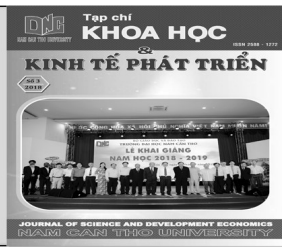
Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn



Heart disease prediction using multilayer perceptron in a dynamic environment

Le Thi My Nhu[1], Ngo Ho Anh Khoi[1], Duong Duy Khanh[2]
[1]*Faculty of Information Technology, Nam Can Tho University*
[2]*Hanoi Tam Anh Hospital*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *cardiovascular diseases, heart failure prediction, multilayer perceptron*

Từ khóa: *bệnh tim mạch, cảm biến đa lớp, dự đoán suy tim*

ABSTRACT

In recent years, the incidence and mortality rates due to cardiovascular diseases have been on the rise globally. This is the primary reason why the main objective of this topic is to investigate techniques aimed at solving the problem of heart disease diagnosis. The research methodology for this topic involves the use of the scientific experimental approach, conducted on the Multilayer Perceptron (MLP) algorithm using the Heart Failure Prediction Dataset as the foundational dataset. This research addresses a highly significant societal issue. If further studied and developed, it has the potential to empower individuals to proactively and effectively prevent heart diseases. The prediction of heart disease has become a crucial field of study, aiding in early detection, risk assessment, and the implementation of preventive measures. This article summarizes several important aspects related to heart disease prediction based on scientific machine learning methods.

TÓM TẮT

Trong những năm gần đây, tỷ lệ mắc và tử vong do bệnh tim mạch ngày càng gia tăng trên toàn cầu. Đây là lý do chính tại sao mục tiêu chính của chủ đề này là nghiên cứu các kỹ thuật nhằm giải quyết vấn đề chẩn đoán bệnh tim. Phương pháp nghiên cứu cho chủ đề này bao gồm việc sử dụng phương pháp thực nghiệm khoa học, được thực hiện trên thuật toán Perceptron đa lớp (MLP) sử dụng bộ dữ liệu dự đoán suy tim làm bộ dữ liệu cơ bản. Nghiên cứu này đề cập đến một vấn đề xã hội rất có ý nghĩa. Nếu được nghiên cứu và phát triển thêm, nó có khả năng trao quyền cho các cá nhân ngăn ngừa bệnh tim một cách chủ động và hiệu quả.

Dự đoán bệnh tim đã trở thành một lĩnh vực nghiên cứu quan trọng, hỗ trợ phát hiện sớm, đánh giá rủi ro và thực hiện các biện pháp phòng ngừa. Bài viết này tóm tắt một số khía cạnh quan trọng liên quan đến dự đoán bệnh tim dựa trên phương pháp học máy khoa học.

1. INTRODUCTION

In recent years, cardiovascular diseases have been confirmed as the leading cause of death globally, including in Vietnam. Heart diseases often progress silently, with many cases going undetected as they do not cause significant pain. Individuals with heart conditions might continue working and engaging in normal activities, leading to complacency. Older adults are particularly vulnerable to heart diseases. The elderly population is susceptible due to gradual damage to blood vessels over time, often resulting in the accumulation of arterial plaques. Recommendations include implementing comprehensive strategies and approaches across the entire population to address the burden of heart disease in Vietnam. Health education, increasing awareness, and modifying behaviors among heart disease patients are crucial for early detection, prevention, and timely treatment. These efforts can help slow down the progression of heart diseases, improve the quality of life for affected individuals, and simultaneously alleviate economic burdens on families and society as a whole.

In the article by Schocken and colleagues, the increasing prevalence of heart failure worldwide, including in developing regions like Vietnam, poses significant challenges for caregivers, researchers, and policymakers. Consequently, prioritizing the prevention of this global catastrophe is of utmost importance. In the article by Ashrafian and colleagues, the authors discuss how neural resistance has successfully reduced the incidence and mortality rates of heart failure. Further discussions delve into therapies based on new mechanisms to enhance metabolic processes and insulin resistance in heart failure (Ashrafian et al., 2007) [1]. In the report by Virani and collaborators, the focus is on the American Heart Association's collaboration with the National Institutes of Health to provide an annual updated statistical report on heart disease, stroke, and cardiovascular risk factors. This statistical update presents the latest data on a range of cardiovascular conditions including stroke, congenital heart disease, arrhythmias, atherosclerosis, coronary artery disease, heart failure, and valve diseases. (Virani et al., 2020) [2]. Trang (2018) [3] focused on assessing clinical and preclinical characteristics in patients admitted for acute heart failure due to localized coronary ischemia. The study investigates factors promoting acute events and their correlations with short-term outcomes. Lastly, Vu Thi Thom addresses the rising incidence and mortality rates of coronary heart disease in Vietnam. Thus, they conducted a cross-sectional descriptive study on 269 university staff members in Hanoi in 2016, aged 20 to 64. The study reveals higher risk factor rates in males, with the university staff having lower rates of hypertension and dyslipidemia compared to the general community. Gender

