I.J. Intelligent Systems and Applications, **2024, 1, 24-40**
Published Online on February 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2024.01.03

AI-powered Predictive Model for Stroke and Diabetes Diagnostic

Ngoc-Bich Le*
School of Biomedical Engineering, International University, Vietnam National University HCM City, HCM City, Vietnam
E-mail: lnbich@hcmiu.edu.vn
ORCID iD: https://orcid.org/0000-0001-7431-0157
*Corresponding author

Thi-Thu-Hien Pham
School of Biomedical Engineering, International University, Vietnam National University HCM City, HCM City, Vietnam
E-mail: ptthien@hcmiu.edu.vn
ORCID iD: https://orcid.org/0000-0001-5808-3214

Sy-Hoang Nguyen
School of Biomedical Engineering, International University, Vietnam National University HCM City, HCM City, Vietnam
E-mail: bebeiu18039@student.hcmiu.edu.vn
ORCID iD: https://orcid.org/0009-0008-8608-3236

Nhat-Minh Nguyen
School of Biomedical Engineering, International University, Vietnam National University HCM City, HCM City, Vietnam
E-mail: SBEIU19015@mp.hcmiu.edu.vn
ORCID iD: https://orcid.org/0009-0004-6120-1225

Tan-Nhu Nguyen
School of Biomedical Engineering, International University, Vietnam National University HCM City, HCM City, Vietnam
E-mail: ntnhu@hcmiu.edu.vn
ORCID iD: https://orcid.org/0000-0003-3343-0886

Received: 29 November 2023; Revised: 31 December 2023; Accepted: 05 January 2024; Published: 08 February 2024

Abstract: Research efforts in the prediction of stroke and diabetes prioritize early detection in order to enhance patient outcomes. To achieve this, a variety of methodologies are integrated. Existing studies, on the other hand, are marred by imbalanced datasets, lack of diversity in their datasets, potential bias, and inadequate model comparisons; these flaws underscore the necessity for more comprehensive and inclusive research methodologies. This paper provides a thorough assessment of machine learning algorithms in the context of early detection and diagnosis of stroke and diabetes. The research employed widely used algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost Classifier, to examine medical data and derive significant findings. The XGBoost Classifier demonstrated superior performance, with an outstanding accuracy, precision, recall, and F1-score of 87.5%. The comparative examination of the algorithms indicated that the Decision Tree, Random Forest, and XGBoost classifiers consistently exhibited strong performance across all measures. The models demonstrated impressive discrimination capabilities, with the XGBoost Classifier and Random Forest reaching accuracy rates of roughly 87.5% and 86.5% respectively. The Decision Tree Classifier exhibited notable performance, with an accuracy rate of 83%. The overall accuracy of the models was evident in the F1-score, a metric that incorporates recall and precision, where the XGBoost model exhibited a marginal improvement of 2% over the Random Forest and Decision Tree models, and 4.25 percent over the last two. The aforementioned results underscore the effectiveness of the XGBoost Classifier, which will be employed as a predictive model in this study, alongside the Random Forest and Decision Tree models, for the accurate identification of stroke and diabetes. Furthermore, combining datasets improves model performance by utilizing relative features. This integrated dataset improves the model's efficiency and creates a resilient and comprehensive prediction

model, improving healthcare outcomes. The findings of this research make a valuable contribution to the advancement of AI-driven diagnostic systems, hence enhancing the quality of healthcare decision-making.

Index Terms: Machine Learning, Predictive Model, Stroke Diagnosis, Diabetes Diagnosis, XGBoost Classifier (XGB), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression.

1. Introduction

1.1. Introduction to the Biomedical Problems

Chronic diseases, such as stroke and diabetes, exert substantial pressure on healthcare systems and provide adverse consequences for a vast number of individuals on a global scale. These diseases exert a substantial influence on both individuals suffering from them and the broader community. The rising incidence of stroke and diabetes has led to an escalating need for precise and prompt diagnostic techniques to efficiently identify and manage these ailments. In recent times, there has been a growing recognition of the potential of Artificial Intelligence (AI) and Machine Learning (ML) approaches in effectively tackling these difficulties. As per the definition proposed by Murphy et al. (2020) [1], stroke is delineated as a focal and abrupt impairment of neurological function arising from the impairment of blood arteries within the central nervous system, manifesting as either infarction or hemorrhage. On a global scale, stroke is recognized as the second most prominent cause of both mortality and impairment. Stroke is associated with a wide range of risk factors, processes, and causes. Hypertension is recognized as the primary modifiable risk factor for stroke, however, its influence may vary depending on the particular subtype. Ischemic strokes mostly arise from small vessel arteriosclerosis, cardioembolism, and major artery atherothrombosis, collectively constituting approximately 85% of occurrences. Young individuals may develop ischemic strokes due to extracranial dissection and other etiologies. The World Stroke Organization (WSO) reports that stroke is the second leading cause of death and the third leading cause of both death and disability worldwide [2]. Over the past thirty years, there has been a notable surge in the prevalence of stroke. This surge is characterized by a 70.0% increase in the frequency of new cases of stroke, a 43.0% increase in the number of stroke-related deaths, a 102.0% increase in the number of existing cases of stroke, and a 143.0% increase in some other relevant measure [2]. It is worth mentioning that a significant percentage of deaths caused by stroke, specifically 86.0%, as well as instances of disability, totaling 89.0%, are concentrated in countries classified as lower-income and lower-middle-income.

The genesis of chronic and diverse diabetes is characterized by multiple factors and a complex nature. Abnormalities in either insulin secretion, insulin action, or both can be ascribed to the occurrence of hyperglycemia. Numerous manifestations of hyperglycemia have been observed to disrupt the metabolic pathways associated with carbs, lipids, and proteins. Chronic hyperglycemia is accountable for a wide range of microvascular and macrovascular complications linked to diabetes, ultimately resulting in the majority of morbidity and death observed in individuals affected by this condition. Hyperglycemia is widely recognized as the principal diagnostic criterion for diabetes [3]. The occurrence of diabetes has become a noteworthy issue in both clinical and public health sectors. Based on the projections made by the International Diabetes Federation (IDF), the estimated number of individuals affected by diabetes was roughly 415 million in 2015. It is anticipated that this figure will increase to 642 million by the year 2040. Diabetes is correlated with substantial financial costs. In 2015, the International Diabetes Federation (IDF) projected that a notable amount of healthcare expenditures, ranging from 5% to 20%, be dedicated to the mitigation of diabetes-related issues. Based on a study conducted by researchers [4], it is anticipated that the financial outlays related to the management of diabetes and its associated complications on a worldwide level will experience a substantial rise to $802 billion by the year 2040. This projection signifies a significant growth compared to the documented expenditure of $673 billion in 2015.

Hewitt et al. (2012) [5] conducted a study examining the relationship between ischemic stroke and diabetes mellitus. The study specifically emphasized the importance of controlling cardiovascular risk factors in patients with diabetes to reduce the occurrence of stroke and improve overall prognosis. A link has been observed between hypertension, dyslipidemia, blood vessel abnormalities, and the occurrence of stroke and diabetes. Moreover, it is important to highlight that individuals with acute stroke often demonstrate abnormal glucose regulation. Nevertheless, it is crucial to emphasize that the application of rigorous glucose control after a stroke has not demonstrated substantial efficacy, as indicated by prior research. Research has demonstrated that effectively managing hypertension and cholesterol levels can serve as a preventive measure against stroke and can result in enhanced results for individuals diagnosed with diabetes.

Healthcare practitioners can augment their capacity to address risk factors, adopt preventive measures, and encourage healthy lifestyles by focusing their attention on these two particular illnesses. The precise prediction of stroke and diabetes carries considerable significance in the field of modern healthcare. The prompt detection of a medical issue enables the application of proactive interventions, customized treatment strategies, and enhanced patient results. It is crucial to prioritize the attainment of accurate prognostic information due to the swift spread and subsequent disruption produced by these diseases. The potential for revolutionizing the healthcare industry lies in the integration of artificial intelligence and machine learning, which enables the identification of risk factors and early detection of diseases. This enables timely interventions, disease reduction, and preservation of lives. The implementation of accurate prediction and prompt

intervention has promise in reducing the prevalence of stroke and diabetes, hence facilitating progress in the field of global health.

1.2. Analysis of Current Solutions for Early Stroke Prediction

Kaur et al. (2022) [6] underscored the necessity of noninvasive and economically viable diagnostic techniques for stroke, given that patients commonly experience transient ischemic episodes (TIA) before the occurrence of a complete stroke. This study focuses on the development of stroke prediction mechanisms through the utilization of processed electroencephalogram (EEG) data. Time series forecasting and performance evaluation commonly employ several approaches such as Long Short-Term Memory (LSTM), Bidirectional LSTM (biLSTM), Gated Recurrent Unit (GRU), and Feedforward Neural Network (FFNN). The results of the study demonstrate that the utilization of brain wave detection can facilitate the early identification of strokes by medical professionals, hence leading to potential life-saving interventions. According to the study, the detection of stroke at an early stage necessitates the utilization of noninvasive methodologies. The proposed approach for predicting strokes utilizes processed electroencephalogram (EEG) data and algorithms based on time series analysis. The experimental findings demonstrate high levels of accuracy, with the GRU model exhibiting superior performance. The findings have the potential to assist medical professionals in the early detection of strokes and enhance patient outcomes.

Another study conducted by Dev et al. (2022) [7] highlights the importance of medical record analysis in the context of stroke therapy and diagnosis. The researchers investigated the impact of risk factors in electronic health records on the prediction of stroke. Statistical techniques and principal component analysis are utilized to identify the most significant predictors of stroke. The primary risk factors associated with stroke encompass advanced age, cardiovascular disease, mean glucose level, and hypertension. The perceptron neural network exhibits superior accuracy and minimal error rates as compared to alternative benchmarking techniques, over a range of input features. These four specific properties contribute to its exceptional performance. The findings are presented based on a balanced dataset that was created using subsampling to mitigate the problem of imbalanced stroke data. This study aims to investigate the issue of dataset discrepancy to enhance stroke prediction accuracy and emphasize the significance of these characteristics in determining patient outcomes.

