approaches at an early stage of stroke prediction. This study applied nine classifiers, including LR, DT, SVM, NN, NB, KNN, AdaBoost, Gradient Boosting, and GP. All the proposed classifiers well performed for stroke risk prediction. The GP was the best approach demonstrating a valid performance with (95.6%) accuracy and (93.0%) F1-score, which reflects the efficiency of the GP model. This is a great achievement by this model. Our study found that machine learning techniques can provide an effective classifier for stroke prediction at the early stage. GP offers an interpretable alternative to complex black-box models, making it valuable for healthcare applications. Moreover, it has proven to be an effective machine learning approach, especially in the medical sector.

As future work, a hybrid classifier can be used by combining GP with a boosting classifier for enhanced performance. Moreover, the proposed model can be evaluated using another dataset with a different size.

Author Contributions: "Aljero: Conceptualization and methodology, writing—original draft preparation, review and editing, data collection, results' analysis and discussion. The author has read and agreed to the published version of the manuscript."

Conflicts of Interest: "The author declare that she has no conflict of interest."

Funding: "This research received no external funding."

Data Availability Statement: "The dataset is available at https://github.com/nmelnyko/data_stroke/blob/main/data_stroke.csv."

Acknowledgements: The authors would like to express their appreciation to the Misurata University, Misurata, Libya

References

[1] C. Tsao, et al, "Heart Disease and Stroke Statistics-2023 update: A report from the American heart association." *Circulation*, vol. 8, no. 147, 2023. https://doi.org/10.1161/CIR.00000000000011

23

[2] G. Kayola, et al. "Stroke Rehabilitation in Low- and Middle-Income Countries: Challenges and Opportunities." *American Journal of Physical Medicine & Rehabilitation*, vol. 2s, no. 102, pp S24-S32, 2023. https://doi.org/ 10.1097/PHM.0000000000002128

[3] S. Xie, et al. "A Comprehensive Analysis of Stroke Risk Factors and Development of a Predictive Model Using Machine Learning Approaches." *Mol Genet Genomics*, vol. 300, no. 18, 2025. https://doi.org/10.1007/s00438-024-02217-3

[4] H. Qassim. "Early Prediction of Stroke Risk Using Machine Learning Approaches and Imbalanced Data." *NTU-JET*, vol. 4, no. 1, Mar. 2025. https://doi.org/ 10.56286/1vf19469

[5] M. Chun, et al. "China Kadoorie Biobank Collaborative Group. Stroke Risk Prediction Using Machine Learning: A Prospective Cohort Study of 0.5 Million Chinese Adults." *J. Am Med Inform Assoc*. vol. 28, no. 8, pp, 1719-1727, 2021. https://doi.org/ 10.1093/jamia/ocab068. PMID: 33969418; PMCID: PMC8324240.

[6] R. Wijaya, F. Saeed, P. Samimi, A. Albarrak, and S. Qasem. "An Ensemble Machine Learning and Data Mining Approach to Enhance Stroke Prediction." *Bioengineering*, vol. 11, no.7, pp, 672, 2024. https://doi.org/10.3390/bioengineering11070672

[7] G. Sailasya, and G. Kumari." Analyzing the Performance of Stroke Prediction Using ML Classification Algorithms." *Int. J. Adv. Comput. Sci. Appl.* 2021, vol, 12, pp, 539–545. http://dx.doi.org/10.14569/IJACSA.2021.0120662

[8] C. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John. "Predicting Stroke from Electronic Health Records." *In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE*, 2019. https://doi.org/ 10.1109/EMBC.2019.8857234. PMID: 31947147

[9] DataHack Available online: https://github.com/nmelnyko/data_ stroke/blob/main/data_stroke.csv. (accessed on 15 January 2026).

[10] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique." *J. Artif. Intell.Res*., vol, 16, pp, 321–357, 2002. https://doi.org/10.1613/jair.953

[11] I. Ahmad, H. Mabuchi, M. Kano, S. Hasebe, Y. Inoue, and H. Uegaki, "Data-Based Fault Diagnosis of Power Cable System:

Comparative Study of k-NN, ANN, Random Forest, and CART.” *In Proceedings of the 18th IFAC World Congress, Milano, Italy, 28 August–2 September 2011*, pp, 12880–12885, 2011. https://doi.org/ 10.3182/20110828-6-IT-1002.01761

[12] M. Suciu, and R. Lung. “A Nash equilibria decision tree for binary classification.” *Appl Intell*, vol. 55, no. 192, 2025. https://doi.org/10.1007/s10489-024-06132-3

[13] A. Solimun, and A. Fernandes. “Ensemble Bagging Discriminant and Logistic Regression in Classification Analysis.” *New Mathematics and Natural Computation*, vol. 21, no. 01, pp, 91-111, 2025. https://doi.org/ 10.37394/23203.2021. 16.64

[14] O. Peretz, M. Koren, and O. Koren. “Naive Bayes classifier–An Ensemble Procedure for Recall and Precision Enrichment.” *Engineering Applications of Artificial Intelligence*, pp, 136, 108972, 2024. https://doi.org/10.1016/j.engappai.2024.108972

[15] F. El-Batta, et al. “AI-Based Monitoring of Solar Panels in Desert Environments: Distinguishing Dust Accumulation for Fault Detection.”. *Wadi Alshatti University Journal of Pure and Applied Sciences*, pp, 340-353, 2026. https://doi.org/10.63318/waujpasv4i1_38

[16] W. Lee, M. Cheon, C. Hyun, and M. Park. “Distance Sensitive AdaBoost Using Distance Weight Function.” *Int Journal of Fuzzy Logic and Intelligent Systems*, vol. 12, no. 2, pp, 143-148, 2012. https://doi.org/ 10.5391/IJFIS.2012.12.2.143

[17] J. Mushava, and M. Murray. “Flexible Loss Functions for Binary Classification in Gradient-Boosted Decision Trees: An Application to Credit Scoring. Expert Systems with Applications.” 238, 121876, 2024. https://doi.org/10.1016/j.eswa.2023.121876

[18] M. Aljero. “Development of Autism Spectrum Disorders Classification Model for Toddlers Using Machine Learning Technique.” *Journal of the Faculty of Education Tripoli*, vol. 1, no. 22, pp, 146–156, 2025. https://doi.org/10.5281/zenodo. 18182345

[19] A. Rezvantalab, H. Safigholi, and S. Karimijeshni. “Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms,”, 2018, arXiv preprint arXiv:1810.10348

[20] L. Ben Dalla, O. Karal, M. EL-Sseid, and A. Alsharif. "An IoT-Enabled, THD-Based Fault Detection and Predictive Maintenance Framework for Solar PV Systems in Harsh Climates: Integrating DFT and Machine Learning for Enhanced Performance and Resilience." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 1, pp. 41-55, 2026. https://doi.org/10.63318/waujpasv4i1_05

[21] S. Alfathi, G. Miskeen, and W. Mremi. "Evaluation and Prediction Performance of Solar Panel and Wind Turbine Systems Using Simulation." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 1, pp. 94-104, 2026. https://doi.org/10.63318/waujpasv4i1_10

[22] E. Almhdi, and G. Miskeen, “Power and Carbon Footprint Evaluation and Optimization in Transitioning Data Centres”, *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 3, no. 2, pp. 221-229, 2025. https://doi.org/10.63318/waujpasv3i2_28

[23] N. Melnykova, et al. “Machine Learning for Stroke Prediction Using Imbalanced Data.” *Sci Rep* 15, 33773, 2025. https://doi.org/10.1038/s41598-025-01855-w