was found to be correlated with overweight and obesity, high blood pressure, and dyslipidemia. (Vu Thi Thom, 2018) [4].

In Colombet (2000) [5], the authors conducted a comparison of three algorithms, CART, Multilayer Perceptron, and logistic regression, to predict cardiovascular risk from real-world data. Estimating multivariate risk is currently required in cardiovascular disease prevention guidelines. Limitations of existing statistical risk models have led to the exploration of machine learning methods. The research data was randomly divided into a training set (n=10,296) and a testing set (n=5,148). The accuracy results for the three algorithms were as follows: 65.9%, 76.0%, 69.1%. In Yan et al (2006) [6], the authors studied a Multilayer Perceptron-based decision support system developed to aid in heart disease diagnosis. The system's input layer comprises 40 input variables categorized into four groups, encoded using proposed encoding schemes. The number of nodes in the hidden layer is determined through layer-wise learning. Each of the 5 nodes in the output layer corresponds to a specific heart disease. In cases of missing patient data, the system employs mean replacement for handling missing values. Furthermore, an enhanced backpropagation algorithm is utilized for system training. A total of 352 medical records were gathered from patients with 5 heart diseases, used for system training and testing. Cross-validation, holdout, and bootstrapping methods were applied to evaluate the system's generalization. The results demonstrate that the proposed MLP-based decision support system achieves highly accurate diagnosis (>90%) with relatively low time overhead (<5%), showcasing its utility in aiding the heart disease diagnostic process.

2. MATERIALS AND METHODS

During the process of data exploration for the research topic, numerous datasets were found (approximately 1163 datasets related to heart disease issues). However, out of these, only four datasets exhibited the highest availability and comprehensive parameter details, including:

- Heart Failure Prediction Dataset by Fedesoriano: This dataset contains 918 rows and 12 columns, encompassing variables such as age, gender, chest pain type, resting blood pressure, maximum heart rate, cholesterol level, etc. (Fedesoriano, 2021) [7].
- Heart Failure Prediction by Larxel (2022) [8]: With 299 rows and 13 columns, this dataset includes variables like age, gender, smoking duration, ischemia status, etc.
- Heart Disease Dataset by Yasser (2021) [9]: Consisting of 304 rows and 14 columns, this dataset covers variables like age, gender, chest pain type, resting blood pressure, exercise-induced chest pain, etc.
- Heart Diseases by Kakaraparthi (2022) [10]: This dataset comprises 303 rows and 14 columns, featuring variables like age, gender, chest pain type, resting blood pressure, maximum heart rate, cholesterol level, major vessels count, etc. Although these datasets are all related to predicting heart diseases, only one dataset meets the requirements of the project due to issues in other datasets such as insufficient data or outdatedness. The dataset must be numerical, have specific class labels, and contain multiple fields to provide objective outcomes. The only dataset meeting these criteria is the Heart Failure Prediction Dataset.

Current databases face a significant issue in adapting to changing data over time. Classical algorithms, on which existing databases are often trained, can only be trained once and need to be re-learned from scratch when data changes. In modern data environments, data evolves continuously, necessitating quick adaptation to these changes. To address this, continuous learning in a non-stable environment has been proposed as an alternative solution. Continuous learning allows databases to learn data continuously in a changing environment, facilitating updates and adjustments to their predictive models. This enhances the database's adaptability to data changes and improves the accuracy of their predictive models.

The term "concept drift" has been widely used in problems in dynamically changing environments like heart prediction. In fact, the concept of drift forms the basis for gradual changes, continuous shifts, and the "forgetting" of previous situations. However, in dynamically changing environments, the complexities often result in scenarios that change rapidly, slowly, or even involve the reappearance of previously disappeared knowledge. In such intricate situations, the notion of "dual dilemma" regarding stability or adaptability gains its full significance. It's important to note that these approaches aren't universally suitable and vary based on real-world conditions.