In their study, Al-Mekhlafi et al. (2021) [8] investigated the efficacy of machine learning (ML), deep learning (DL), and a combined ML-DL approach in the detection of stroke and cerebral hemorrhage. The study included medical records and magnetic resonance imaging (MRI) data. The medical records dataset was subjected to dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE) and feature selection employing Recursive Feature Elimination (RFE) techniques to generate diabetes and obese features. The accuracy, precision, recall, F1 score, sensitivity, and specificity of the SVM, KNN, Decision Tree, Random Forest, and Multilayer Perceptron classification algorithms were evaluated. The Random Forest algorithm demonstrated superior performance, achieving an accuracy, precision, recall, and F1 score of 99%. The MRI dataset was analyzed using the AlexNet architecture and a hybrid approach combining AlexNet with Support Vector Machines (SVM). The hybrid model, which combines the AlexNet architecture with a Support Vector Machine (SVM), exhibited superior performance compared to the standalone AlexNet model. The hybrid model achieved an accuracy of 99.9%, sensitivity of 100%, specificity of 99.8%, and an Area Under the Curve (AUC) value of 99.86%. Machine learning (ML), deep learning (DL), and hybrid methodologies have demonstrated efficacy in the detection and prediction of stroke and cerebral hemorrhage, thereby emphasizing their significance in the realm of biological research and therapeutic interventions.

The work conducted by Emon et al. (2020) [9] aimed to employ machine learning techniques for the early prediction of strokes. Blood pressure, body mass index (BMI), cardiovascular disease, average glucose levels, smoking status, history of stroke, and age have been identified as prognostic factors. Ten classifiers, including Logistics Regression, Stochastic Gradient Descent, and Decision Tree Classifier, are trained utilizing the aforementioned features. The primary objective of this project is to improve the precision of stroke prediction to facilitate the identification and prevention of strokes by medical professionals and patients. The algorithm utilized stroke-related information to train ten classifiers. The following machine learning models were trained individually: Logistic Regression, Stochastic Gradient Descent, Decision Tree, AdaBoost, Gaussian, Quadratic Discriminant Analysis, Multi-layer Perceptron, K-Neighbors, Gradient Boosting, and XGBoost. The technique of weighted voting was employed to combine the results of these crucial classifiers to maximize precision. The study on stroke prediction demonstrated a 97% accuracy rate with the utilization of a weighted voting classifier. The classifier demonstrated superior stroke prediction capabilities compared to the basal classifiers. The weighted voting classifier outperformed other classifiers in terms of false positive and false negative rates, as reported by previous studies. The results of this study demonstrate that the utilization of a weighted voting classifier yields reliable predictions in the context of stroke detection. Consequently, it is recommended that medical professionals and patients consider incorporating this approach into their practices as a means of early identification and prevention.

In a separate investigation, Sirsat et al. (2020) [10] discovered that machine learning (ML) can effectively and efficiently forecast the outcomes of stroke patients. The utilization of machine learning (ML) and deep learning (DL) is becoming more prevalent in this field; nonetheless, further investigation is required in certain domains. To address the classification of machine learning (ML) methods for brain injury, this study categorized them into four distinct functional or equivalent groupings. The study encompassed a total of 39 research conducted between the years 2007 and 2019. The results obtained from a comprehensive analysis of 10 papers focused on stroke-related research indicate that the Support Vector Machine (SVM) model demonstrated superior performance compared to other models. The majority of scholarly

articles primarily concentrate on stroke diagnosis, whereas there is a noticeable scarcity of literature about stroke treatment. This discrepancy suggests the existence of a research void in the field. Additionally, it was shown that studies on stroke frequently utilized computed tomography (CT) imaging. Both Support Vector Machines (SVM) and Random Forests were found to be effective machine-learning approaches in each category. This study on stroke highlights the significance of employing machine learning (ML) approaches. This work elucidates the utilization of machine learning (ML) in the domain of stroke research. It identifies areas of research that require more investigation and proposes appropriate models and datasets by conducting a comprehensive analysis of the existing literature.

Dritsas et al. (2022) [11] dedicated their efforts to the development of a machine-learning framework aimed at predicting the long-term risk of stroke. The machine learning algorithms encompassed in this set are Naive Bayes, Random Forest, Logistic Regression, K-Nearest Neighbors, Stochastic Gradient Descent, Decision Tree, Multilayer Perceptron, Majority Voting, and Stacking. The implementation and assessment of each algorithm were conducted using the most effective approaches and measures available. According to the study, layering demonstrates superior performance in terms of AUC, precision, recall, F-measure, and precision when compared to other classification approaches. The use of layering results in an Area Under the Curve (AUC) value of 98.9%. Additionally, the F-measure, precision, and recall metrics achieve scores of 97.4%, while the accuracy metric reaches 98%. The layering method demonstrates a strong predictive capability in assessing the risk of stroke, as evidenced by the presented data. The performance and validity of classifiers in stroke prediction can be assessed using evaluation metrics such as AUC, F-measure, and accuracy measurements. The findings indicate that the utilization of the stacking method yields superior outcomes in the context of stroke risk prediction. The study additionally proposes the incorporation of deep learning into the existing machine learning architecture as a means of enhancing its performance. This suggests a compelling approach to enhance the predictive capabilities of the framework. Furthermore, the researchers aim to utilize brain CT scan images to evaluate the efficacy of deep learning algorithms in predicting occurrences of strokes.

1.3. Analysis of Current Solutions for Diabetes Prediction

In their study, Chowdary et al. (2023) [12] put forth a proposal to identify individuals at high risk for type 1 diabetes and implement preventive interventions aimed at delaying the clinical progression of the disease. Interventions implemented during different stages of type 1 diabetes are designed to mitigate risk, reduce occurrence, or reinstate the functionality of beta cells. The research demonstrates that type 1 diabetes is associated with subclinical damage to pancreatic beta cells, and recent advancements in risk identification have facilitated the implementation of preventive clinical trials. The incorporation of familial history, genetic, immunological, and metabolic markers into screening methodologies enhances the accuracy of risk assessment. The strategy outlined in this study incorporates primary, secondary, and tertiary stages of prevention. In cases of overt type 1 diabetes, the implementation of tertiary prevention strategies aims to either restore the function of beta cells or prevent the occurrence of complications. Primary prevention is the proactive measures used to avoid the occurrence of type 1 diabetes in individuals who do not exhibit any damage to their pancreatic beta cells.

The classifier-based diabetes prediction and diagnostic decision support system was proposed by Vijayan et al. (2015) [13] in their study. The proposed system employs the AdaBoost algorithm with a Decision Stump as the underlying classifier for classification. The primary classifiers utilized for accuracy verification encompass Support Vector Machine, Naive Bayes, and Decision Tree. The primary objective of this study is to conduct a comparative analysis of AdaBoost's accuracy with other classifiers. The AdaBoost technique, while utilizing a decision tree as its foundation classifier, exhibits a superior accuracy rate of 80.72 percent. This accuracy surpasses that of other classifiers such as Support Vector Machines, Naive Bayes, and Decision Tree. The utilization of the AdaBoost algorithm in the proposed decision support system shows promise in its ability to accurately forecast and identify cases of diabetes.

In their study, Sonar et al. (2019) [14] developed a precise method for predicting the risk of diabetes by employing various algorithms such as Decision Tree, Artificial Neural Network (ANN), Naive Bayes, and Support Vector Machine (SVM). The aforementioned methodologies yielded Decision Tree models exhibiting an 85 percent precision rate, Naive Bayes models with a precision rate of 77 percent, and Support Vector Machine models with a precision rate of 77.3 percent. The aforementioned classification methodologies exhibit a high degree of precision, hence indicating their potential to accurately forecast the risk of diabetes. This methodology has the potential to identify and provide timely intervention for individuals, hence enhancing the management of diabetes and mitigating the occurrence of associated problems.

Arianna et al. (2018) [15] developed prediction models for type 2 diabetes complications under the MOSAIC project, funded by the European Union, utilizing a data mining pipeline. The pipeline encompassed the profiling of clinical centers, the targeting of predictive models, as well as the development and validation of those models. The Random Forest (RF) algorithm was employed to address the issues of missing data and class imbalance. Additionally, Logistic Regression with step-by-step feature selection was utilized to make predictions regarding the development of retinopathy, neuropathy, and nephropathy at different time intervals. The models included variables such as gender, age, time since diagnosis, BMI, Hba1c, hypertension, and smoking propensity. The ultimate specialized models employed distinct variables for each temporal and complexity scenario, resulting in an accuracy rate of 0.838%. These models can accurately forecast Type 2 Diabetes Mellitus (T2DM) complications and can be readily implemented within clinical settings.

1.4. Proposed Solutions

The research mentioned above emphasizes the critical importance of early detection in order to prevent the severe consequences of stroke and diabetes, allowing for life-saving interventions and improving patient outcomes as a whole. Scholars are presently investigating a wide range of approaches, including the examination of medical records, the analysis of EEG data, and the integration of machine learning (ML) and deep learning (DL) methods, in order to address the complexities associated with the prediction of stroke and diabetes. Despite these developments, a thorough examination of the literature uncovers specific deficiencies. Numerous studies showcase datasets that are deficient in diversity, which may result in the development of biased models. In specific research endeavors, imbalanced datasets, characterized by the dominance of one class over the other, present obstacles. Furthermore, the lack of exhaustive comparisons with alternative models or established benchmarks impedes the ability to assess the efficacy of the model in a thorough manner. Significantly lacking is research that addresses the simultaneous detection of diabetes and stroke; this indicates that the subject merits further investigation.

