Journal of Computer Science

Original Research Paper

Stroke Risk Assessment with Classical ML Methods

[1]Mohammad Aref Alshraideh, [2]Najwan Alshraideh, [3]Abedalrahman Alshraideh, [2]Yara Alkayed, [4]Yasmin Al Trabsheh, [2]Heba Alshraideh and [5]Bahaaldeen Alshraideh

[1]Department of Artificial Intelligence, the University of Jordan, Jordan
[2]Department of Internal Medicine, the University of Jordan, Jordan
[3]Internal Medicine, East Midlands Deanery, NHS, England, UK
[4]Clinical Attache, United Lincolnshire Hospitals, NHS, England, UK
[5]Department of Special Surgery, Division of Urology, the University of Jordan, Jordan

Article history
Received: 19-08-2024
Revised: 02-10-2024
Accepted: 18-10-2024

Corresponding Author:
Department of Artificial Intelligence, the University of Jordan, Jordan
Email: mshridah@ju.edu.jo

Abstract: Stroke, often caused by a disruption in the supply of essential oxygen, blood, and nutrients to the brain, represents a significant global health challenge. Due to limited resources, developing countries like Ethiopia face unique obstacles in identifying and treating strokes. This study explores the potential of Machine Learning (ML) techniques to predict stroke risk and facilitate early detection and intervention. By doing so, it aims to reduce the burdens of disability, mortality, and healthcare costs associated with strokes. In this research, we utilized four machine learning models: Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest Classifier (RFC). These models were developed using a dataset from Kaggle, which contained information from 5,110 individuals and various attributes related to each person. Our methodology followed a systematic approach that included data understanding, preparation, experimentation, rectifying inconsistencies, removing duplicates, and resolving errors within the dataset. The ML models were created and rigorously assessed within the Anaconda Python programming environment, with performance evaluation conducted through Confusion Matrix analysis. Our findings revealed that the Random Forest Decision Tree classifier outperformed the others, boasting an accuracy rate of 99.3%. The support vector machine was closely behind at 96.63%, while the k-nearest neighbor and stochastic gradient descent achieved acceptable accuracy. Consequently, we recommend the utilization of the random forest decision tree classifier for further stroke risk prediction endeavors.

Keywords: Stroke Risk, Stroke Risk Prediction, Random Forest, SGD, SVM

Introduction

A stroke is a medical condition that can cause damage to brain cells, leading to disabilities or even death. A stroke, also known as a cerebrovascular accident, is a neurological disorder that can occur due to either ischemia or hemorrhage in the brain's arteries. It typically causes many motor and cognitive impairments, significantly impacting functionality. Globally, stroke affects approximately 16 million individuals annually and is associated with substantial societal costs.

Strokes can occur due to blood vessel blockage by clots or rupture, preventing the brain from receiving essential oxygen and nutrients (Party, 2012).

The World Health Organization (WHO) indicates that` stroke ranks as the second primary cause of mortality globally, responsible for approximately 11% of all deaths (Katan and Luft, 2108). However, up to 80% of strokes can be prevented through early prediction (Gaines *et al*., 2015).

Given the high fatality rate associated with strokes and the economic burden they impose, efforts to reduce their occurrence are essential. For instance, stroke-related healthcare costs in the USA amounted to $193.1 billion from 2011-2012, with an additional $123.5 billion lost in productivity due to premature deaths (Mozaffarian *et al*., 2016).

Traditional stroke prevention methods include periodic blood tests, a balanced diet, exercise, smoking cessation, alcohol moderation, stress management, and relaxation techniques. While these methods are widely acknowledged for their health benefits, many individuals

struggle to adhere to them due to a lack of motivation or time constraints.

Hence, there is a growing need for nontraditional methods to reduce stroke risk. One promising approach is harnessing machine learning to predict strokes based on individual health information. Such a system could empower individuals to take preventive actions without needing medical consultations, clinics, or costly healthcare expenses, which is particularly valuable for underserved populations.

A machine learning-based stroke prediction system could be integrated into mobile apps or websites, allowing people to input personal information and receive stroke risk assessments from the comfort of their homes. Additionally, healthcare professionals could use this tool alongside other diagnostic tests to enhance stroke detection.

Machine learning has gained significant traction in diagnosing and predicting various medical conditions, such as skin cancer (Ahmad *et al*., 2020; Alshraideh, 2020; Farhan *et al*., 2015; Salah *et al*., 2011), heart disease (Ul-Haq *et al*., 2018); (Shboul *et al*., 2022) and Parkinson's disease (Alshraideh *et al*., 2024). However, stroke prediction in machine learning is still in its infancy, with some prior research efforts (Mahesh *et al*., 2020).

This study aims to identify an effective machine-learning model for foreseeing stroke risk. The model will provide individuals with necessary precautions, medications, and lifestyle recommendations to reduce the probability of stroke. It will also help manage high blood pressure and address other risk factors associated with stroke.

This study addresses a supervised binary classification task. It is supervised because the Dataset used is labeled and its binary classification predicts either a '1' for individuals at risk of stroke or a '0' for others.

I utilized a dataset from Kaggle (2024) to achieve this goal. It consists of information on 5110 male and female individuals with various features related to each person. Further details about these features will be presented in section III. The Dataset's' stroke' feature is the training and testing label.

Figure (1) demonstrates the most recent worldwide health estimates from 2000-2016, emphasizing ischemic heart disease and stroke as the top contributors to mortality and disability (Heart *et al*., 2015). The American Heart Association identifies stroke as a significant health issue because of its elevated mortality rate (Salah *et al*., 2011). Moreover, the expense associated with stroke-related hospitalizations is on the rise (Johnson *et al*., 2016). Consequently, there is an increasing need for advanced technologies to support clinical diagnosis, treatment, event prediction, effective therapeutic recommendations, and the design of rehabilitation programs (Katan and Luft, 2108).

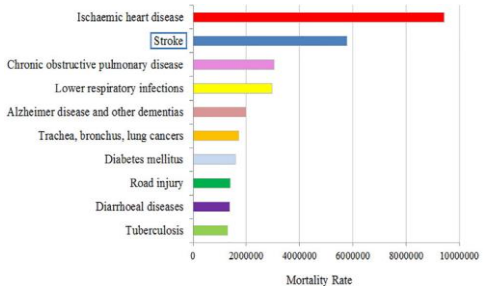

Fig. 1: This figure shows estimated mortality rankings by a factor based on data recorded by (Karimi *et al*., 2021)

Early stroke detection plays a pivotal role in efficient treatment, and Machine Learning (ML) can offer invaluable assistance in this regard. To achieve this, Machine learning is a transformative technology that enables healthcare professionals to make well-informed medical decisions and precise forecasts. In recent decades, extensive research has been devoted to utilizing machine learning for stroke diagnosis, focusing on enhancing accuracy and efficiency.