The "Sliding Windows" approach is one of three methods proposed to address concept drift. This method involves considering the evolution of concepts in a non-stable environment by using a "sliding window" as seen in the FLORA approach. The principle involves updating the model at each time step using the most recent training data defined by a window of time or data count. This approach can reclassify data into "groups" (based on temporary window data) or update models if online learning methods permit. The challenge lies in determining the window size, which is often fixed for practical applications. The dataset is temporary, reflecting diverse emerging standards, potentially leading to changes in dataset classification and indicating an unstable environment. Consequently, traditional algorithms designed for static environments aren't viable. Incremental learning algorithms that accommodate dynamic environments must be used.

Considering the mentioned characteristics and drawing from historical research on various methods, the upcoming research will employ the Sliding Windows method, which is the most suitable approach. This method will be combined with the Multilayer Perceptron algorithm.

The Heart Failure Prediction Dataset is provided by the author FEDESORIANO, who is a member of the Kaggle website. The dataset was last updated on September 11, 2022, according to the Vietnam time zone. Upon downloading, the dataset is in raw form, which cannot be directly used for this study. It requires substantial transformation to convert various parameters and characteristics into a usable format. Raw data needs preprocessing before it can be applied to the software training process.

The following transformations are performed:

- Sex Attribute: M representing male is converted to "1", and F representing female is converted to "0".
- ExerciseAngina Attribute: Y is changed to "1", N is changed to "0".

- ChestPainType Attribute: ASY is converted to 1, TA to 2, ATA to 3, NAP to 4.
- FastingBS Attribute: If fasting blood sugar > 120, it's converted to "1", otherwise "0".
- RestingECG Attribute: Normal is changed to 1, ST to 2, LVH to 3.
- ST_Slope Attribute: Up is changed to 1, Flat to 2, Down to 3.

Other attributes (Age, RestingBP, MaxHR, Oldpeak, Cholesterol) are retained as they are numeric. After these transformations, the dataset must be saved in ".csv" format. Each row should be entered in a single cell, with parameters and labels separated by commas. Parameters should be placed before the label as per system regulations.

The processed dataset will have 12 features, totaling 918 data entries, with attributes like Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope, and Prediction Label.

For the experimental setup, the dataset is divided into two parts: a training dataset (642 data entries, 70% of the original data) and a testing dataset (276 data entries, 30% of the original data). The positions of these data entries are shuffled randomly for each experiment and are reshuffled after training.

The experiment is conducted through an indirect experimental model (Batch Learning, batch size = 641), where each step involves one data entry. Increasing the batch size equals 70% of the original data and helps improve accuracy, particularly for large, complex datasets. However, it's time-consuming.

All achieved results are based on Balanced Accuracy, which is particularly useful for evaluating binary classifier performance in imbalanced classes. It's more complex than traditional Accuracy calculations. Balanced Accuracy addresses the issue of accuracy skewing when classes are imbalanced (e.g. one data entry in class A, 999 in class B). It calculates percentages based on true negatives, true positives, false negatives, and false positives.

The formula for Balanced Accuracy used in these experiments is:

First, reference the confusion matrix:

	True Ground Truth Labels (1)	**Incorrect Ground Truth Labels (0)**
True positive predictions (1)	TP	FP
False positive predictions (0)	FN	TN

It's important to have:

Correct, TPR (True Positive Rate) is also known as the sensitivity or recall, and it represents the ratio of true positive predictions (correctly detecting the positive class) to the actual positive instances in the dataset. It's calculated using the formula: TPR = $\frac{TP}{TP+FN}$

The TNR (True Negative Rate) is also known as specificity and represents the ratio of true negative predictions (correctly detecting

the negative class) to the total actual instances of the negative class in the dataset. It's calculated using the formula: $\text{TNR} = \frac{TN}{TN+FP}$

After obtaining these two metrics, Balanced Accuracy is calculated using the formula: $Balance\ Accuracy = \frac{TPR+TNR}{2}$

3. RESULTS AND DISCUSSION

Using artificial intelligence methods, specifically algorithms like the Multilayer Perceptron, in combination with the Sliding Window technique, we have applied a comparative approach. Utilizing charts, we compare the average results of the algorithm. The experimental average results of the algorithm are represented in the chart below (Figure 1).