This study aims to develop prediction models using well-established machine learning (ML) techniques for the early identification of stroke and diabetes. The selected methodologies consist of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and XGBoost Classifier. These techniques have been chosen based on their established efficacy in predicting the aforementioned medical disorders, as substantiated by previous scholarly investigations. By directly confronting these obstacles, the research endeavor seeks to provide significant contributions in the advancement of machine learning models that not only effectively handle the intricacies of specific ailments, but also establish a comprehensive framework for the concurrent detection of stroke and diabetes. The innovation of this approach rests not only in the successful resolution of challenges presented by individual datasets, but also in their integration to form a more resilient and comprehensive prediction model. Consequently, this method provides a practical option for enhancing healthcare outcomes.

The successful deployment of machine learning algorithms for stroke and diabetes prediction hinges upon the acquisition of a carefully curated dataset. The dataset should include relevant information and accurately labeled examples of both stroke and diabetes cases. The effectiveness of the models' training and evaluation procedures depends on the use of appropriate measures, such as accuracy, precision, recall, and F1 score, to ensure a thorough evaluation of their predictive abilities.

To ensure the dissemination of the advantages derived from this predictive undertaking to the ultimate recipient, the last phase of the project will entail the creation of a website that is designed with a focus on user accessibility and ease of use. This platform will not only enhance the prediction processes but also function as a vital tool for healthcare professionals and individuals alike, thereby contributing to proactive healthcare management.

Our research is detailed in the manuscript's next sections. The "Materials and Methodology" section (Section 2) covers the implementation pipeline (2.1), data preparation (2.2) explaining how we handled and curated the datasets, machine learning models (2.3) used for prediction, evaluation metrics (2.4) assessing model performance, and production (2.5) – deploying our models in real-world settings. In the "Results and Discussion" section (Section 3), the model training process (3.1) is examined to explain the training regimen. Section 3.2 discusses performance measures including our models' accuracy, precision, recall, and F1-score. Section 4 concludes the manuscript by summarizing major findings, analyzing their ramifications, and suggesting future research.

2. Materials and Methodology

2.1. Implementation Pipeline

It is advisable to provide a comprehensive overview of the experimental workflow, as seen in Fig.1, before diving into the details. The details of each of the pipeline's five main stages are essential to building a strong basis for the others.

- Beginning with the initial phase, we carefully selected and examined appropriate datasets obtained from pre-existing collections. Using the WEKA toolkit, our objective was to analyze the datasets and identify the most optimal and likely relationships.
- In the second step of the process, the data preprocessing phase is responsible for addressing the issue of missing values and rectifying any imbalances in the data. Concurrently, a precisely produced stratified-split dataset is created. The utilization of this dataset serves the twin objective of mitigating overfitting and preventing data leaking, hence safeguarding the integrity of subsequent model training.
- The third phase is characterized by the hyperparameter optimization procedure, which represents a crucial stage including a rigorous exploration of the most optimal hyperparameters. As a consequence, a collection of meticulously optimized hyperparameters is obtained for each of the five chosen machine-learning models designated for training objectives.
- The assigned hyperparameters are subsequently utilized throughout the training procedure, ensuring the creation of models that possess both robustness and a high level of sensitivity to the intricacies present within the data. To determine the effectiveness of the trained models, a thorough evaluation is conducted in the following step, utilizing assessment criteria including F1 score, precision, accuracy, and recall. The rigorous assessment

procedure guarantees that the models adhere to the intended performance criteria. After doing this evaluation, the model that demonstrates the best level of efficiency is chosen to be implemented in the upcoming phase.

- The last stage of this complex procedure entails the implementation of the selected model, completely linked with a web-based backend and frontend framework. The purpose of this integration is to optimize patient access, guaranteeing that the model's predictive capabilities are both efficient and easily available to the end user in a user-friendly manner.

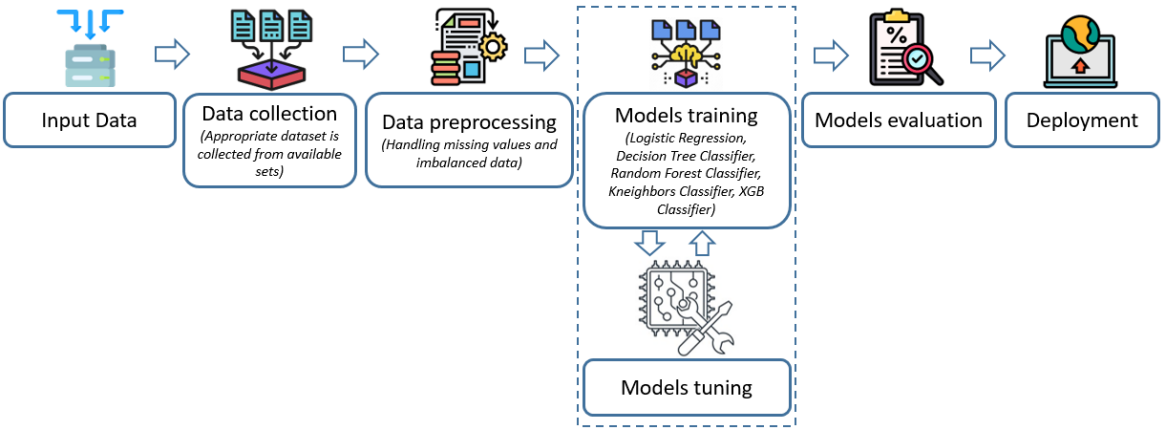

Fig.1. Experiment Pipeline (Step 1: Data collection; Step 2: Data preprocessing; Step 3: Model training; Step 4: Model evaluation; Step 5: Deployment)

2.2. Data Preparation

In this research study, we utilized two carefully selected datasets obtained from Kaggle, which provided a substantial amount of information for our predictive models. Each dataset contributed unique aspects to our analysis. The initial dataset, referred to as the Stroke Prediction dataset [16], demonstrated its significance as a useful source of insights. The dataset consists of 10 feature variables and 1 target variable. It exhibits a balanced distribution of classes, which provides a strong basis for conducting detailed analysis and training durable models. Simultaneously, our investigation also encompassed the Diabetes Prediction dataset [17], which is a storehouse of data that holds similar importance and possesses distinct characteristics. With a total of 17 feature variables and 1 target variable, this dataset provides a broader range of dimensions for further investigation. Similar to its stroke-focused counterpart, the Diabetes Prediction dataset exhibits a notable balance in the distribution of classes, hence augmenting the dependability and applicability of our predictive models. To present a more comprehensive analysis, Table 1 presents a comprehensive overview of the technical details of both datasets. This table provides insights into the distinct features, their respective attributes, and the general structure of each dataset. The thoughtful selection and thorough comprehension of our datasets highlight the rigorous methodology employed in constructing a strong basis for the creation of our prediction models.

To thoroughly examine the complexities of the datasets, we judiciously utilized the multifaceted function info() [18]. The utilization of this analytical tool enabled a thorough analysis of the data contained within both the Stroke Prediction dataset and the Diabetes Prediction dataset. The findings of this investigation are comprehensively documented in Fig.2 and Fig.3, which show a graphical depiction of the dataset attributes and offer significant observations regarding the variables, their classifications, and the presence of any potential data gaps or irregularities. The utilization of the info() function in a systematic manner is an essential component of our data exploration process. It facilitates a comprehensive comprehension of the structures within the datasets, hence establishing a solid foundation for making informed decisions throughout the following stages of developing our predictive model.

```
RangeIndex: 40910 entries, 0 to 40909
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   sex                40907 non-null  float64
 1   age                40910 non-null  float64
 2   hypertension       40910 non-null  int64
 3   heart_disease      40910 non-null  int64
 4   ever_married       40910 non-null  int64
 5   work_type          40910 non-null  int64
 6   Residence_type     40910 non-null  int64
 7   avg_glucose_level  40910 non-null  float64
 8   bmi                40910 non-null  float64
 9   smoking_status     40910 non-null  int64
 10  stroke             40910 non-null  int64
dtypes: float64(4), int64(7)
```