To optimize results, I experimented with more than four machine-learning models, including Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest Classifier (RFC). I performed comprehensive data preprocessing and fine-tuned Hyperparameter to optimize accuracy.

Related Work and Background

A stroke is a vascular abnormality that occurs in the brain and can result in neurological symptoms such as muscle weakness, numbness, and, in severe cases, potential fatality. The World Health Organization (WHO) defines a stroke as a "sudden disturbance of cerebral function lasting more than 24 h or resulting in death, without an evident non-vascular cause". Strokes are categorized into two primary types: Ischemic and hemorrhagic.

Ischemic strokes occur when blood flow to the brain cells is obstructed, causing damage and eventual cell death. This blockage can be due to a blood clot in a blood vessel, known as an ischemic stroke, or a rupture within a blood vessel. Stroke risk factors encompass lifestyle elements such as diet, smoking, obesity, physical inactivity, and alcohol consumption, along with family history, genetics, age, gender, drug use, race, oral contraceptive use, geographic location, seasonal and climatic conditions, and socioeconomic influences. Medical conditions like atrial fibrillation and high blood pressure also play a significant role in stroke risk. In Addis Ababa, the capital of Ethiopia, approximately 68% of adults are estimated to have one or more risk factors for cardiovascular disease. Parkinson’s disease arises from the loss of dopamine-producing brain cells, leading to motor symptoms after 60-80% of these cells are lost. Researchers are seeking early non-motor indicators to halt disease progression before movement symptoms appear. Using machine learning, the study found that support vector machines with recursive feature

elimination achieved a 93.84% accuracy in diagnosing Parkinson's disease with a minimal set of voice features (Karapinar Senturk, 2020).

Hemorrhagic strokes are divided into two primary types: Intracerebral hemorrhages, often associated with conditions like hypertension, cerebral amyloid angiopathy, or degenerative arterial disease, and subarachnoid hemorrhages, typically caused by the rupture of an aneurysm. Key risk factors for hemorrhagic stroke include advanced age, heavy alcohol consumption, and hypertension. In younger individuals, cocaine use is a significant risk factor for cerebral hemorrhage. Common symptoms of hemorrhagic stroke include focal neurological deficits, vomiting, drowsiness, neck stiffness, and seizures.

Neurological symptoms of an ischemic stroke typically appear suddenly but can also develop gradually in some cases; often referred to as a "stroke-in-progress," the specific symptoms of an ischemic stroke vary depending on the location of the blockage and the collateral blood flow. However, weakness on one side of the body (hemiparesis) is a joint presentation, particularly in older individuals. Atherosclerotic ischemic strokes typically occur abruptly without prior warning and are more frequent in older populations. It is worth noting that ischemic strokes constitute around 80% of all strokes and are on the rise in developing countries due to unhealthy lifestyles (Salah *et al*., 2011).

Business Understanding

Comprehending the business objectives and requirements from a business perspective is paramount when leveraging data mining techniques to transform data into actionable knowledge. As a result, we have conducted a thorough exploration and highlighted vital insights that will aid us in comprehending, defining, and analyzing the problem at hand most effectively.

Classical machine learning (Classical ML) refers to the traditional or conventional set of machine learning techniques and algorithms that have been in use for several decades. These techniques are typically based on statistical principles and mathematical models to make predictions or decisions based on data.

Classical ML includes a variety of algorithms and methods, such as.

Linear regression: Used for predicting a continuous numerical value based on one or more input features. It is often used in regression problems.

Logistic regression: This is primarily used for binary classification problems. The goal is to predict one of two classes (e.g., yes/no, true/false).

Decision trees are tree-like structures used for classification and regression tasks. They make decisions by splitting data based on feature values.

Random forest: An ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.

Support Vector Machines (SVM): A supervised learning algorithm for classification and Regression. It finds the optimal hyperplane that best separates data points.

K-Nearest Neighbors (K-NN) is a simple yet effective algorithm for classification and Regression that makes predictions based on the majority class or the average of the k-nearest data points.

Naïve bayes: A probabilistic algorithm based on Bayes' theorem commonly used for classification tasks, particularly in natural language processing.

K-Means clustering: An unsupervised learning algorithm used for clustering and pattern recognition.

Principal Component Analysis (PCA): A dimensionality reduction technique used to reduce the number of features while retaining the most essential information.

Gradient boosting: This is an ensemble learning method that constructs a sequence of decision trees, with each tree iteratively improving predictions by correcting errors from previous iterations. Classical ML techniques are well-established and widely used in various fields, including healthcare, finance, natural language processing, and image analysis. While they remain relevant and practical, recent advancements in deep learning have introduced more complex and robust approaches, sometimes collectively referred to as "modern" or "deep learning" techniques. These newer methods have gained popularity in solving complex tasks involving large datasets, but classical ML remains valuable for many practical applications, especially when interpretability and transparency are critical.

Machine Learning (ML) techniques have gained popularity in medical prediction for their remarkable accuracy. Researchers have employed various ML models in this domain.

In a study referenced by Mahesh *et al*. (2020), the research team employed Decision Tree models, Naïve Bayes models, and Neural Networks for stroke prediction. Their Dataset shared features similar to our study. The researchers constructed and trained these three models after data preprocessing, including data cleaning and Encoding into numerical values. Impressively, they achieved satisfactory levels of accuracy.

Another notable work cited in Nwosu *et al*. (2019) involved using electronic medical records, albeit with highly imbalanced data. To address this challenge, the researchers employed resampling techniques. Their study included decision trees, random forests, and multilayer perceptrons as part of their model selection. The Multilayer Perceptron yielded the highest accuracy at 75.02%.

In Goyal (2017), researchers explored the potential of Deep Neural Networks in stroke prediction. They compared the performance of this approach with that of a Support Vector Machine (SVM) and the Naïve Bayes model. Their study introduced an Integrated Machine Learning Approach to enhance stroke prediction."