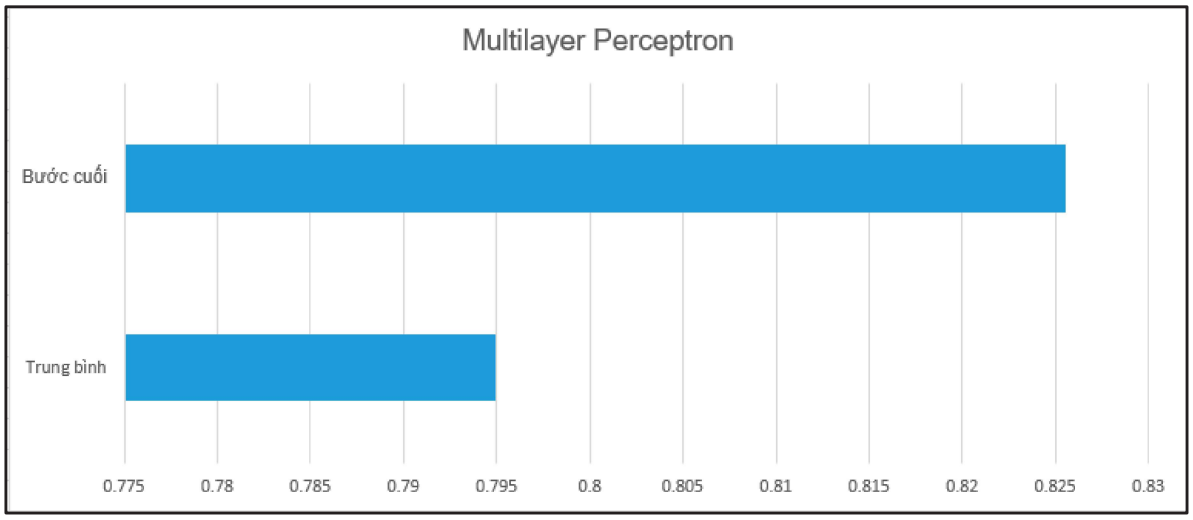


Figure 1. Chart comparing the average percentage and final step of the Multilayer Perceptron algorithm

Analyzing the data from the chart, we can deduce the following insights about the average percentage and the final step of the Multilayer Perceptron algorithm:

The data illustrates that the Multilayer Perceptron achieves a stable and remarkable performance, with an accuracy rate at the final step exceeding 80% (specifically 82.55%), and an average accuracy of 79.49%. Apart from comparing the accuracy percentage between the average and final steps, it's also possible to assess the algorithm in a more comprehensive and detailed manner by examining each step. This allows for a clearer evaluation of the algorithm's progress through the accuracy progression chart of the Multilayer Perceptron presented below (Figure 2).

