Research on development of online system for stroke prediction

Nguyen Trong Tai[1,2,*], Nguyen Le Hai Dang[1,2a], Huynh Du Kien Hung[1,2a]

[1]Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology (HCMUT)
268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
[2]Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
[a] Undergraduate Student
[*]Corresponding Author: nttai@hcmut.edu.vn

Abstract

This paper presents a system for online stroke prediction. Firstly, a stroke prediction model is built using Random Forest machine learning method. Next, a mobile app user interface is created to collect patients' input data. Based on the trained model and the collected input data, the stroke risk for patients can be predicted. As a result, patients can take preventive measures to avoid severe outcomes from stroke. The training results indicate that the prediction accuracy exceeds 90% on the training dataset and 84% on the testing dataset. The trained machine learning model is then deployed on server as Application Programming Interface. This allows users to access and check their stroke risk from the mobile application.

***Keywords**: stroke, machine learning, prediction, online stroke prediction, stroke prediction server.*

Abbreviations

RF	Random Forest
ML	Machine Learning
API	Application Programming Interface
KNN	K-Nearest Neighbors
SMOTE	Synthetic Minority Over-sampling Technique
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
WHO	World Health Organization

Tóm tắt

Đột quỵ là một vấn đề sức khỏe nghiêm trọng xảy ra khi có sự gián đoạn trong nguồn cung cấp máu của não, do tắc nghẽn hoặc vỡ một mạch máu. Sự gián đoạn này làm mất hơi oxy và dưỡng chất cần thiết cho tế bào não, dẫn đến một loạt các hậu quả, bao gồm liệt, suy nói, suy giảm trí tuệ và thậm chí là tử vong, phụ thuộc vào mức độ và vị trí tổn thương não. Theo Tổ chức Y tế Thế giới (WHO), đột quỵ là một trong những nguyên nhân hàng đầu gây tử vong và tàn tật trên toàn cầu. Sự nhận biết sớm các dấu hiệu cảnh báo của đột quỵ có thể ngăn chặn sự nghiêm trọng của đột quỵ. Các dấu hiệu này bao gồm huyết áp cao, hút thuốc lá, tiểu đường, béo phì, cholesterol cao, rối loạn nhịp tim, thiếu vận động, tiêu thụ rượu quá mức và tiền sử gia đình về đột quỵ hoặc bệnh tim.
Bài báo này tập trung vào phân tích một tập dữ liệu chứa các yếu tố nguy cơ quan trọng của bệnh nhân và sau đó kết luận xem cái nào trong số chúng quan trọng hơn trong việc dự đoán khả năng mắc đột quỵ. Sau đó, mô hình học máy sẽ được xây dựng dựa trên tập dữ liệu này để tìm ra xem kết luận có đúng không và sử dụng mô hình này vào việc dự đoán đột quỵ. Dựa trên các nghiên cứu liên quan đến việc ứng dụng mô hình máy học trong chuẩn đoán đột quỵ, mộ hình phân loại rừng ngẫu nhiên (Random Forest) là mộ hình cho kết quả huấn luyện với các chỉ số đánh giá tốt và được lựa chọn để triển khai cho hệ thống dự báo online. Mục đích của hệ thống này là cho phép người dùng có thể kiểm tra nguy cơ đột quỵ của bản thân để có thể khám chữa kịp thời. Mô hình chuẩn đoán online bao gồm các phần cứng thu thập dữ liệu trực tiếp từ bệnh nhân, giao diện qua điện thoại, và server nhận dự liệu, chạy thuật toán dự báo và trả kết quả.
Mặc dù còn một số hạn chế về tập dữ liệu huấn luyện, kết quả của nghiên cứu đạt được ý tưởng đề ra. Tạo nên tảng cho quá trình thu thập dữ liệu hoàn thiện hơn cũng nhưng hướng tới việc phát triển các thiết bị phục vụ đo chỉ số sức khỏe liên quan.