Recent studies have delved into the application of machine learning and statistical models for predicting stroke risk, offering valuable insights into this critical area of healthcare. Here, we provide a concise summary of the key findings from these studies.

In Kansadub *et al.* (2015), research used demographic data to predict stroke occurrence. The study compared the performance of three machine-learning classification algorithms: Decision Tree, Naïve Bayes, and neural network. Decision tree emerged as the most accurate model, boasting an accuracy rate of 0.75. Neural networks, while slightly less accurate, demonstrated superiority in terms of safety by minimizing False Positives (FP). Understanding the crucial balance between accuracy and safety in stroke prediction is vital for patient outcomes.

Karimi *et al.* (2021), aimed to develop a machine learning-based approach for predicting stroke risk in individuals displaying symptoms or risk factors. The study employed a Support Vector Machine (SVM) with various kernel functions. The linear kernel function stood out with the highest accuracy, achieving 91%. The research suggested expanding this methodology to larger datasets to enhance performance further. The success of SVM in stroke prediction underscores its potential in early diagnosis and risk assessment.

Dritsas and Trigka, (2022) developed a comprehensive framework for long-term stroke risk prediction utilizing a range of machine learning algorithms. The combination of stacking machine learning and Random Forest models demonstrated superior performance, with high accuracy, sensitivity, specificity, Precision, and F1 score. The study highlighted the importance of utilizing diverse models to address the complex task of stroke prediction effectively.

Zheng *et al.* (2015), study aimed to create a predictive risk model for stroke over one year, leveraging Electronic Medical Records (EMR) and clinical notes. Logistic regression models yielded high c-statistics for retrospective (0.892) and prospective (0.887) predictions. The model's effectiveness in identifying stroke risk within a large, independent cohort underscores its potential for real-world applications. Implementing this model in real-time population monitoring platforms can provide healthcare providers with early warnings, enable timely intervention, and improve patient outcomes.

Their study published in Sultan *et al*. (2017) conducted a cross-sectional survey at Ethiopia's emergency center, focusing on stroke types, risk factors, and clinical presentations. They found that hemorrhagic strokes were more prevalent than ischemic strokes, which contrasts with patterns observed in more developed countries. Common risk factors identified in the study include hypertension, cardiac disease, and diabetes, highlighting the importance of managing these conditions to reduce the risk of stroke. The study also shed light on the challenges and opportunities for stroke risk assessment in Low and Middle-Income Countries (LMICs), where healthcare resources may be limited.

In summary, these studies collectively highlight the promise of machine learning and statistical models in stroke prediction. They offer diverse approaches and valuable insights into understanding and mitigating stroke risks. Further research in this domain is encouraged to enhance accuracy and extend the applicability of these models, especially in resource-constrained healthcare settings. The findings underscore the potential for early stroke risk identification and timely intervention to improve patient outcomes.

Methods and Data Handling

Figure (2) illustrates the proposed architecture for a dataset used to predict the risk of stroke in individuals. The architecture consists of the following steps:

1. Data collection: Patient data is gathered from web sources and organized into a dataset containing information on individuals with and without a history of stroke
2. Data preprocessing: The Dataset is cleaned and preprocessed to ensure that all attributes are in a consistent format
3. Data splitting: The Dataset is split into a training dataset (70%) and a test dataset (30%). The training dataset is used to train the stroke prediction model and the test dataset is used to evaluate the model's performance.
4. Model training: The training dataset is used to train three different classification models: Stochastic Gradient Descent (SGD), Random Forest Classifier (RND), and K-Nearest Neighbor and Vector Machine (SVM). The model with the best performance is selected as the final stroke prediction model.
5. Model evaluation: The test dataset is utilized to assess the performance and accuracy of the chosen stroke prediction model
6. Stroke risk prediction: The stroke prediction model predicts the risk of stroke in new patients

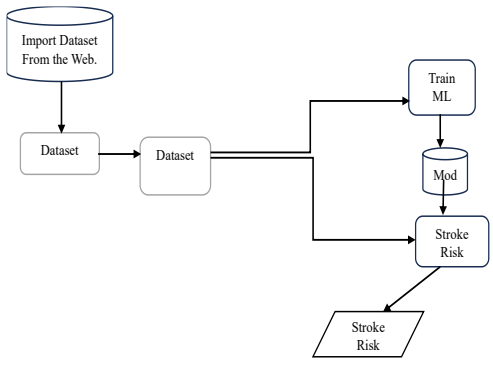

Fig. 2: The proposed architecture for the stroke prediction model

The proposed architecture is a simple and effective way to predict the risk of stroke in individuals. It can be used to identify high-risk patients who can benefit from preventive interventions.

The Dataset used in this study predicts whether a human will likely get a stroke based on several input attributes mentioned in detail below. Each row in the Dataset provides relevant information about one person. The Dataset contains 5110 observations with 12 attributes and is from Kaggle (2024).

Table (1) describes the initial set of the twelve attributes and a brief meaning of each attribute.

Body Mass Index (BMI) is a metric used to estimate body fat based on an individual's height and weight. It is widely employed to determine whether a person's weight is appropriate for height. BMI is calculated by dividing a person's weight in kilograms by the square of their height in meters.

Higher BMI values typically indicate a more significant amount of body fat. People with a high BMI are at increased risk for specific health problems, such as heart disease, stroke, type 2 diabetes, and some types of cancer.

BMI is used to broadly define different weight groups in adults 20 years old or older. Table (2) shows the other weight groups based on BMI.

Sirsat *et al*. (2020) incorporated additional features, such as atrial fibrillation, lifestyle factors, and others, alongside the features used in this project to enhance the predictive capabilities of their models.

After I explained the initial structure of the Dataset, it was time to explain the steps I took to process the data to achieve the best-targeted results. Data processing is required in most machine-learning projects because it yields better results. Additionally, data processing is needed to prepare the data for training. For example, I should convert some categorical attributes into numerical values. Moreover, sometimes, some attributes need to be added to get a better correlation, or some non-useful qualities are dropped, among other steps.

The steps taken in data processing depend on the Dataset under study. It differs from one Dataset to another. Also, some researchers may take different steps in different orders to get the best results.

First, the ”id” attribute does not affect the results because each person has a unique id value. Therefore, it is better to drop this column from the dataset.