Fig.2. Stroke prediction dataset information

Table 1. Datasets' attribute description

Attribute name	Possible values	Description
Age	Years	Represents the age of participants.
Gender	1: male; 0: female	Indicates the gender of the participants.
Hypertension	The patient has ever had hypertension (1) or not (0)	Indicates whether participants have hypertension, with a prevalence of 21.39%.
Heart_disease	Patient has ever had heart_disease(1) or not (0)	Indicates whether participants suffer from heart disease, with a prevalence of 12.77%.
Ever married	Patient married (1) or not (0)	Represents the marital status of the participants, with 82.15% being married.
Work type	Patient job type: 0 - Never_worked, 1 - children, 2 - Govt_job, 3 - Self-employed, 4 - Private	Represents the participants' work status, with categories including children, private, self-employed, govt_job, and never_worked.
Residence type	Patient area: 1 - Urban, 0 - Rural	Represents the participants' living status, with categories including urban (51.50%) and rural (48.50%).
Avg glucose level (mg/dL)	Patient average blood sugar level (55.1-272)	Captures the average glucose level of the participants.
BMI (Kg/m2)	Body Mass Index (11.5-92)	Captures the body mass index of the participants.
Smoking Status	1 - smokes, 0 - never smoked	Represents the smoking status of the participants, with categories including smoke (48.87%), and never smoked (51.13%).
Smoking condition	The patient has ever smoked more than 100 cigarettes in their lives (1) or not (0)	Indicates whether the participants had ever smoked more than 100 cigarettes in their lives, with 47.52% of prevalence.
Stroke	The patient has ever experienced a stroke (1) or not (0)	Indicates if the participant has previously experienced a stroke, with a prevalence of 50.01%.
High cholesterol	0 = no high cholesterol 1 = high cholesterol	Indicates whether participants have a high cholesterol level, with a prevalence of 52.56%.
Cholesterol check	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years	Indicates whether participants had cholesterol checks in 5 years, with a prevalence of 97.51%.
Heart disease or attack	Coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes	Indicates whether participants had coronary heart disease (CHD) or myocardial infarction (MI), with a prevalence of 14.78%.
Physical activity	Physical activity in the past 30 days - not including job 0 = no 1 = yes	Indicates whether participants had physical activity in the past 30 days - not including job, with a prevalence of 70.29%.
Fruit consuming	Consume Fruit 1 or more times per day 0 = no 1 = yes	Indicates whether participants had consumed fruit one or more times per day, with a prevalence of 61.17%.
Vegetable consuming	Consume Vegetables 1 or more times per day 0 = no 1 = yes	Indicates whether participants had consumed vegetables one or more times per day, with a prevalence of 78.87%.
Heavy consumption of alcohol	0 = no 1 = yes	Indicates whether participants had consumed alcohol heavily, adult men >=14 drinks per week and adult women>=7 drinks per week, with a prevalence of 4.27%.
General Health	Scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor	the participants indicated their general health based on a scale of 1 - 5, 1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor.
Mental health	Days of poor mental health scale 1-30 days	Represents the day in the poor mental health of the participants, with the scale from 1 - 30 days.
Physical health	Physical illness or injury days in the past 30 days scale of 1-30	Represents the day in physical illness or injury of the participants in the past 30 days, with the scale from 1 - 30 days.
Difficulty walking	Does the patient have serious difficulty walking or climbing stairs? 0 = no 1 = yes	Indicates whether participants had serious difficulty walking or climbing stairs, with a prevalence of 25.27%.
Diabetes	0 = no diabetes, 1 = diabetes	Indicates if the participant has previously experienced diabetes, with a prevalence of 50.00%.

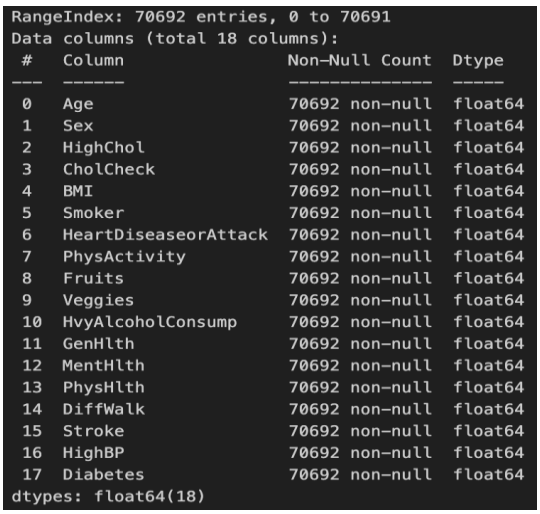

Fig.3. Diabetes prediction dataset information

After the dataset was extracted, a crucial step in the data analysis process was visualizing the interrelationships between variables. Correlation is a statistical measure that assesses the degree of the association between two variables. The aforementioned component is a crucial determinant in comprehending the relationship between variables and can be classified as either positive or negative. A positive correlation indicates a propensity for both variables to exhibit simultaneous increases. For example, the empirical evidence suggests a significant correlation between the number of study hours and academic test results, indicating that an augmentation in study hours is likely to yield enhanced performance in examinations. In contrast, a negative correlation indicates that as one variable increases, the other variable decreases. An inverse correlation exists between smoking behavior and life expectancy, wherein an escalation in smoking habits is associated with a decline in overall life expectancy.

Researchers utilize a correlation matrix as a tool to store and exhibit relationships between multiple variables, enabling a comprehensive evaluation and assessment of correlations. A correlation matrix is a graphical representation that illustrates the correlation coefficients between a set of variables [19]. The application of this approach allows for the identification of variable combinations that demonstrate a significant level of similarity, as well as the discovery of hidden associations that were not previously apparent [19]. Researchers and analysts seeking a comprehensive comprehension of the interrelationships between different variables can benefit from utilizing correlation matrices.

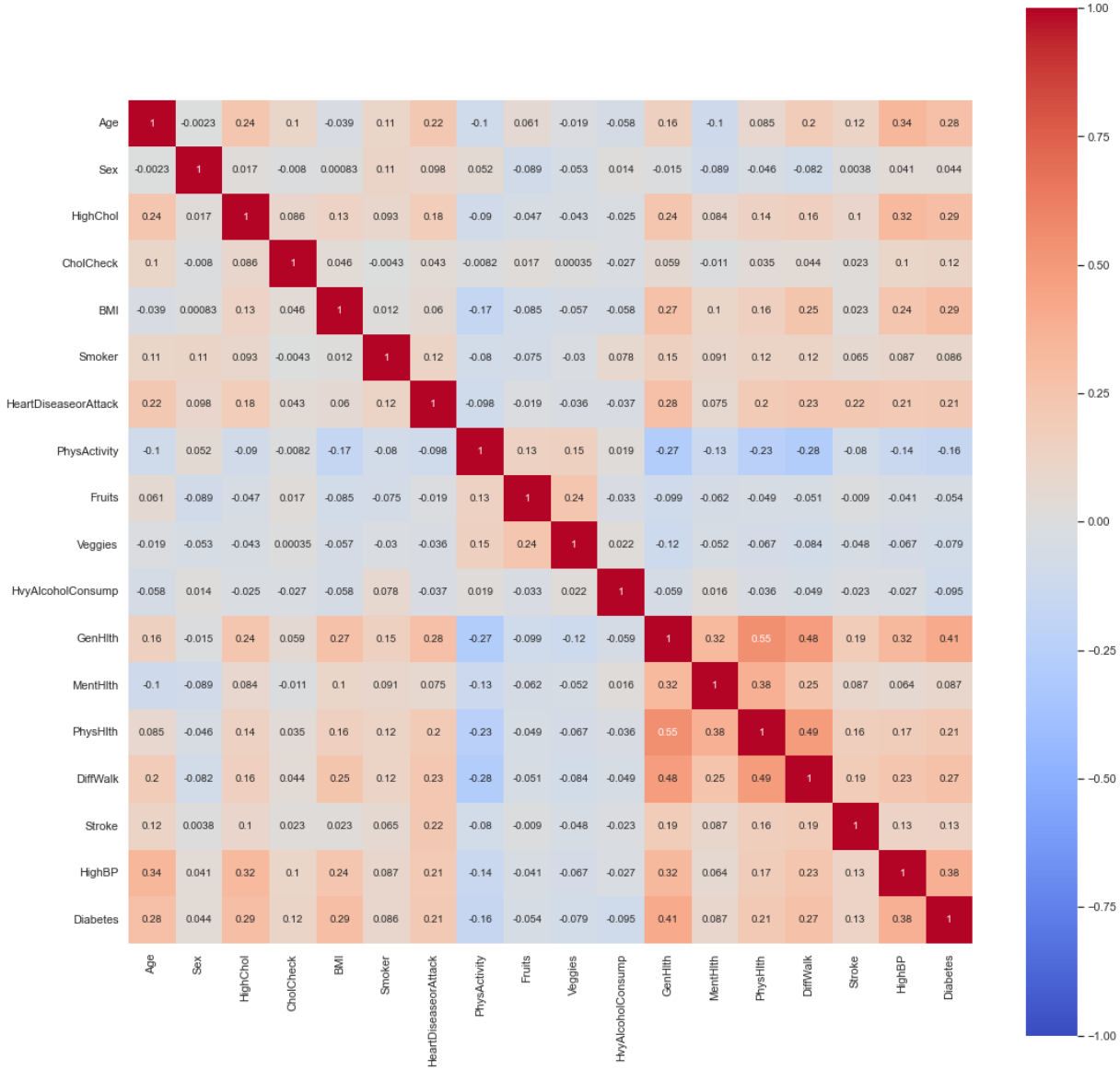

Fig.4. Correlation matrix heatmap of diabetes prediction dataset

Heatmaps provide a visual representation that facilitates the evaluation of the strength of relationships between numerical data. Graphs play a crucial role in the identification of relationships between variables and the quantification of their magnitude. Correlation graphs often consist of a matrix of numerical variables, where each variable is represented by a unique column. The rows inside the matrix serve as a representation of pairs of variables, while the values included within the cells indicate the size of their link. Positive values indicate a positive association, whereas negative values

indicate a negative correlation. Correlation heatmaps are a useful tool for quickly and effectively identifying relationships between variables. By applying color gradients, these heatmaps enable the study of potential connections and enhance analytical and decision-making processes. In addition, correlation heatmaps are useful tools for detecting abnormalities and understanding both linear and nonlinear relationships between variables, hence improving the overall understanding of the interconnections within a particular dataset. The process of examining and representing the datasets is carried out by utilizing the heatmap() function provided by the Seaborn library. The heatmap() function is a robust tool that aids in the visualization of a correlation matrix heatmap. The utilization of this visualization technique provides valuable insights into the interconnectedness of variables within datasets, as illustrated in Fig.4 and Fig.5.