1. Introduction

Stroke is a serious medical condition and a leading cause of death and disability [1]. It happens when blood flow to a part of the brain is blocked or a blood vessel in the brain bursts, causing brain cells to die. Key risk factors for stroke include high blood pressure, diabetes, high cholesterol, smoking, obesity, and lack of exercise. Age, gender, and family history also play roles [2, 3]. Prompt medical attention is crucial in the event of a stroke, as quick intervention can minimize damage and improve the chances of recovery. According to the World Health Organization (WHO), about 15 million individuals suffer from stroke. Among them, 5 million lose their lives, and an additional 5 million face lasting disabilities, imposing challenges on both families and communities. While strokes are infrequent in people under 40, when they do occur, high blood pressure is often the primary cause. Notably, stroke affects approximately 8% of children with sickle cell disease. Preventing stroke involves managing these risk factors through healthy lifestyle choices, medications, and regular medical check-ups.

Due to the severe consequences of strokes, stroke prediction and prevention are essential. Several methods and techniques are employed to address this need. Imaging techniques, such as MRI and CT scans, help doctors diagnose strokes and plan treatments by providing detailed images of the brain. Real-time monitoring, through wearable devices and mobile apps, continuously tracks health indicators like blood pressure and heart rate, supplying data for early warning systems. Recently, machine learning (ML) and artificial intelligence (AI) have gained popularity and attracted significant research interest. ML uses computer algorithms to analyze large datasets and predict stroke risk more accurately than traditional

Received: 15 July 2024; Accepted: 28 September 2024.

methods. These algorithms consider a wide range of factors, including health records and genetic data, enhancing the precision of stroke risk assessment.

Several risk factors contribute to the likelihood of experiencing a stroke including hypertension, smoking, drinking, age, gender, body mass index (BMI), average glucose level, frequency of regular physical activity, and family history of stroke. With the collection of these data combined with machine learning algorithms, prediction models of stroke can be made with high accuracy in order to help forecasting the possibility of having a stroke for people. As a result, early treatments could possibly be made to prevent irreversible damage or even death [1-4]. Therefore, numerous researches have been made using machine learning to predict the likelihood of having a stroke. In research [5], the authors used four different machine learning algorithms on a publicly available dataset on Kaggle. The result models achieved decent accuracy led by random forest algorithm with 96 percent accuracy, followed by decision tree, voting classifier and logistic regression with the accuracy of 94, 91 and 79, respectively. Previous studies have employed algorithms such as K-Nearest Neighbors (KNN) and Decision Tree [6, 7], with accuracy rates ranging from 73% to 96% depending on the dataset used.

The leading cause of stroke is typically high blood pressure, also known as hypertension. Hypertension can damage and weaken the blood vessels over time, making them more susceptible to the formation of blood clots or ruptures, which are the primary mechanisms behind strokes. Therefore, it is crucial to find the correct method for measuring blood pressure to notify patients if they have hypertension [3, 8].

Although numerous studies have focused on using algorithms to improve the accuracy of stroke prediction [2, 3, 6, 9-12], the application and implementation of these models remains limited. In this study, beyond training and evaluating a machine learning model, the primary objective is to develop an online deployment mechanism that can be accessed by anyone, anywhere. This will allow users to assess their initial health status and create early examination plans based on the model's predictions. By integrating the model into a mobile app, the research aims to facilitate early detection and intervention, ultimately improving health outcomes.

The primary tasks of this research are detailed as follows: First, a model of stroke prediction is trained and evaluated. Based on stroke database [13], the stroke Machine Learning (ML) model is trained and evaluated base on random forest classification machine learning algorithm. Next, the trained ML model is deployed on a server as an online prediction service. Finally, a mobile app is developed to communicate with the server, collect user data, and send it to the prediction server. Based on the collected data and the predictions generated by the model, the user's stroke risk is assessed, and the results are sent back to the mobile app.

The remainder of the paper is organized as follows: Section 2 focuses on dataset and reprocessing of data. Section 3 presents the training results using the Random Forest model. Next, Section 4 describes the construction of the server and mobile app. Finally, discussions and conclusions are provided in Section 5.