Second, the “BMI” attribute has missing data. This can be handled in several ways, like dropping columns or rows or filling in the missing data. I adopted filling in the missing data due to the importance of such attributes and because the Dataset is relatively tiny. I filled in the missing values by the mean value of the BMI.

Next, the “gender” attribute is well distributed between males and females. However, an odd value called “other” was listed only once. Therefore, I filtered out this value.

Table 1: An explanation of features with brief descriptions

ID	It is a unique identifier of the patient
Gender:	Whether "Male," "Female," or "Other." It looks like "other" can be an odd value to be filtered later because we have only one person with the gender “other.” The data has 2994 females and 2115 males
Age:	Age of the patient. Has continuous ages up to 82 years
Hypertension:	The values are either 0 (count = 4612) if the patient does not have hypertension or 1 (count = 498) if the patient has hypertension
Heart_disease	: The values are either 0 (count = 4834) if the patient has no heart disease or 1 (count = 276) if the patient has a heart disease
Ever_married:	the values are either "No" (count = 1757) or "Yes" (count = 3353)
Work_type:	The possible values are "children" (count = 687), "Govt_job" (count = 6570) for those who work in the government, "Never_worked" (count = 22), "Private" (count = 2925) for those who work in the private sector, and "Self-employed" (count = 819) for those who work by themselves and not employed by others
Residence_type:	"Rural" (count = 2514) or "Urban" (count = 2596)
Avg_glucose_level	: average glucose level in blood. The normal glucose level is less than 100 mg per deciliter Kaggle (2024)
BMI:	It is the body mass index. In general, this attribute is for people 20 years old and older, and it is translated into four categories, and these categories are the same for men and women, as shown in
Smoking_status:	The possible values are "formerly smoked" (count = 855), "never smoked" (count = 1892), "smokes" (count = 789), or "Unknown" (count =1544). The Unknown status means that the information is unavailable for this patient
Stroke:	The possible values are either 1 (count = 249) if the patient had a stroke or 0 (count = 4861) if not. This is the label of the data that will try to predict

Table 2: Body mass index

BMI	Weight status
Below 18.5	Underweight
18.5–24.9	Normal or healthy weight
25.0–29.9	Overweight
30.0 and Above	Obese

Furthermore, in the medical field, there are several attributes whose exact value is not distinguishable from other values except by the range in which these values fall. For example, when somebody does a blood test for Vitamin D, the result of the test is a value *X*. Also, a description explains several ranges: Below 10 ng/mL is considered low, (10-30) ng/mL is deemed insufficient, (30-150) ng/mL is considered sufficient and above 150 ng/mL is considered toxic. Therefore, the values 35 and 40 are normal and have the same meaning (Heart *et al*., 2015). Also, in such cases, instead of keeping the values as they are, I divide them into four ranges.

In the dataset under study, I have several attributes that will be handled, like the vitamin D example above, which includes glucose level, BMI, and age. Below is the description of each of them.

Regarding the glucose level, I divided it into three ranges: Low for levels below 90, typical for the range (90-160), and above 160, which is high. This means that, from a medical point of view, glucose levels 90 and 95 have the same meaning and will have the same stroke probability.

Regarding the BMI values, according to the values reported by Shafer *et al*. (2009), I divided them into four ranges: values (0-19) are considered underweight, values (19-25) are considered ideal, values (25-30) are considered overweight and values (30 and above) are considered obese.

Regarding the age attribute, the medical field considers the ages as ranges. Therefore, I divided the ages into four ranges: The ages (0-18) are considered children, the ages (18-45) are considered adults, the ages (45-60) are considered adults and the ages (60 and above) are considered elderly.

As Table (3) shows, there is a significant imbalance in the distribution of the label we are studying (i.e., stroke). In such cases, we need to solve the issue. This can be done using over-sampling techniques.

In this context, the term "label" signifies whether an individual has had a stroke, with "1" indicating they have not and "0" meaning they have.

The over-sampling step is performed using the SMOTE method. SMOTE selects a sample from the dataset and identifies its kkk nearest neighbors within the feature space.

The newly generated point is determined by drawing a vector between one of the kkk neighbors and the current data point. After that, it multiplies this vector by a random number x, which lies between 0 and 1. The result will be added to the current data point to create the new data point. The result after over-sampling is shown in Table (4), where several points with label one equal the number of points with label 0. The resulting data after oversampling contains 9720 rows, as shown in Table (4).

To make the data suitable for most machine learning algorithms, categorical attributes in the data frame were transformed into numerical ones. Two techniques were employed for this purpose: One-Hot Encoding and Ordinal Encoding. The key distinction lies in how they handle categorical attributes.

Table 3: Data Stroke load distribution

Label	Count
0	4861
1	249

Table 4: Distribution of stroke data after applying over-sampling techniques

Label	Count
0	4860
1	4860

One-Hot encoding: This method converts a categorical attribute into multiple attributes equal to the number of categories within the attribute. These new attributes are binary, representing the presence or absence of each category.

Ordinal encoder: In contrast, the Ordinal Encoder transforms the categorical attribute into a single numerical attribute with values corresponding to the categories. The original categorical attributes are retained, but a single numerical attribute replaces them with distinct values for each category.

These encoding techniques are pivotal in data preparation for machine learning algorithms that require numerical input. They enable the effective utilization of categorical information while ensuring compatibility with the selected models.

The process of One-Hot encoding was explicitly applied to eight attributes, including three binned attributes: 'avg_glucose_level,' 'BMI,' and 'age,' along with five original categorical attributes: 'smoking_status,' 'ever_married,' 'Residence_type,' 'gender,' and 'work_type.' To implement this, we utilized the 'get_dummies' function from the Pandas Library to convert the data into binary values (0 and 1). This transformation expands the Dataset to include 30 columns, each representing a specific category within the encoded attributes. By doing so, One-Hot Encoding enhances the model's ability to capture the precise aspects of the data while maintaining its numerical compatibility.

Conversely, we also experimented with applying Ordinal Encoding to the data. This approach was considered because One-Hot Encoding substantially increased the number of attributes by 250%, resulting in almost zero values. Ordinal Encoding might offer a more efficient alternative. However, the outcomes from the One-Hot Encoded data were superior.

It is worth noting that One-Hot Encoded data consists entirely of binary values (0 and 1), whereas Ordinal Encoded data has a narrower range of values. We attempted scaling the Ordinal data using a standard scaler to explore potential improvements. Nonetheless, the results achieved through One-Hot Encoding remained superior to those obtained from Ordinal Scaling.