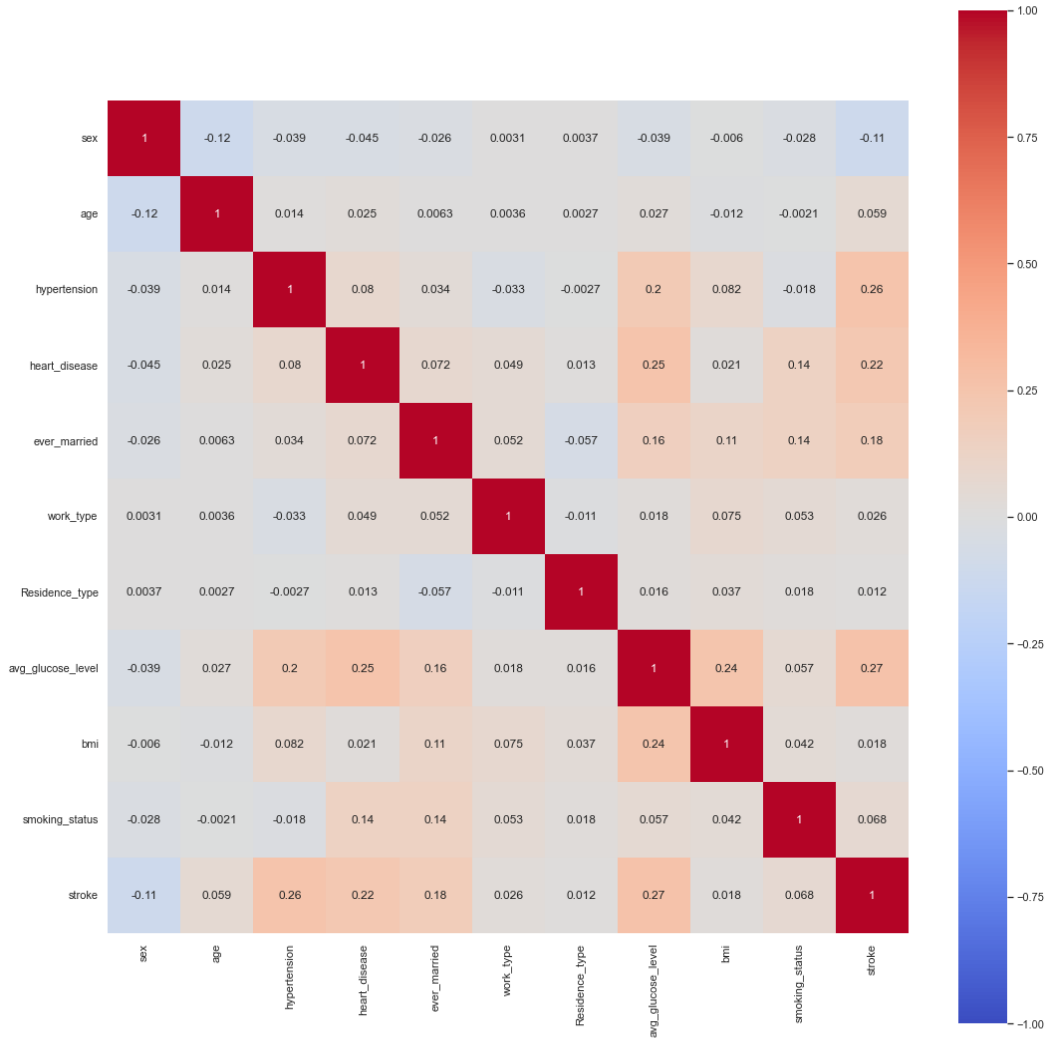

Fig.5. Correlation matrix heatmap of stroke prediction dataset

The process of partitioning a dataset into a training set and a test set can be easily achieved by employing the train_test_split() function provided by the Scikit-learn module in the Python computer language [20]. The function indicated above enables the automation of the splitting process and ensures accurate data division. The train_test_split() function requires the input features (X) and their corresponding target labels or outcomes (y) to be provided as parameters to partition the dataset. In addition, this approach offers additional alternatives that enable enhanced manipulation of the division procedure. Several elements that should be considered include the size of the test, the random state, and the stratification. The test size parameter represents the proportion of the dataset that is allocated to the test set. For instance, when the test_size parameter is assigned a value of 0.2, it partitions 20% of the available data for the test set, while assigning the remaining 80% to the training set. The inclusion of the random state parameter's optional argument enhances the reproducibility of outcomes. The data will exhibit consistent division patterns when the function is executed with a predetermined random state value.

Furthermore, when dealing with classification tasks that involve categorical target labels, the stratified parameter can be employed to ensure that each class is proportionally represented in both the training and test datasets. This method facilitates the maintenance of a proportional distribution of classes within each set. When the function is used with appropriate parameters, it produces segregated datasets referred to as X_train, X_test, y_train, and y_test. The training sets, referred to as X_train and y_train, include a subset of the input features and target labels that are employed in the

process of training the machine learning model. The test sets, referred to as X_test and y_test, consist of the remaining data used to evaluate the model's performance.

The train_test_split() method is employed to streamline the data splitting process, ensuring precise division of the data. This, in turn, facilitates the training and testing of the model on separate and uncorrelated datasets. This methodology facilitates the assessment of the model's ability to generalize to unseen data and accurately quantify its performance.

The utilization of cross-validation is a crucial method in this research, as it allows for the evaluation of the effectiveness and applicability of the machine learning models utilized. The dataset is divided into various subsets, commonly referred to as folds. Each fold is used for training the models, while the remaining data is used for validation. This process is repeated for each fold, allowing for iterative training and validation. This methodology assists in addressing concerns associated with overfitting and offers a more precise assessment of the models' capabilities. The utilization of 5-fold cross-validation achieves a harmonious equilibrium between computing efficiency and robust performance assessment, guaranteeing the models' ability to generalize effectively across diverse subsets of the data. The merits of this approach reside in its capacity to provide a thorough assessment of model performance, enhancing the reliability and applicability of the findings across various datasets.

2.3. Machine Learning Models

This section presents a comprehensive overview of the models that have been incorporated into the classification framework for assessing the prevalence of stroke and diabetes. Various classifiers were utilized for this objective, such as Logistic Regression [21], Decision Tree [22], Random Forest [23], K-Nearest Neighbors (KNN) [24], and XGBoost Classifier [25]. When faced with the decision between machine learning and deep learning approaches, it is crucial to consider various factors to determine the most appropriate technique. By conducting a thorough analysis of key factors including data accessibility, model complexity, computational resource requirements, feature engineering process, problem domain, model interpretability, data utilization efficiency, model execution and integration, training time, and resource limitations, a prudent decision can be made by taking into account the specific requirements of the task at hand. Based on the evaluation of the aforementioned parameters, it is recommended to prioritize the application of Machine Learning over Deep Learning in the particular context. Machine Learning provides numerous advantages in various contexts, such as situations with limited datasets, the requirement for simplified models, the necessity for interpretability, constraints on processing resources, and a wide range of problem domains. In addition, it offers a more efficient deployment procedure, decreased training time, and decreased reliance on substantial data. On the other hand, Deep Learning demonstrates a higher level of appropriateness for complex tasks that involve large amounts of annotated data, such as the identification of images and sounds. However, the increased complexity, longer training period, and higher demand for computational resources impose constraints on its applicability in resource-constrained environments and in scenarios where interpretability is of utmost importance.

The inclusion of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost Classifier in our research is based on a thorough evaluation of their relevance in tackling binary classification problems. Logistic Regression functions as a fundamental model, offering a clear comprehension of the associations between features and the target variable. The Decision Tree, Random Forest, and XGBoost Classifier are well-known algorithms recognized for their ability to effectively handle intricate linkages and interactions among features, hence providing improved prediction skills. The K-Nearest Neighbors algorithm, due to its straightforwardness and emphasis on local patterns, serves as a suitable supplement to ensemble approaches. The primary objective of this ensemble is to encompass a wide range of algorithms in order to provide a thorough examination of predictive strategies for stroke and diabetes prediction. The integration of different models enables our research to identify the advantages and disadvantages of each, finally enhancing a comprehensive and knowledgeable examination.

2.4. Evaluation Metrics

In the domain of classification, it is widely acknowledged that assessment metrics can be classified into three distinct categories: threshold metrics, probability metrics, and ranking metrics [26]. These metrics evaluate the efficacy of classifiers with varying purposes. While providing a comprehensive evaluation, individual scores have the potential to mask nuanced behavioral complexities as they facilitate the comparison and examination of data. Researchers often employ threshold and ranking criteria to evaluate the efficacy of classifiers. These indications have proven to be valuable in three distinct evaluation scenarios. The fundamental objective of utilizing evaluation metrics is to evaluate the ability of trained classifiers to generalize. The quantification and summarization of the performance of these classifiers on previously unseen data during the testing phase is how this is accomplished. The statistic commonly utilized to evaluate the capacity of a model to generalize is accuracy or error rate. This statistical measure evaluates the degree of reliability associated with predictions made on instances that have not been previously seen. In addition, the utilization of assessment metrics is of paramount importance in the model selection process as they assist in identifying the most appropriate classifier among a variety of choices, with the ultimate goal of attaining the best performance on future unseen data. Evaluation metrics are of utmost importance in the training of classifiers as they act as discriminators to select the most optimal answer from a set of generated solutions. The utilization of the measure of accuracy allows for the differentiation of solutions and the identification of the most advantageous one generated by a specific classification algorithm. This procedure ensures that only the most optimal model is evaluated using previously unobserved data.

This paper aims to thoroughly investigate optimal solutions in the context of classification training within the domain of binary classification problems. To assess the effectiveness of our models in this endeavor, we utilized the comprehensive evaluation provided by a confusion matrix [26], which is a robust tool for analyzing classification performance. The detailed analysis of this evaluation is presented in Table 2, which provides a broad range of model evaluation parameters. The evaluation metrics considered in this study encompass not only the traditional measure of Accuracy, but also delve into the nuances of Precision (P), Sensitivity, Specificity, Recall (R), and F1-score (F1). Every parameter has a unique function in elucidating the intricacies of our models' performance. The term "True Positive" (TP) refers to examples that have been accurately categorized within the correct class. On the other hand, "False Positive" (FP) refers to instances that have been incorrectly classified as belonging to the true class, when in fact they belong to a different class. In the context of classification, the term "False Negative" (FN) pertains to situations where examples belonging to the actual class are inaccurately classified as belonging to a different class. The intricacies discussed here are derived from the results shown in the confusion matrix, which are visually depicted in Fig.7 and Fig.8. The comprehensive evaluation approach presented in this study serves to enhance our comprehension of the performance of the models, while also establishing a solid foundation for identifying the strengths and areas for development in our binary classification efforts.