2. Dataset and Selection of Machine Learning algorithm

2.1. Brief dataset analysis

In this research the dataset is utilized from Harvard Dataverse Collection [13] The dataset includes 43400 subjects The dataset contains 11 features including gender, age, hypertension, heart disease, marriage, work type, residence type, average glucose level, BMI, smoking status and stroke status. From the dataset, some basic information about stroke rate with correspond to feature can be analyzed. Fig. 1 to Fig. 3 depict the stroke rate with respect to gender, age and hypertension features. From Fig. 1, it can be seen that the stroke rate is almost same in gender feature. It can be seen from Fig. 2 that above the age of 50, people tend to have a stroke. In addition, people with hypertension also have a higher risk of suffer a stroke according to Fig. 3.

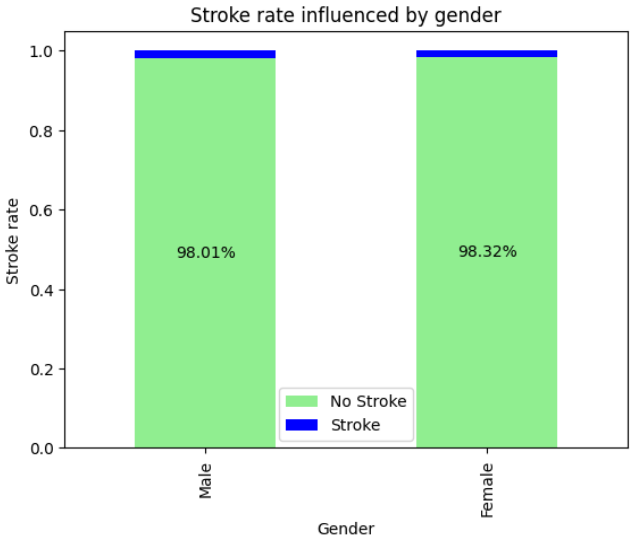

Fig. 1. Stroke rate influenced by gender.

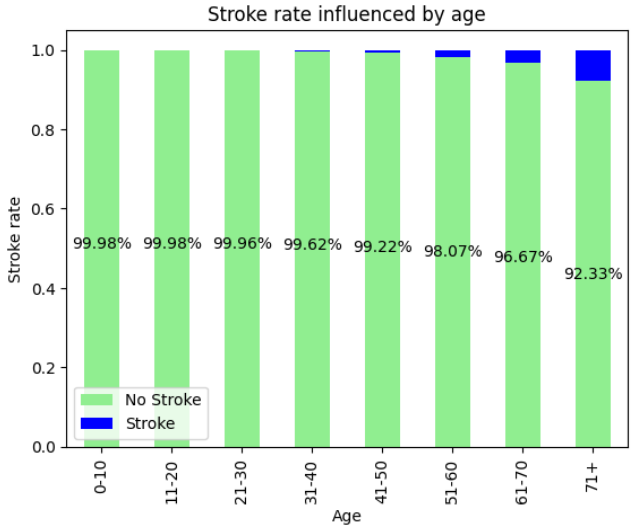

Fig. 2. Stroke rate influenced by age group.

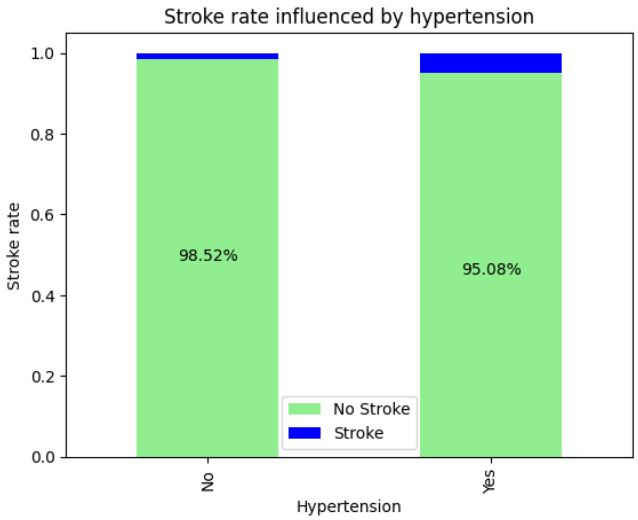

Fig. 3. Stroke rate influenced by hypertension.