Although the Kaggle dataset, with 5110 observations and 12 attributes, provides a solid foundation for stroke risk prediction, several limitations must be considered to ensure the model's validity and generalizability:

- Class imbalance: The dataset exhibits a significant class imbalance, with 4861 non-stroke cases and only 249 stroke cases. This imbalance may lead to a model biased toward predicting non-stroke cases, resulting in more false negatives. Although we used the Synthetic Minority Over-sampling Technique (SMOTE) to address this imbalance, such methods have limitations and may introduce synthetic data that could affect the model's interpretability.

- Demographic representation: The dataset may not adequately represent all demographic groups, such as different age brackets, socioeconomic statuses, or ethnic backgrounds. If the dataset over-represents certain groups (e.g., males vs. females or urban vs. rural populations), the model may perform better for those groups and poorly for others. This could lead to biased predictions when applied to populations not well-represented in the training data, potentially putting these populations at risk. The dataset's limited geographic and demographic scope can undermine the model's generalizability in real-world settings.

Future work must focus on acquiring more comprehensive datasets to address these limitations. These should encompass diverse populations, additional risk factors, and longitudinal data. Furthermore, the application of fairness-aware learning methods.

Model Evaluation

This section describes how the Stroke Risk Prediction model was evaluated.

Cross-validation is a robust technique for assessing the model's generalization capability and preventing overfitting. This study used k-fold cross-validation to ensure the stroke prediction model performs well on unseen data.

K-fold cross-validation process: The dataset was split into k equal subsets (folds), where k = 10. The model was trained on k-1 folds and the remaining fold was used for testing. This process was repeated k times, with each fold serving as the test set exactly once. The final model performance was then averaged across all folds. By using this technique, we ensured that the model was evaluated on the entire dataset without relying on a single train-test split. This approach minimizes the risk of the model overfitting to the training data and offers a more reliable estimate of its performance on unseen data. Cross-validation ensures that every data point is used for training and testing, reducing bias in the model evaluation. Also, the variance of model performance is reduced compared to a single train-test split, providing a more reliable estimate of generalization.

While cross-validation was the primary validation technique, we also implemented a hold-out validation strategy. This method divides the dataset into 70% for training and 30% for testing. The training set is utilized to build and train the model, while the test set is reserved for the final evaluation of the model's performance after optimization through cross-validation.

We used a confusion matrix (Karimi *et al*., 2021) to assess the performance of a classification model. This is a standard tool for measuring a classification model's performance on a test dataset with known actual values (Karimi *et al*., 2021).

Confusion matrices are crucial assets in predictive analysis within machine learning. They summarize a classifier's correct and incorrect predictions, especially in binary classification tasks.

The model aims to predict one of two potential outcomes in binary classification tasks. For instance, in a binary classification scenario regarding stroke prediction, the task could involve determining whether a patient has had a stroke.

The Confusion Matrix consists of four elements:

- True Positives (TP): The count of cases where the model accurately identifies a positive outcome
- False Positives (FP): The count of cases where the model mistakenly identifies a positive outcome
- True Negatives (TN): The count of cases where the model accurately identifies a negative outcome
- False Negatives (FN): The count of cases where the model mistakenly identifies a negative outcome

The Confusion Matrix can be used to compute various metrics that evaluate the performance of a classification model. One of the most commonly used metrics is.

Accuracy: This measures the percentage of correct predictions the model makes. It is calculated as the ratio of correctly classified instances (True Positives + True Negatives) to the total number of instances (True Positives + True Negatives + False Positives + False Negatives). This metric provides an overall measure of the model's effectiveness in classification tasks:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: The percentage of correct optimistic predictions. This determines whether the model is reliable or not:

$$Precision = \frac{TP}{TP + TN}$$

- Recall in machine learning refers to the percentage of actual positive instances correctly identified by the model. A higher recall value suggests that the model correctly identifies the most favorable cases, resulting in more true positives and potentially more false positives, leading to lower overall accuracy. Conversely, a lower recall value indicates more false negatives, where cases that should be identified as positive are incorrectly labeled as negative. In practical terms, higher Recall means the model is better at correctly identifying positive cases. In comparison, lower Recall implies the model may miss positive cases, being more specific when labeling a case as positive:

$$Recall = \frac{TP}{TP + FN}$$

A harmonic mean of precision and recall, the F1-score reaches its maximum when precision equals recall.

$$F1\text{-}score = \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}}$$

The F1 score can be challenging to interpret, so it is often combined with other evaluation metrics to provide a more complete picture. Without this, it can be challenging to determine whether the classifier is optimizing for Precision or Recall. To prevent overfitting, several techniques were employed during model development:

- Regularization: Techniques like L2 regularization were applied in the Stochastic Gradient Descent (SGD) model to prevent it from becoming too complex and overfitting the training data
- Early stopping: For models like Stochastic Gradient Descent (SGD), we implemented early stopping to halt training if performance on the validation set did not improve after a certain number of epochs

Experiment Setup

To ensure a successful experiment, it is crucial to establish a well-equipped environment that will aid your machine-learning endeavors. Here is a breakdown of the setup you need to consider.

Programming Language Selection: It is recommended that you choose a suitable programming language like Python, which provides a wide range of machine learning libraries and frameworks. Python libraries such as sci-kit-learn, TensorFlow, and PyTorch offer the necessary tools and functions for modeling and training your machine-learning algorithms.

Integrated Development Environment (IDE) or Text Editor: Set up an IDE or text editor to streamline the coding and debugging processes. This will enhance your efficiency in developing and fine-tuning your models.

The implemented machine specifications include an HP computer type with 8 GB RAM, a Core i5 processor, and the Windows 10 operating system.

Results and Discussion

This section presents and discusses the achieved results. Indeed, this study shows several tuning steps, but not all of them. Several machine-learning models have also been tried on this data.

As an initial step, our objective was to determine whether to employ One-Hot Encoding or Ordinal Encoding. This decision stemmed from the observation that One-Hot Encoding expanded the Dataset to encompass 30 attributes, with many primarily populated by zeros. Consequently, Ordinal Encoding might offer a more efficient approach.

Multiple versions of the Dataset were generated to arrive at the ultimate decision. Specifically, the Dataset labeled Version 2 was subjected to One-Hot Encoding, the Dataset labeled Version 3 underwent Ordinal Encoding and the Dataset labeled Version 4 was Ordinal Encoded and subsequently scaled using standard scaling. This comparative approach allowed for a thorough evaluation of the performance of the encoding techniques.