Table 2. Metrics for classification problem

Metrics	Formula	Description
Accuracy	$\frac{TP+TN}{\text{TP + TN + FP + FN}}$	The accuracy metric assesses, in general, the proportion of correct predictions relative to the total number of instances evaluated.
Precision (P)	$\frac{TP}{TP+FP}$	Precision is the ratio of correctly predicted positive patterns to the total predicted positive patterns in a cohort.
Sensitivity	$\frac{TP}{TP+FN}$	This metric is used to measure the proportion of correctly classified positive patterns.
Specificity	$\frac{TN}{TN+FP}$	This metric is used to measure the proportion of incorrectly classified negative patterns.
Recall (R)	$\frac{TP}{TP+FN}$	Recall is used to measure the proportion of correctly classified positive patterns.
F1-score (F1)	$2\frac{P\times R}{P\times R}$	This metric is the harmonic mean of recall and precision values.

2.5. Production

Upon entering the deployment phase, the complexities of the inference production process are depicted in Fig.6. This crucial stage in our technique involves the application of cutting-edge technologies, specifically Docker and Amazon Web Services (AWS), taking the lead. Docker plays a vital role in our deployment strategy by facilitating the packaging of code and its complex dependencies into self-contained and easily transportable containers. These containers provide exceptional adaptability, beyond the constraints of certain platforms, allowing them to be executed in a wide range of contexts, regardless of the underlying hardware or operating system. This not only enhances the efficiency of deployment but also strengthens the model's ability to adapt to different computational environments. The deployment of this orchestration takes place on the resilient infrastructure provided by Amazon Web Services (AWS). By utilizing the capabilities of serverless container execution on the AWS platform, our model assumes a prominent role inside an environment specifically engineered for effortless scalability and optimal resource use. The strategic deployment architecture employed not only guarantees the accessibility and availability of our application but also strategically situates it to leverage the whole capabilities of cloud-based services to achieve improved performance and enhance the user experience.

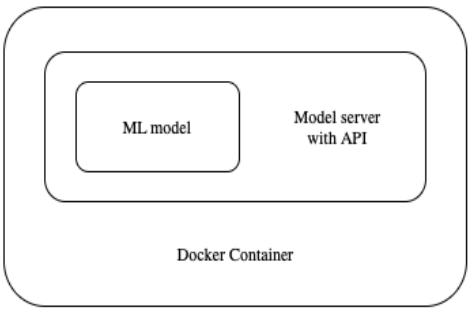

Fig.6. The architecture of production

3. Results and Discussion

3.1. Model Training

In the persistent quest for attaining optimal performance in the field of machine learning, the careful selection and deep understanding of algorithm parameters arise as crucial factors to be taken into account. Exploring the complexities

of parameter analysis for well-established models such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost Classifier transcends mere task-oriented efforts and instead entails a thorough and comprehensive investigation. This undertaking beyond surface-level analysis, providing a deep understanding of the intricate mechanisms behind these systems. This thorough examination has a twofold aim. To begin with, this enhances our comprehension of the fundamental mechanisms that govern the operation of each algorithm, revealing the complexities that define their predictive capabilities. Additionally, and potentially of greater significance, it provides us with helpful suggestions for optimizing model performance. Through the process of navigating the intricate network of algorithmic parameters, we can discover the essential factors that, when carefully manipulated, can enhance the prediction powers of these models. The importance of this procedure is heightened within the framework of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost Classifier, where the emphasis is placed on the precise identification of stroke and diabetes. The comprehensive examination conducted in this study serves as the foundation for the development of an artificial intelligence-based predictive model, specifically designed to excel in the complex field of medical diagnostics. The utilization of a meticulously designed strategy for parameter optimization enables us to establish a foundation for a more sophisticated, streamlined, and ultimately impactful predictive model. This model holds great potential for making substantial advancements in the diagnostic fields of stroke and diabetes.

3.2. Performance Metrics

In this particular section, a comprehensive analysis is presented, providing insight into the results obtained from our efforts. Fig.7 functions as a visual representation, depicting a graphical sequence of the accuracy score, precision, recall, and F1 score attained by each machine learning model. The visual depiction serves as a crucible for comparison analysis, allowing for the examination of the different degrees of performance demonstrated by various algorithms. Within the framework of the stroke prediction dataset, a significant revelation emerges. Both the XGBoost and Decision Tree models demonstrate exceptional performance, exhibiting the greatest scores in four crucial evaluation metrics: accuracy, precision, recall, and F1-score. The remarkable effectiveness of these algorithms in accurately predicting occurrences of stroke is emphasized by this significant constancy. In contrast, when the focus is moved to the diabetes prediction dataset, the XGBoost model emerges as the most prominent, surpassing its peers in terms of overall performance. The results of this study highlight the exceptional performance of the XGBoost model in accurately predicting diabetes, establishing it as a prominent contender in this field.

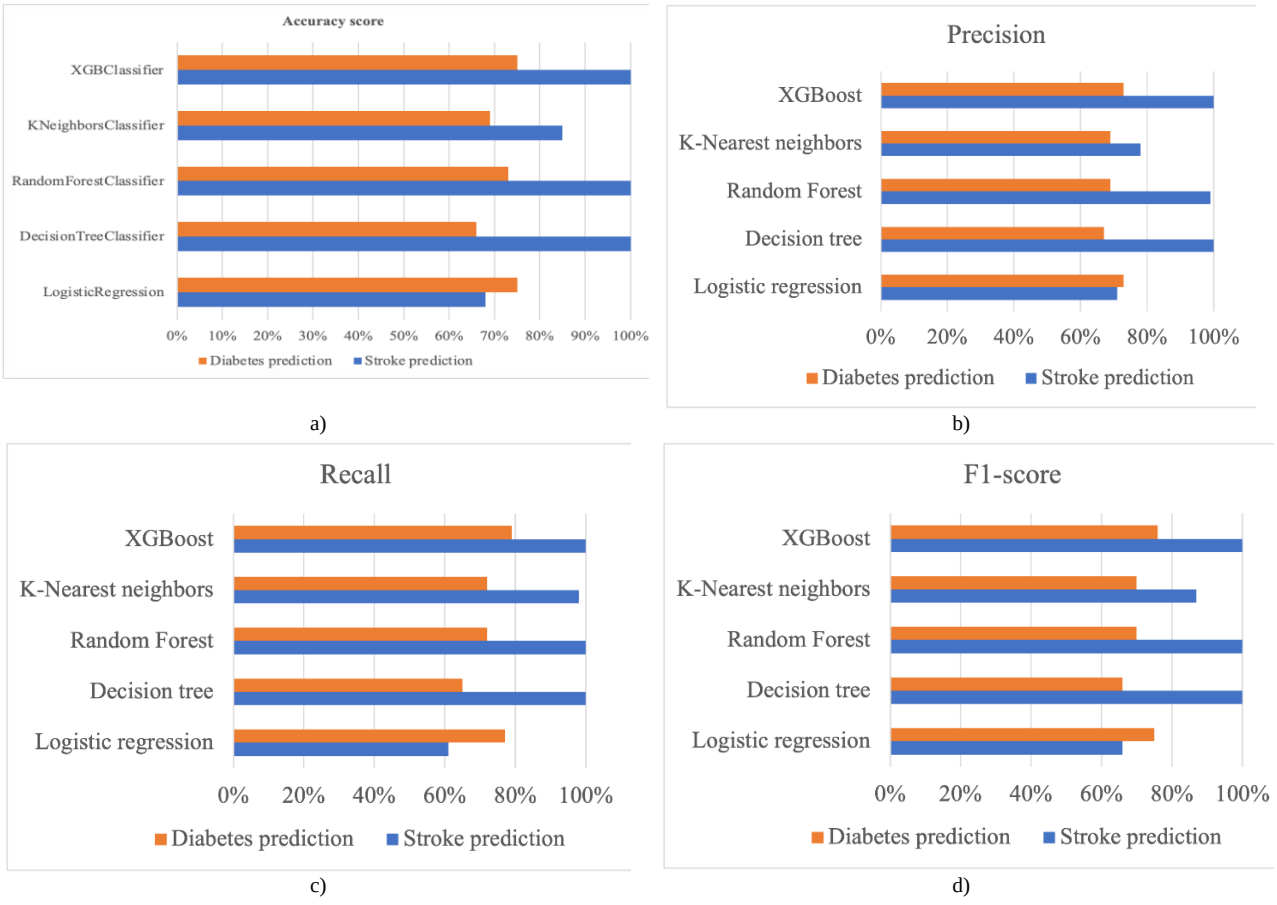

Fig.7. a) Machine learning algorithms performance on accuracy_score for both datasets, b) Machine learning algorithms performance on precision_score for both datasets, c) Machine learning algorithms performance on recall for both datasets and d) Machine learning algorithms performance on F1-score for both datasets

In addition to evaluating model performance measures, we do a comprehensive analysis of the datasets, exploring the complexities associated with feature relevance. The stroke prediction dataset emphasizes the 'bmi_index' feature, highlighting its crucial importance in the predictions made by the decision tree model, as illustrated in Fig.8a. The significance of body mass index (BMI) in influencing the accuracy of the model's ability to identify prospective stroke cases is highlighted by this finding. In contrast, the analysis of the diabetes prediction dataset reveals the significant influence of the 'high blood pressure' variable on the predictive performance of the XGBoost model, as illustrated in Fig.8b. This specific discovery enhances the significance of hypertension in predicting diabetes, emphasizing its crucial contribution to the accuracy of the model in identifying those who are at risk.