2.2. Dataset processing

The data preprocessing process includes the following steps: converting Categorical data to Numerical type; handle and remove data with NaN values; divided into feature set and result set, clustering features “age”, “avg_glucose”, “bmi”; Normalize the data (used for Logistic Regression model) and finally balance the data using the SMOTE method (Synthetic Minority Over-sampling Technique).

In particular, data normalization is the technique of avoiding overfitting. Specifically in the code, our team uses 'StandardScaler' to normalize the features to the same value range, with a mean value of 0 and a variance of 1. This can help the model learn the relationship between features without being unduly influenced by large variations between them.

2.3. Selecting of Machine Learning model

Regarding stroke prediction model training, many research studies and algorithms have been investigated. Some studies can be referenced, such as [3, 11, 12, 14, 15].

According to [15], random Forest is reviewed as one of the best machine learning models for stroke prediction using the Kaggle dataset, as shown in Table 1.

Table 1: Comparisons Random Forest method and other Approaches with refer to [15]

Reference	Investigated Learning Approaches	Results
[6]	LR, DT, RF, KNN, SVM, and NB Classification	Naïve Bayes performed best in the task that gave an accuracy of 82%
[8]	LR, DT, KNN, RF, NB	Highest accuracy =95.4% utilizing Random Forest
[16]	DT, LR, NB, KNN, RF, ANN, SVM, XGBoost (Second highest accuracy achieved)	Highest accuracy= 92.32% achieved using Random Forest (AUC= 0.975)
[9]	DT, KNN, LR, NB, RF, SVM and neural network	Highest accuracy =93% using Random Forest

In addition, based on processed data, three ML models were investigated to verify in this research: Logistic Regression, Decision Tree, and Random Forest. The training results with the relevant metrics are shown in Table 2. Among these models, Random Forest delivered the best performance.

Table 2: Comparisons of 3 difference ML methods

Metrics	Machine learning algorithm		
	Logistic Regression	Decision Tree	Random Forest
Accuracy (%)	77	84	91.9
Precision (%)	77	85	92
Recall (Sensitivity) (%)	77	84	97.1
F1-score (%)	77	84	92

Based on both references and the evaluated training results, Random Forest is selected for further investigation in this research. The following section will detail the Random Forest method and its training outcomes.

3. Random forest and training result

3.1. Introduction for used algorithm

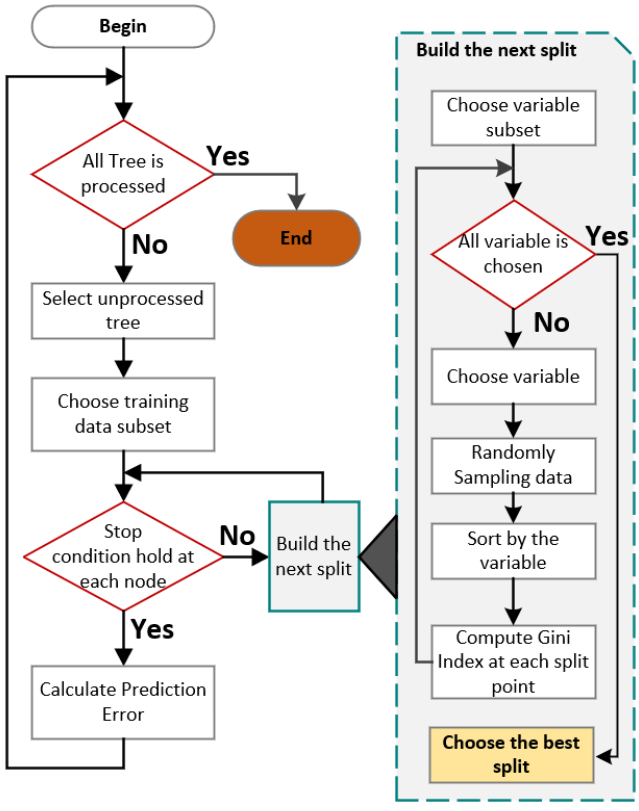

Fig. 4. Random forest algorithm diagram.