Both Stochastic Gradient Descent (SGD) and Support Vector Machine (SVM) were employed on these datasets, consistently demonstrating the superiority of the One-Hot Encoding method. Consequently, our exclusive focus will be on the One-Hot Encoded dataset for all subsequent experiments.

As a second step, we addressed the imbalance issue in the stroke attribute. We experimented with both the One-Hot Encoded data without oversampling and with oversampling. The results indicated a significant enhancement in performance when oversampling was applied. Therefore, oversampling will be consistently employed for the remaining results and analyses.

After implementing the data processing steps outlined in section III, we adopted One-Hot Encoding and oversampling. Additionally, our approach employed a training/test split ratio of 30% for testing and 70% for training.

From a clinical perspective, these findings present the potential for developing decision-support tools to help healthcare professionals identify high-risk patients early. By enhancing stroke prediction models, clinicians may be able to introduce targeted interventions for at-risk individuals, thus reducing the incidence of stroke. The model's 95% accuracy indicates its suitability for real-world scenarios where early detection is crucial for stroke prevention and management.

Comparing classification algorithms

I conducted a thorough performance analysis comparing three machine learning algorithms using different metrics such as accuracy, Precision, Recall, F1-score, and FAR. The results consistently showed that the RF prediction models outperformed the SGD, K-NN, and SVM models, as shown in Figs. (3-6).

Figure (3) illustrates the accuracy of the developed stroke risk prediction model, evaluated using four classifiers: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Stochastic Gradient Descent (SGD). RF achieved the highest accuracy compared to SVM, K-NN, and SGD. However, SVM exhibited relatively better accuracy than SGD and the K-NN algorithm.

The RF algorithm emerged as the top-performing among the others based on the extensive experimental results.

Figure (4) contrasts the precision values of four classifiers Stochastic Gradient Descent (SGD), Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) for predicting stroke and non-stroke cases.

Precision is measured for both categories, represented by the blue bars for non-stroke and the orange bars for stroke.

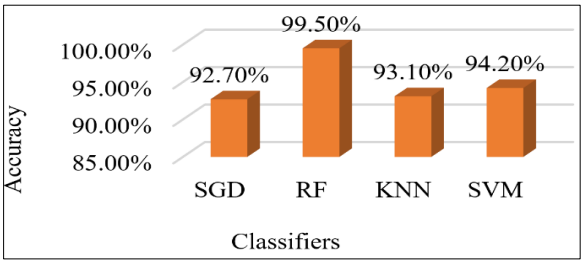

Fig. 3: Performance accuracy of the four classifiers

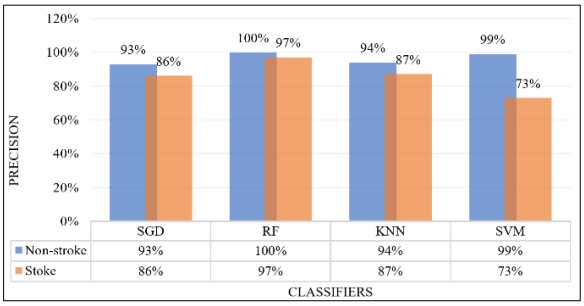

Fig. 4: Precision of the four classifiers across both classes

The Random Forest classifier demonstrates the highest performance, achieving a precision of 100% for non-stroke and 97% for stroke predictions. Similarly, the Support Vector Machine performs well in identifying non-stroke cases, with a precision of 99%, but shows a notable drop to 73% for stroke cases. The KNN classifier performs with an accuracy of 94% for non-stroke and 87% for stroke predictions. Finally, the SGD classifier achieves a precision of 93% for non-stroke and 86% for stroke predictions.

Overall, RF outperforms the other models in terms of precision, particularly in identifying stroke cases, while SVM shows a significant disparity between stroke and non-stroke precision. These results highlight the strengths and limitations of each classifier in stroke prediction.

Figure (4) illustrates the precision performance of four different machine-learning classifiers Stochastic Gradient Descent (SGD), Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) for forecasting stroke and non-stroke cases. Precision is shown separately for both categories, with the blue bars representing non-stroke precision and the orange bars representing stroke precision:

- SGD achieves a precision of 93% for non-stroke predictions and 86% for stroke predictions, indicating a relatively balanced performance across both categories
- RF stands out with perfect precision of 100% for non-stroke predictions and 97% for stroke predictions, demonstrating superior performance in both cases.
- KNN shows a precision of 94% for non-stroke cases and 87% for stroke cases, reflecting a solid performance but slightly lower precision for stroke predictions than non-stroke
- SVM performs well in non-stroke prediction with a precision of 99%, but its precision drops significantly to 73% for stroke predictions, indicating a challenge in accurately predicting stroke cases

In summary, the Random Forest classifier exhibits the best overall precision, particularly for stroke prediction, while SVM demonstrates a significant gap in performance between the two categories. The results suggest that while specific models, like RF, are robust across both categories, others may require further tuning to improve stroke prediction accuracy.

Figure (5) compares the recall values for four classifiers: Stochastic Gradient Descent (SGD), Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) in predicting stroke and non-stroke cases. Recall, which measures the ability of the model to identify all positive instances correctly, is represented by blue bars for non-stroke and orange bars for stroke predictions:

- SGD achieves a recall of 93% for non-stroke predictions and 89% for stroke predictions, indicating a balanced ability to correctly identify both categories, with a slightly better performance for non-stroke cases
- RF performs exceptionally well, with a recall of 100% for non-stroke predictions and 98% for stroke predictions, making it the best-performing classifier for recall across both categories.
- KNN shows a recall of 95% for non-stroke predictions and 88% for stroke predictions, maintaining high recall but slightly lower performance in identifying stroke cases than non-stroke
- SVM exhibits a high recall of 99% for non-stroke cases but significantly drops to 73%, indicating difficulty in correctly identifying stroke instances

In summary, the Random Forest classifier again outperforms the other models regarding recall, particularly in identifying stroke cases, while SVM shows a pronounced gap between non-stroke and stroke recall. This suggests that while specific models, like RF, are highly effective in correctly classifying both cases, others may need further tuning to improve their ability to detect stroke cases. Figure (6) depicts the F1-score performance of four classifiers: Random Forest (RF), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD) and K-Nearest Neighbors (K-NN). RF consistently attains the highest F1 score across all classes compared to the other models. Nevertheless, SVM demonstrates superior F1 scores compared to the K-NN and SGD algorithms. Based on these performance results, the Random Forest (RF) decision tree emerges as the preferred choice for predicting stroke risk.