The aforementioned discoveries go beyond simple statistical results; they function as indicators that guide our focus toward the essential characteristics that hold substantial sway within the models. Furthermore, these observations serve as guiding principles for further improvement, providing a fertile environment for enhancing the predictive capabilities of our algorithms and promoting a continual progression toward increased diagnostic accuracy.

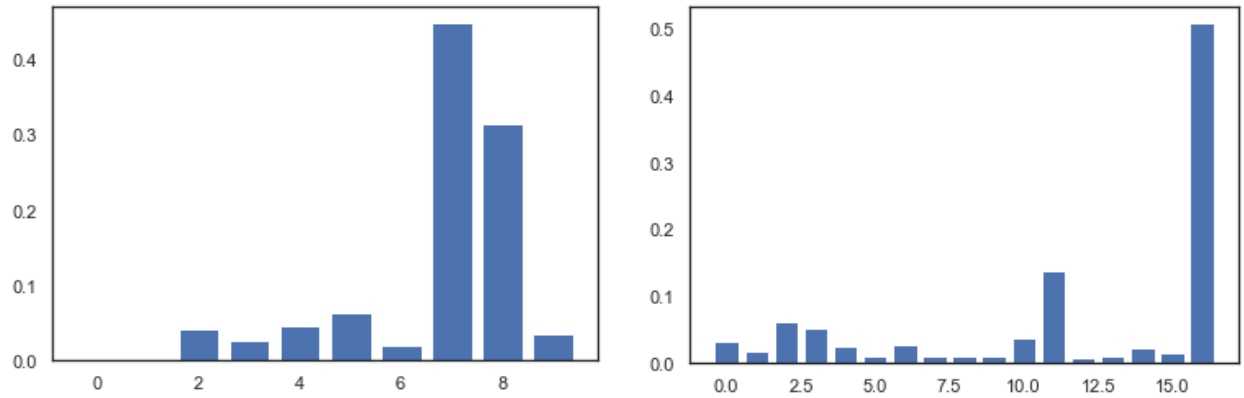

Fig.8. a) Feature importance plotting on Stroke prediction, and b) Feature importance plotting on Diabetes prediction

In addition to the informative column chart, we have included a thorough table that presents the average performance measures, namely accuracy, precision, recall, and F1-score, for each algorithm over both datasets, thereby enhancing our analytical capabilities. The presented tabular representation enhances our analysis, providing a detailed understanding of the performance of the models at a class-specific level, surpassing superficial observations. Within the domain of accuracy, these measures reveal the models' proficiency in reducing false positive errors, providing insight into their precision-focused skills. Concurrently, recall measures assess the effectiveness of decreasing false negative mistakes, highlighting the models' ability to accurately detect instances that should not be overlooked. The F1-score is a metric that combines precision and recall, providing a comprehensive evaluation of the overall performance of models.

Table 3. Machine learning algorithms average performance by accuracy, precision, recall, and f1-score

Studies	Models	Accuracy	Precision	Recall	F1-score
This study	LogisticRegression	0.715	0.715	0.7125	0.7125
	DecisionTreeClassifier	0.83	0,8325	0,8325	0,8325
	RandomForestClassifier	0.865	0,855	0,85	0,855
	KNeighborsClassifier	0.77	0,785	0,775	0,7725
	XGBClassifier	**0.875**	**0,875**	**0,875**	**0,875**
Ojha et. al. (2023) [27] (stroke prediction)	Naïve Bayes	0.756	0.719	0.756	0.734
	AdaBoost	0.801	0.756	0.801	0.755
	Decision Table	**0.821**	**0.796**	**0.821**	**0.784**
	Random Forest	0.801	0.751	0.801	0.744
	K_NN	0.756	0.719	0.756	0.734
Chauhan_et. al. (2023) [28] (Diabete prediction)	**Decision tree classifier**	**0.759**	NA	NA	NA
	Logistic regression	0.746	NA	NA	NA

Table 3 serves as a complete compilation, providing a thorough overview of the average performance demonstrated by our set of five machine learning algorithms. Once again, the XGBoost model demonstrates its superiority by exhibiting the highest average performance across the various measures. The consistent superior performance demonstrated by the XGBoost model not only serves as evidence of its effectiveness but also establishes it as a prominent example in the field of predictive analytics. Additionally, it is important to emphasize that the performance of the XGBoost model surpasses the benchmarks established by previous studies. The outstanding performance gained in this study is further highlighted

by comparing it with the results of Ojha et al. (2023) [27] in stroke prediction and Chauhan et al. (2023) [28] in diabetes diagnosis. The results of this study not only confirm the reliability of our model but also represent a significant leap in predictive modeling, with potential implications for improving the diagnosis of stroke and diabetes.

Upon delving into the complexities of our model results, it becomes evident that the XGBoost Classifier stands out as the most exceptional performer. It demonstrates a remarkable level of average accuracy, precision, recall, and F1-score, with each metric achieving a noteworthy value of 87.5%. A comparative examination of the mean performance across all methodologies reveals the consistent excellence of the Decision Tree, Random Forest, and XGBoost classifiers. These models demonstrate strong discriminatory ability, as seen by their impressive levels of accuracy. The XGBoost Classifier and Random Forest algorithms demonstrate high accuracy rates of around 87.5% and 86.5%, respectively, indicating their proficiency in effectively classifying situations. The Decision Tree Classifier has notable performance, with a commendable accuracy rate of 83%. When examining the F1-score, a metric that combines precision and recall, it is observed that the XGBoost model slightly outperforms both Random Forests and Decision Tree models. The observed improvements in the F1-score, specifically by 2% and 4.25% respectively, further establish the dominance of the XGBoost model in delivering a well-rounded assessment of the overall correctness of the model on the provided dataset. Nevertheless, these outstanding performances raise a series of issues. The remarkable levels of accuracy, precision, recall, and F1-score, which closely approach the optimal value of 1.00, give rise to relevant inquiries. A comprehensive analysis of these algorithms necessitates a contextual perspective on the dataset utilized. The crucial elements of a dataset are variety, representativeness, and class balance, especially when working with imbalanced datasets that can potentially distort performance measurements. The potential for overfitting, a condition characterized by excessive alignment between a model and its training data, is amplified in scenarios of particularly strong performance. Addressing these difficulties requires a comprehensive and varied approach. Specifically, cross-validation has emerged as a reliable method for assessing the consistency and generalizability of models. Regularization is a technique that is employed to effectively handle the complexity of a model, hence mitigating the risks associated with overfitting. The utilization of feature selection and dimensionality reduction techniques aids in simplifying the complexities of the dataset. The robustness of our findings is strengthened by many measures, including the maintenance of balanced datasets, validation using external data, comparisons with other models, and consideration of interpretability and explainability. Through the strategic implementation of these tactical maneuvers, scholars can enhance the dependability and importance of their research outcomes, while also fostering a deep comprehension of the effectiveness of these models in practical situations.

This study provides a substantial addition to the current corpus of literature by acknowledging and surmounting common limitations in current investigations regarding the prediction of stroke and diabetes. To begin with, it acknowledges the necessity of adopting a holistic strategy that integrates the rapid detection of diabetes and stroke. This innovative element enhances the overall comprehension of predictive models pertaining to chronic diseases. Although certain research studies may encounter difficulties associated with imbalanced or undiversified data, our methodology involves the integration of two separate datasets: one for stroke prediction and the other for diabetes prediction. By incorporating this information, we are able to not only increase the depth of our examination but also gain a more comprehensive understanding of the complex interplay between these health conditions. Through the adoption of this novel viewpoint, our research endeavors to furnish a predictive model that is both more precise and practical, thereby potentially transforming approaches to early detection in the healthcare industry.

4. Conclusions

In summary, our research endeavor has explored the complex domain of predictive modeling for stroke and diabetes, utilizing a wide range of machine learning methods. The commencement of our rigorous investigation involved conducting a comprehensive examination of algorithm parameters, with a particular focus on their crucial contribution to attaining the highest level of model performance. The following utilization of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost Classifier models on Stroke and Diabetes Prediction datasets revealed their strengths and intricacies. The XGBoost Classifier demonstrated remarkable performance in the comprehensive evaluation criteria, which encompass accuracy, precision, recall, and F1-score. The model showcased exceptional accuracy on both datasets and also showed notable precision, recall, and F1 score, establishing itself as a strong competitor in the field of medical diagnosis. The results of our study, supported by visual representations and statistical analysis, not only shed light on the unique capabilities of each model but also highlight the key factors that influence their predictions. The inclusion of significant attributes, such as the 'bmi_index' in the prediction of stroke and 'high blood pressure' in the prediction of diabetes, enhances our comprehension of the decision-making mechanisms employed by the models. Nevertheless, the exceptional performance of the model necessitates a careful examination of possible limitations, such as the balance of the dataset, the risk of overfitting, and the capacity to generalize the findings. The significance of cross-validation, regularization, feature selection, and other strategic techniques is underscored to tackle these issues and improve the resilience of predictive models.

The novel element of detecting stroke and diabetes concurrently fills a significant gap in the academic literature by providing a more all-encompassing and cohesive methodology. Additionally, our research endeavors to rectify shortcomings such as unbalanced datasets and insufficient comparisons with alternative models, with the ultimate goal of augmenting the dependability and importance of predictive models within the healthcare domain. This study establishes

a fundamental basis for subsequent investigations and emphasizes the capacity of integrating varied datasets to develop more resilient and practical prognostic models for chronic ailments.

The combination of algorithmic precision, thorough investigation of feature importance, and strategic evaluation of models establishes our prediction models, particularly the XGBoost Classifier, as powerful tools for proactive healthcare management. As the integration of artificial intelligence and healthcare progresses, this study not only makes a valuable contribution to the field of predictive modeling but also lays the groundwork for future advancements in the early identification and prevention of stroke and diabetes, thereby promoting a healthier and more knowledgeable society.

Acknowledgment

This research was funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under Grant No. DS2023-28-02.