Random Forest (RF) is an ensemble learning technique that builds multiple decision trees and aggregates their predictions to improve the overall accuracy and robustness of the model. This method relies on bagging (bootstrap aggregating), where each tree is trained on a different random subset of the training

data. During tree construction, a random subset of features is selected at each node, ensuring that the decision trees are diverse and less correlated with each other. By averaging the predictions from these individual trees (or taking the majority vote in classification tasks), Random Forest reduces the risk of overfitting and increases the model's generalization ability. Additionally, RF is particularly effective for high-dimensional datasets and can handle both categorical and numerical data. Fig. 4 illustrates the operation of the Random Forest algorithm.

3.2. Training results

In order to archive the training, validation and testing results, the dataset, which comprises 434,000 samples, are divided into three sets for this research. Set 1, accounting for 72% of the dataset, is used for training. Set 2, representing 8% of the dataset, serves for validation. The remaining data is allocated for testing.

The training procedure in this research is conducted using Anaconda, which supports machine learning tasks. Additionally, Spyder IDE is employed for its intuitive interface, facilitating the development and debugging of Python code.

Fig. 5 illustrates the training results using the Random Forest model, which achieved an accuracy of 91.92%, a sensitivity of 97.1%, a precision of 92%, and an F1-score of 92%. These scores, all above 90%, indicate the strong performance of the trained model. The precision score means that 92% of the stroke predictions are correct, while the sensitivity (recall) score shows that 97.1% of actual stroke cases are identified. Additionally, the F1-score of 92% reflects the overall robustness of the model.

```
Classification Report of Random Forest Model:
                precision    recall  f1-score   support

           0       0.97      0.87      0.92      6503
           1       0.88      0.97      0.92      6282

    accuracy                           0.92     12785
   macrro avg       0.92      0.92      0.92     12785
weighted avg        0.92      0.92      0.92     12785

Feature_age: 0.3525059035420994
Feature_avg_glucose level: 0.13417216348213312
Feature_bmi: 0.1277705742318928
Feature_smoking_status: 0.10674197807756586
Feature_work_type: 0.09720219377597078
Feature_ever_married: 0.04751262503746178
Feature_gender: 0.04139152399523287
Feature_Residence_type: 0.04013309920149503
Feature_hypertension: 0.03161679832372796
Feature_heart_disease: 0.02095314033242054
Accuracy: 0.9192804067266328
Sensitivity: 0.971028334925183
Specificity: 0.8692910964170383
AUC: 0.9201597156711108
```

Fig. 5. Random forest training result.

When considering the impact of individual features, age contributes the most at 35.25%. From the dataset, both age and average glucose level are collected in detail, but age has the highest impact. Other subgroups, however, are represented as discrete or logical data, which results in lower contributions compared to the average glucose level subgroup. Although blood pressure and heart disease are known to be key factors in stroke risk, in this dataset, heart disease has the least impact at 2.09%. This is likely due to the dataset's limitations, where both hypertension and heart disease statuses are represented as binary (true/false), reducing their influence on the model. It would be beneficial if the dataset included more detailed information on blood pressure, heart disease symptoms, or heart rate, which could improve the model's ability to capture the nuances of these risk factors.

3.3. Trained Model evaluation

To evaluate the trained model from Section 3.2, it was tested on the test set. In Fig. 6, the confusion matrix of the testing results is shown. It indicates that stroke cases were correctly predicted (True Positives, TP) for 5,957 samples. Samples without stroke were incorrectly predicted as having a stroke (False Positives, FP) for 1,450 cases. Non-stroke cases were correctly predicted (True Negatives, TN) for 4,875 samples, and stroke cases were incorrectly predicted for non-stroke samples (False Negatives, FN) for 504 cases.