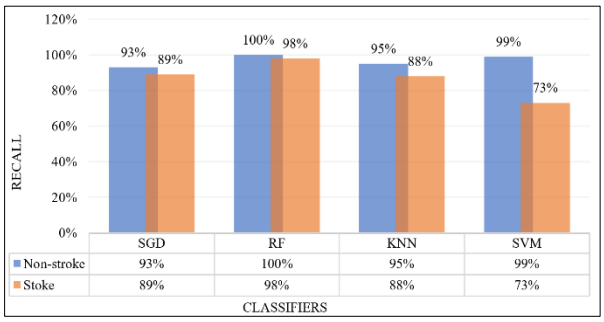

Fig. 5: Recall the three classifiers for both classes

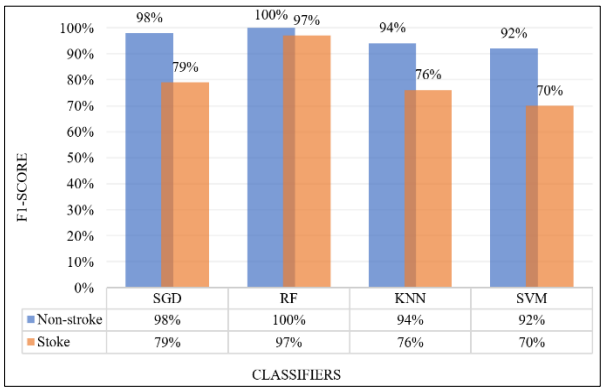

Fig. 6: F1 scores for the three classifiers in both classes

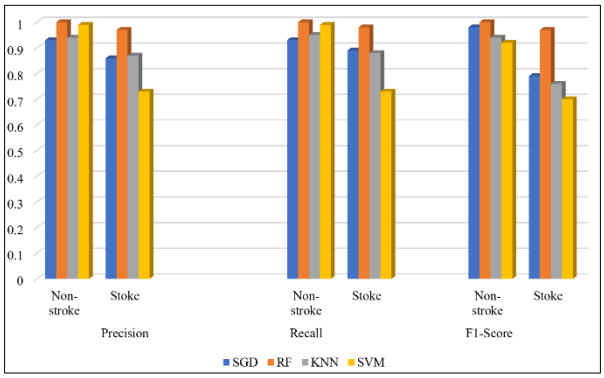

Fig. 7: A comparison of the performance metrics for all four models

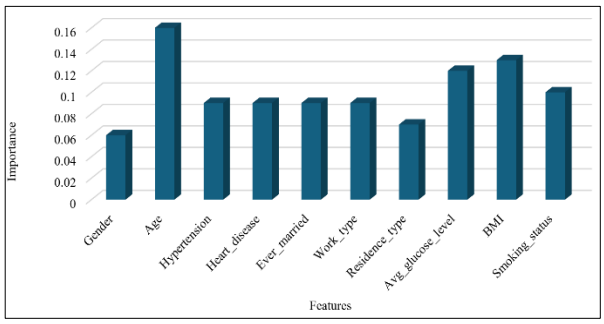

Fig. 8: Feature importance plot

Figure (7) compares the performance metrics for all four models: Precision, recall, and F1-score.

We analyzed feature importance using Random Forest feature importance values to gain insights into which factors contribute most to stroke prediction:

- Random forest feature importance: The model was trained, and feature importance scores were calculated. The results (Fig. 8) indicate that age, BMI, and average glucose level were the most significant factors in predicting stroke risk, while attributes such as residence type and gender had a relatively lower impact

Using machine-learning models for stroke prediction in healthcare presents several ethical challenges, including bias, privacy, transparency, accountability, and impact on patient care. To ensure fairness, data used for model training should represent diverse populations and be balanced to avoid bias. Robust privacy and data protection measures, including anonymization, encryption, and regulatory compliance, are crucial to safeguarding sensitive patient information. Transparency and explainability are essential for trust and informed decision-making, requiring interpretable AI techniques. While models can assist clinicians, responsibility for patient care remains with healthcare professionals, necessitating clear guidelines for accountability. Finally, deploying ML models must consider equitable access to technology and prevent over-reliance, ensuring patient care is not compromised. Strategies like bias auditing, transparency tools, data protection measures, and human oversight can address ethical considerations, promoting fair, responsible, and impactful ML use for stroke prediction.

Conclusion

In this study, we developed a machine-learning model to predict the occurrence of stroke using a dataset obtained from Kaggle, which included 5110 individuals and 12 attributes. After applying various data preprocessing techniques, we experimented with multiple machine learning models, including Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Random Forest (RND), and K-Nearest Neighbor (KNN). The SGD, SVM, and KNN models achieved a promising accuracy of 95%, demonstrating the potential of machine learning in assisting early stroke prediction.

Although these results are promising, future research could explore additional strategies to enhance the model's predictive accuracy and generalizability. First, expanding the dataset by incorporating additional features such as lifestyle habits, family medical history, and exercise routines may enhance the model's ability to capture keystroke predictors. Additionally, obtaining a larger and more balanced dataset with equal representation of stroke and non-stroke cases would mitigate the reliance on

oversampling techniques, thereby improving the robustness of the model.

Another potential research direction involves feature engineering, merging attributes like average glucose level, heart disease history, and age. This could reveal stronger correlations with stroke risk and improve model performance. Moreover, exploring alternative machine learning models, such as Logistic Regression, Artificial Neural Networks (ANN), or deep learning models, may offer additional insights and improvements. Continued Hyperparameter tuning and model optimization could also lead to better predictive performance.

In conclusion, the current model provides a solid foundation, but these future research directions could significantly enhance its accuracy, generalizability, and real-world applicability in reducing stroke incidence.

Acknowledgment

The authors would like to express their sincere gratitude to the Editor and the anonymous reviewers for their valuable time, constructive feedback, and insightful comments. Their rigorous evaluation and thoughtful guidance greatly enhanced this study's clarity, quality, and academic rigor.

Funding Information

The author did not receive support from any organization for the submitted work.