Conflict of Interest

The authors declare that they have no competing interests.

References

[1] Murphy S. J., Werring D. J., “Stroke: causes and clinical features”, *Medicine (Abingdon)*, Vol. 48, No. 9, pp. 561-566, ISSN 1357-3039, 2020. DOI: 10.1016/j.mpmed.2020.06.002.

[2] Feigin V. L., Brainin M., Norrving B., “World Stroke Organization (WSO): Global Stroke Fact Sheet 2022”, *Int J Stroke*., Vol. 17, No. 4, pp. 18-29, 2022. DOI:10.1177/17474930211065917.

[3] Mujeeb Z. B., Aga S. S., Saniya N., “Pathophysiology of diabetes: An overview”, *Avicenna J Med*, Vol. 10, No. 04, pp. 174-188, 2020. DOI: 10.4103/ajm.ajm_53_20.

[4] William H. H., “Diabetes Mellitus in Developing Countries and Underserved Communities”, *Springer Cham*, ISBN : 978-3-319-41557-4, 2017. DOI:10.1007/978-3-319-41559-8.

[5] Jonathan H., Luis C. G., María del C. F., Cristina S., "Diabetes and Stroke Prevention: A Review", *Stroke Research and Treatment*, Vol. 2012, Article ID 673187, 6 pages, 2012. DOI:10.1155/2012/673187.

[6] Mandeep K., Sachin R. S., Kirti W., Farzana A., "Early Stroke Prediction Methods for Prevention of Strokes", *Behavioural Neurology*, Vol. 2022, Article ID 7725597, 9 pages, 2022. DOI: 10.1155/2022/7725597.

[7] Dev S., Wang H., Nwosu C. S., Jain N., Veeravalli B., John D., “A predictive analytics approach for stroke prediction using machine learning and neural networks”, *Healthcare Analytics*, Vol. *2*, 100032, 2022. DOI:10.1016/j.health.2022.100032.

[8] Al-Mekhlafi Z. G., Senan E. M., Rassem T. H., Mohammed B. A., Makbol N. M., Alanazi A. A., Almurayziq T. S. and Ghaleb F. A., “Deep Learning and Machine Learning for Early Detection of Stroke and Haemorrhage”, *Computers, Materials and Continua*, Vol. 72, No. 1, pp. 775 - 796., 2022. DOI:10.32604/cmc.2022.024492.

[9] Emon M. U., Keya M. S., Meghla T. I., Rahman M. M., Mamun M. S. A. and Kaiser M. S., "Performance Analysis of Machine Learning Approaches in Stroke Prediction," *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2020, pp. 1464-1469. DOI: 10.1109/ICECA49313.2020.9297525.

[10] Sirsat M. S., Fermé E., Câmara J., “Machine Learning for Brain Stroke: A Review”, *J Stroke Cerebrovasc Dis*, Vol. 29, No. 10, 105162. DOI: 10.1016/j.jstrokecerebrovasdis.

[11] Dritsas E., Trigka M., “Stroke Risk Prediction with Machine Learning Techniques”, *Sensors (Basel)*, Vol. 21, No. 13, pp. 4670, 2022. DOI: 10.3390/s22134670.

[12] Chowdary M. K., Kumar K. A., Ganesh C., Turaka R., Rao B. D. and Naik S. L., "Multiple Disease Prediction by Applying Machine Learning and Deep Learning Algorithms," *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2023, pp. 502-510. DOI: 10.1109/ICICCS56967.2023.10142766.

[13] Vijayan V. V. and Anjali C., "Prediction and diagnosis of diabetes mellitus — A machine learning approach," *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, Trivandrum, India, pp. 122-127, 2015. DOI: 10.1109/RAICS.2015.7488400.

[14] Sonar P. and JayaMalini K., "Diabetes Prediction Using Different Machine Learning Approaches," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 367-371, 2019. DOI: 10.1109/ICCMC.2019.8819841.

[15] Dagliati A., Marini S., Sacchi L., Cogni G., Teliti M., Tibollo V., De Cata P., Chiovato L., Bellazzi R., “Machine Learning Methods to Predict Diabetes Complications”, *J Diabetes Sci Technol*., Vol. 12, No.2, pp. 295-302, 2018. DOI: 10.1177/1932296817706375.

[16] Stroke Prediction Dataset. Available online: https://www.kaggle.com/code/prasadshingare/diabetes-hypertension-and-stroke-prediction/data?select=stroke_data.csv (accessed on 06 July 2023).

[17] Diabetes Prediction Dataset. Available online: https://www.kaggle.com/code/prasadshingare/diabetes-hypertension-and-stroke-prediction/data?select=diabetes_data.csv (accessed on 06 July 2023).

[18] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.info.html, (accessed on 06 July 2023)

[19] Dziuban, C. D., Shirkey, E. C., “When is a correlation matrix appropriate for factor analysis? Some decision rules”, *Psychological Bulletin*, Vol. 81, No.6, pp. 358–361, 1974. DOI:10.1037/H0036316.

[20] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed on 22nd Nov 2023).

[21] LaValley M. P., “Logistic regression”, *Circulation*, Vol. 117, No. 18, pp. 2395–2399, 2008. DOI: 10.1161/CIRCULATIONAHA.106.682658.

[22] Breiman L., Friedman J., Stone C. J., and Olshen R. A., "Classificationalgorithms and regression trees", Taylor & Francis, ISBN: 0412048418, 1984.
[23] Breiman L., "Random Forests", *Machine Learning* , Vol 45, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.
[24] Kramer O., "K-Nearest Neighbors", *In: Dimensionality Reduction with Unsupervised Nearest Neighbors, Intelligent Systems Reference Library*, Vol 51, 2013. DOI: 10.1007/978-3-642-38652-7_2.
[25] Chen T., Guestrin C., "XGBoost", *in: Proceedings of the 22ndACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, USA, 2016. DOI:10.1145/2939672.2939785.
[26] Hossin M., Sulaiman M. N., "A review on evaluation metrics for data classification evaluations", *International journal of data mining & knowledge management process*, Vol.5, No.2, 2015. DOI: 10.5121/ijdkp.2015.5201.
[27] Trailokya Raj Ojha, Ashish Kumar Jha, "Analyzing the Performance of the Machine Learning Algorithms for Stroke Detection", *International Journal of Education and Management Engineering*, Vol.13, No.2, pp. 27-35, 2023.
[28] Chauhan R., Goel A., Kaur H., Alankar B., "Machine Learning: An Analytical Approach for Pattern Detection in Diabetes", In: Kumar, R., Verma, A.K., Sharma, T.K., Verma, O.P., Sharma, S. (eds) Soft Computing: Theories and Applications, *Lecture Notes in Networks and Systems*, Vol. 627, Springer, Singapore, 2023. DOI: 10.1007/978-981-19-9858-4_12.

Authors' Profiles

Ngoc-Bich Le received his B.S. degrees at Bach Khoa University, Vietnam, and his Master's and Ph.D. in Mechatronics Science from Southern Taiwan University of Science and Technology – Taiwan in 2004, 2007, and 2010, respectively. He is currently a full-time lecturer in the School of Biomedical Engineering, International University, Vietnam National University Ho Chi Minh City, Vietnam. He published several papers in preferred Journals such as J. Biomedical Microdevices, J. Microfluidics and Nanofluidics, J. Sensors and Actuators, and many Vietnamese Engineering Books in automation, CAD, and mold design. He also presented various academic as well as research-based papers at several national and international conferences. His articles focus on Medical devices, MEMs, Microfluidics, Robotics, and AI.

Pham Thi Thu Hien received the B.S. degree in mechatronics from the Ho Chi Minh City University of Technology-Vietnam National University, Ho Chi Minh City, Vietnam, in 2003 and the M.S. and Ph.D. degrees in mechanical engineering from Southern Taiwan University of Technology and National Cheng Kung University, Tainan, Taiwan, in 2007 and 2012, respectively. She is currently a Head of Biomedical Photonics Lab in School of Biomedical Engineering, International University-Vietnam National University HCMC, Ho Chi Minh City, Vietnam. Her research interests are in the areas of polarized light-tissue studies, polarimetry, optical techniques in precision measurement to determine the optical properties of bio-samples (glucose, collagen, and tumor) or applied Artificial Intelligence (AI) models for cancer detection (skin, liver, blood, and breast), noninvasive glucose measurement, cell/tissue characterization, and laser/LED applications in treatment.

Sy-Hoang Nguyen received his B.S. degree in Biomedical Engineering and Biomechanics at International University, Vietnam National University Ho Chi Minh City, Vietnam, in 2023. His current research interest is wearable devices, applied Artificial Intelligence (AI) models for diseases early detection.

Nhat-Minh Nguyen received his B.S. degree in Biomedical Engineering and Biomechanics at International University, Vietnam National University Ho Chi Minh City, Vietnam, in 2023. He is currently full-time master student in the School of Biomedical Engineering, International University, Vietnam National University Ho Chi Minh City, Vietnam. His current research interest is bio 3D printing technology, wearable devices, applied Artificial Intelligence (AI) models for diseases early detection.

Tan-Nhu NGUYEN received a Ph.D. in Biomedical Engineering and Biomechanics at Université de Technologie de Compiègne, France, in 2020. His current research interest is muscle modeling coupled with a serious game for facial rehabilitation. He is currently a full-time lecturer in the School of Biomedical Engineering, International University, Vietnam National University Ho Chi Minh City, Vietnam.

How to cite this paper: Ngoc-Bich Le, Thi-Thu-Hien Pham, Sy-Hoang Nguyen, Nhat-Minh Nguyen, Tan-Nhu Nguyen, "AI-powered Predictive Model for Stroke and Diabetes Diagnostic", International Journal of Intelligent Systems and Applications(IJISA), Vol.16, No.1, pp.24-40, 2024. DOI:10.5815/ijisa.2024.01.03