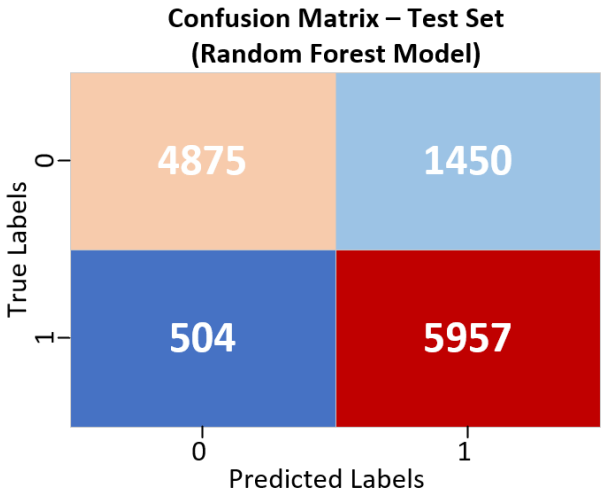

Fig. 6. Confusion matrix for test set.

From the confusion matrix, the evaluation metrics on test dataset are determined as:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

$$\text{F1\_Score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Based on above formulate, the metrics of trained model on the test data are indicated with Accuracy = 84.7%, Precision = 80.4%, Recall = 92.2%, F1 Score = 85.8%. These metrics are quite high and over 80%, indicating that the trained model can effectively predict stroke situations in real-world scenarios. The precision score shows that 80.4% of the stroke predictions are correct, while the sensitivity (recall) score shows that 92.2% of actual stroke cases are identified in the test data. This capability enables users to assess their own stroke risk with confidence.

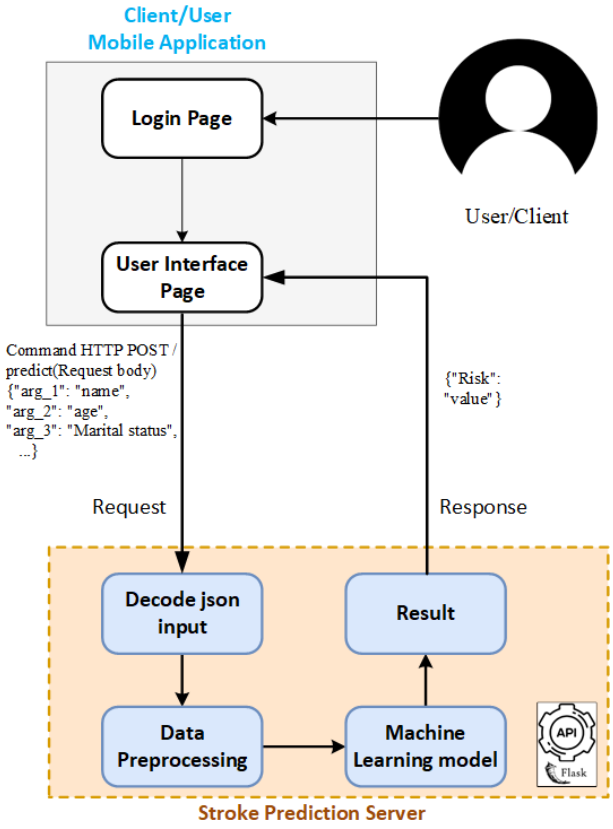

Fig. 7. Online Stroke Prediction Architecture.

4. Development of online stroke prediction system

The architecture of the online prediction system is illustrated in Fig. 7. This structure comprises two main parts: the client/user application and the prediction server. The client application collects user data and sends it to the server. The prediction server preprocesses the data, estimates the stroke risk using a machine learning model, and sends the result back to the client. Communication between the user application and the prediction server is conducted via the HTTP protocol.

4.1. Android Mobile Application

An Android mobile application has been developed for end-users using Android Studio. The UI module allows users' profiles and data to be uploaded to the server. The application has been built with the Flutter framework and Dart programming language. Upon the first login, profile data, including the name, age, and other relevant details, is provided by users. After navigating to the prediction page, detailed instructions on using the stroke risk prediction feature are received by users. Registration and login are required to access the stroke prediction function. User login data is stored in Google Cloud Firestore via a REST API. The login and instruction interfaces of the mobile application are depicted in Fig. 8.