Author's Contributions

Mohammad Aref Alshraideh: A programmer who contributed to writing certain sections of the document.

Najwan Alshraideh, Abedalrahman Alshraideh, Yara Alkayed, Heba Alshraideh, Yasmin Al Trabsheh, and Bahaaldeen Alshraideh: Responsible for providing the dataset, writing and supervising the entire process.

Ethics

This study was conducted according to all relevant ethical guidelines and standards governing research. The data employed in this study were obtained from website sources. No human subjects were directly involved in the research, and no personally identifiable information was collected. The authors affirm that all analyses, interpretations, and conclusions presented herein reflect impartial scientific inquiry and adhere strictly to principles of integrity, transparency, and academic honesty.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Disclosure

This research stems from a master's student thesis. All the collaborating doctors actively participated in and contributed to this study.

Reference

Ahmad, E. F., Alshraideh, M., & Fram, K. A. (2020). Clinical Decision Support System for Diagnosing Gynecological Diseases. *Journal of Theoretical and Applied Information Technology*, *98*(16).

Alshraideh, M. (2020). Kidney Disease Predictor Based on Medical Decision Support System. *Trends in Technical & Scientific Research*, *04*(5), 555646. https://doi.org/10.19080/ttsr.2020.04.555646

Alshraideh, M., Alshraideh, N., Alshraideh, A., Alkayed, Y., Al Trabsheh, Y., & Alshraideh, B. (2024). Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital. *Applied Computational Intelligence and Soft Computing*, *2024*(1), 1–16. https://doi.org/10.1155/2024/5080332

Dritsas, E., & Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. *Sensors*, *22*(13), 4670. https://doi.org/10.3390/s22134670

Farhan, S., Alshraideh, M., & Mahafza, T. (2015). A Medical Decision Support System for ENT Disease Diagnosis using Artificial Neural Networks. *International Journal of Artificial Intelligence and Mechatronics*, *4*(2), 45–54.

Gaines, K., Commiskey, P., & Bridges, A. (2015). Abstract T P380: Blood Pressure Management among Stroke Patients: The Impact of a Comprehensive Post-Stroke Model in Louisiana. *Stroke*, *46*(suppl_1), tp380. https://doi.org/10.1161/str.46.suppl_1.tp380

Goyal, M. (2017). Prediction of Stroke Using Deep Learning Model. *Neural Information Processing*, 774–781. https://doi.org/10.1007/978-3-319-70139-4_78

Heart, N., Lung, & Institute, B. (2015). *Stroke*. https://www.nhlbi.nih.gov/health-topics/stroke

Johnson, W., Onuma, O., Owolabi, M., & Sachdev, S. (2016). Stroke: A Global Response is Needed. *Bulletin of the World Health Organization*, *94*(9), 634-634A. https://doi.org/10.2471/blt.16.181636

Kansadub, T., Thammaboosadee, S., Kiattisin, S., & Jalayondeja, C. (2015). Stroke Risk Prediction Model Based on Demographic Data. *2015 8th Biomedical Engineering International Conference (BMEiCON)*, 1–3. https://doi.org/10.1109/bmeicon.2015.7399556

Kaggle. (2024). Datasets. https://www.kaggle.com/datasets

Karapinar Senturk, Z. (2020). Early Diagnosis of Parkinson's Disease Using Machine Learning Algorithms. *Medical Hypotheses*, *138*, 109603. https://doi.org/10.1016/j.mehy.2020.109603

Karimi, Z. (2021). *Confusion Matrix*. Mahatma Gandi Central University, https://www.researchgate.net/publication/355096788_Confusion_Matrix#fullTextFileContent [on line 3/5/2024]

Katan, M., & Luft, A. (2018). Global Burden of Stroke. *Seminars in Neurology*, *38*(2), 208–211. https://doi.org/10.1055/s-0038-1649503

Mahesh, K. A., Shashank, H. N., Srikanth, S., & Thejas, A. M. (2020). Prediction of stroke using machine learning. *In Conference Paper· June*.

Mozaffarian, D. (2016). Heart disease and stroke statistics—2016 update: a report from the American Heart Association. *Circulation*, *133*(4), 338–360.

Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting Stroke from Electronic Health Records. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5704–5707. https://doi.org/10.1109/embc.2019.8857234

Party, S. W. (2012). *National Clinical Guideline for Stroke*. Royal College of Physicians.

Salah, B., Alshraideh, M., Beidas, R., & Hayajneh, F. (2011). Skin Cancer Recognition by Using a Neuro-Fuzzy System. *Cancer Informatics*, *10*, CIN.S5950. https://doi.org/10.4137/cin.s5950

Shafer, K. J., Siders, W. A., Johnson, L. K., & Lukaski, H. C. (2009). Validity of Segmental Multiple-Frequency Bioelectrical Impedance Analysis to Estimate Body Composition of Adults Across a Range of Body Mass Indexes. *Nutrition*, *25*(1), 25–32. https://doi.org/10.1016/j.nut.2008.07.004

Shboul, L., Fram, K., Sharaeh, S., Alshraideh, M., Shaar, N., & Alshraideh, N. (2022). Male and Female Hormone Reading to Predict Pregnancy Percentage Using a Deep Learning Technique: A Real Case Study. *AI*, *3*(4), 871–889. https://doi.org/10.3390/ai3040053

Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine Learning for Brain Stroke: A Review. *Journal of Stroke and Cerebrovascular Diseases*, *29*(10), 105162. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162

Sultan, M., Debebe, F., Azazh, A., & Hassen, G. W. (2017). Epidemiology of stroke patients in Tikur Anbessa specialized hospital: emphasizing clinical characteristics of hemorrhagic stroke patients. *Ethiopian Journal of Health Development*, *31*(1), 13-17. https://www.ajol.info/index.php/ejhd/article/view/167760

Ul-Haq, A. (2018). A Hybrid Intelligent System Framework for Predicting Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, *2018*(1), 3860146. https://doi.org/10.1109/access.2020.3016062

Zheng, L., Wang, Y., Hao, S., Sylvester, K. G., Ling, X. B., Shin, A. Y., Jin, B., Zhu, C., Jin, H., Dai, D., Xu, H., Stearns, F., Widen, E., Culver, D. S., Alfreds, S. T., & Rogow, T. (2015). Risk prediction of stroke: A prospective statewide study on patients in Maine. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 853–855. https://doi.org/10.1109/bibm.2015.7359796