User information is supplied through the information form; after all data fields are filled, the data is uploaded to the server. The result is then determined by the trained model on the server. Subsequently, the result is sent back to the user and displayed on the interface. Fig. 9 shows the user interfaces for information filling and prediction results.

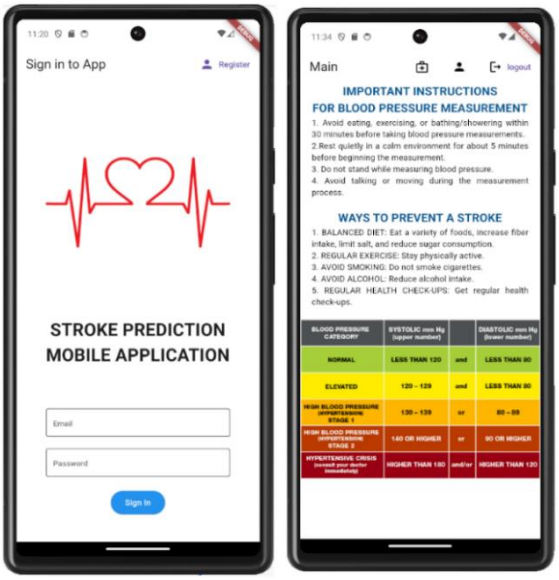

Fig. 8. Android app user interface.

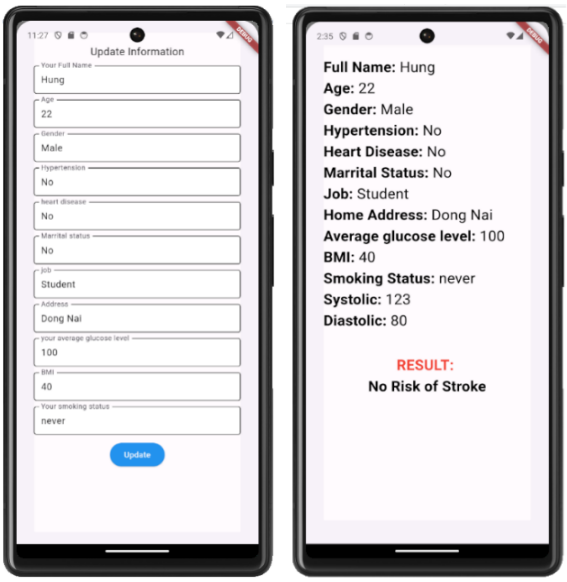

Fig. 9. User data update and predict result interface

4.2. Stroke Prediction Server

The server is responsible for receiving requests from the user application, processing data, and responding with the predicted stroke situation to the user. In this research, the server has been

developed using Visual Studio Code, and the machine learning model has been trained using the Spyder IDE.

The best-performing trained ML model from Section III is deployed using the Flask framework as an API. User data, including essential information, is sent from the Android application via an API request. Subsequently, the relevant data needed for the prediction process is extracted, normalized, and input into the predictive model. The predicted outcomes are then sent back and displayed within the user's application interface.

In this research, the server is executed locally. To enable access from anywhere, the server needs to be placed in a public domain. Therefore, Ngrok platform is used to overcome this challenge. Ngrok is a cross-platform application that creates secure tunnels to a localhost machine, allowing the exposure of a local development server to the Internet with minimal effort. After developing the Android application and deploying the server, the testing results are displayed in Fig. 9.

5. Conclusion and discussion

In this paper, an online stroke prediction system is developed and investigated. Using Harvard's dataset, a machine learning model for stroke prediction is built with the Random Forest algorithm. The trained model achieves an accuracy of 92% during the training phase and 84% during testing. An online prediction server is then established to process client data and return stroke risk results. The machine learning model is deployed on this server using the Flask framework as an API. Additionally, an Android application is developed to facilitate interaction between users and the server.

Through this mobile application, users can input their information and receive predictions of stroke risk. Furthermore, users can receive daily advice for maintaining a healthy lifestyle and ongoing stroke risk predictions, helping them take preventive measures and raise awareness.

A key advantage of our system is the integration of the trained model into an online server, paired with a mobile application, which allows users to interact with the model remotely and receive real-time stroke risk predictions. By utilizing an online server, the prediction model can be easily updated and improved. Despite the promising results, this study has several limitations. First, the dataset used may not fully represent the general population, introducing potential bias into the model’s predictions. The data were collected from a specific cohort, which may limit the model's generalizability to other demographics or regions. Additionally, the binary classification of certain risk factors, such as hypertension and heart disease, reduces the granularity of these features, potentially affecting the model's ability to capture subtle variations in stroke risk. Finally, this system has not yet been validated in real-world clinical environments.

Future work should focus on validating the model in real-world settings to ensure its practical applicability and expanding the system to incorporate more diverse data sources. Moreover, while our system primarily focuses on structured data from patients, further enhancements could involve integrating real-time data from wearable devices for continuous monitoring, a feature present in some of the latest stroke prediction systems.

Acknowledgement

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant **number C2021-20.**

We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for this study.

References

[1] S. A. Chohan, P. K. Venkatesh, and C. H. How, "Long-term complications of stroke and secondary prevention: an overview for primary care physicians," (in eng), *Singapore Med J*, vol. 60, no. 12, pp. 616-620, Dec 2019.

[2] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics*, vol. 2, p. 100116, 2022/11/01/ 2022.

[3] A. Hassan, S. Gulzar Ahmad, E. Ullah Munir, I. Ali Khan, and N. Ramzan, "Predictive modelling and identification of key risk factors for stroke using machine learning," *Scientific Reports*, vol. 14, no. 1, p. 11498, 2024/05/20 2024.

[4] B. Borsos, C. G. Allaart, and A. van Halteren, "Predicting stroke outcome: A case for multimodal deep learning methods with tabular and CT Perfusion data," *Artificial Intelligence in Medicine*, vol. 147, p. 102719, 2024/01/01/ 2024.

[5] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches," *Journal of Healthcare Engineering*, vol. 2021, no. 1, p. 7633381, 2021/01/01 2021.

[6] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 12, no. 6, 2021.

[7] S.-F. Sung *et al.*, "Developing a stroke severity index based on administrative data was feasible using data mining techniques," *Journal of Clinical Epidemiology*, vol. 68, no. 11, pp. 1292-1300, 2015/11/01/ 2015.

[8] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Analytics*, vol. 2, p. 100032, 2022/11/01/ 2022.

[9] L. García-Temza, J. L. Risco-Martín, J. L. Ayala, G. R. Roselló, and J. M. Camarasaltas, "Comparison of Different Machine Learning Approaches to Model Stroke Subtype Classification and Risk Prediction," in *2019 Spring Simulation Conference (SpringSim)*, 2019, pp. 1-10.

[10] R. Gurjar, S. K, N. C, S. Sathish, and R. S, "Stroke Risk Prediction Using Machine Learning Algorithms," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 20-25, 07/01 2022.

[11] S. Mushtaq and K. S. Saini, "A Review on Predicting Brain Stroke using Machine Learning," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2023, pp. 667-673.

[12] A. Byna, M. Modi Lakulu, and I. Yusuf Panessai, "Current critical review on prediction stroke using machine learning," *Bulletin of Electrical Engineering and Informatics; Vol 13, No 5: October 2024DO - 10.11591/eei.v13i5.7435*, 10/01/ 2024.

[13] M. M, "Replication Data for: Prediction of Cerebral Stroke," DRAFT VERSION ed: Harvard Dataverse, 2021.

[14] A. S. Niharika Patil, "Stroke Prediction Using Machine Learning," *Journal of Research in Engineering and Computer Sciences*, vol. 2, no. 1, pp. 61-72, 2024.

[15] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," *European Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 23-30, 01/20 2023.

[16] J.-A. Tavares, *Stroke prediction through Data Science and Machine Learning Algorithms*. 2021.