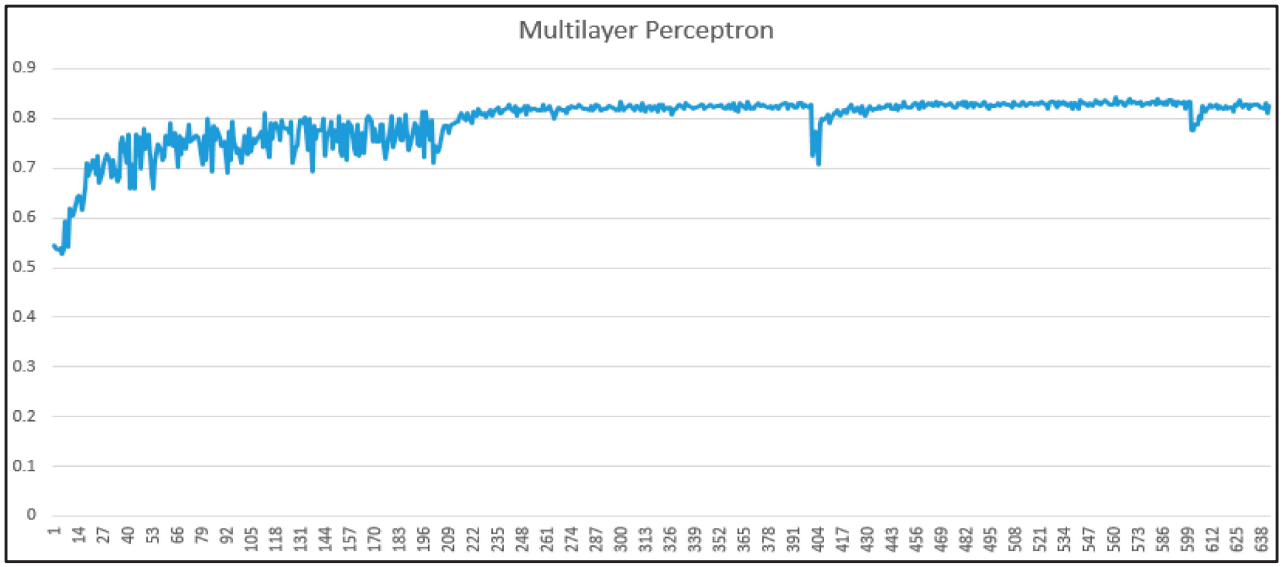


Figure 2. The accuracy progression chart of the Multilayer Perceptron algorithm

Observing the accuracy progression chart of the Multilayer Perceptron algorithm, we notice that the algorithm starts with a relatively low point, reaching only 55.45% accuracy. However, the stability has gradually increased over time. Specifically, at the beginning, the algorithm achieves an average accuracy of 55.45%, but by step 9, it has risen to 61.80%. Subsequently, there is a continuous increase in accuracy over the following steps, maintaining stability. Looking at the chart, we can see that the Multilayer Perceptron reaches its highest ratio at step 561, achieving an accuracy rate of 84.11%, which is quite remarkable. It's evident that the greatest growth in accuracy occurs in the phase from step 546 to 561, as illustrated in the graph below (Figure 3).

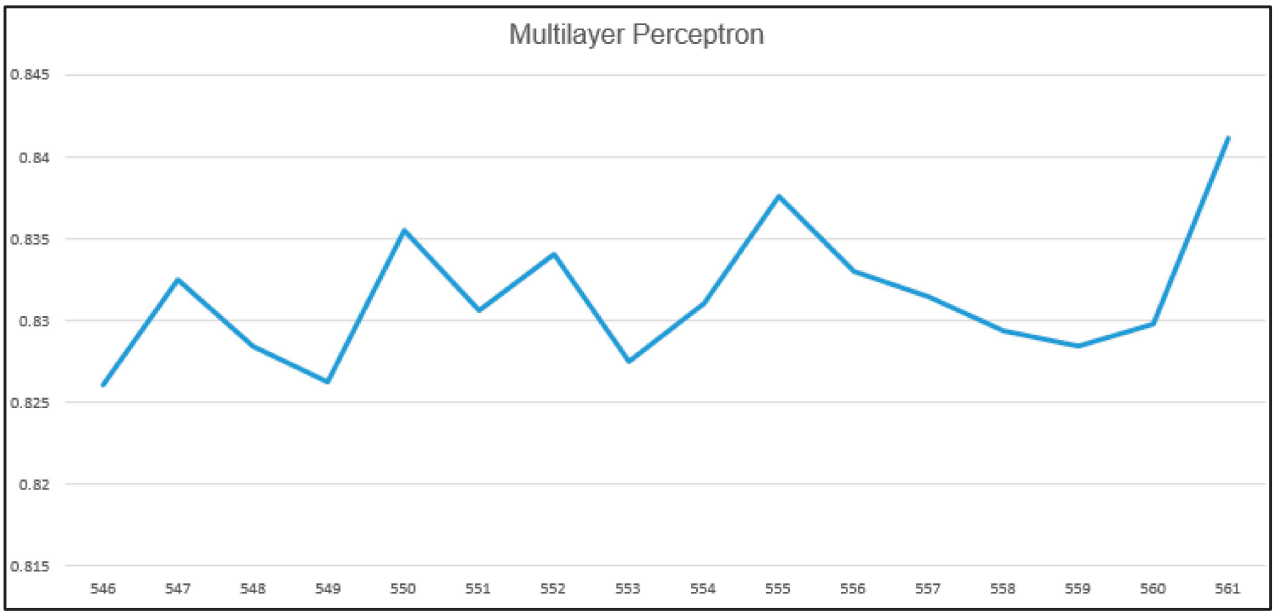


Figure 3. The highest level stage of the Multilayer Perceptron algorithm (Batch 546-561)

Examining the chart, we can clearly observe the steady and consistent increase of the Multilayer Perceptron algorithm's performance, consistently reaching above the 80% mark. The Multilayer Perceptron consistently proves itself to be a solid algorithm, even though there are some stages where the results are not favorable. However, the algorithm quickly manages to achieve a good state and maintains stability, as depicted in the graph below (Figure 4).

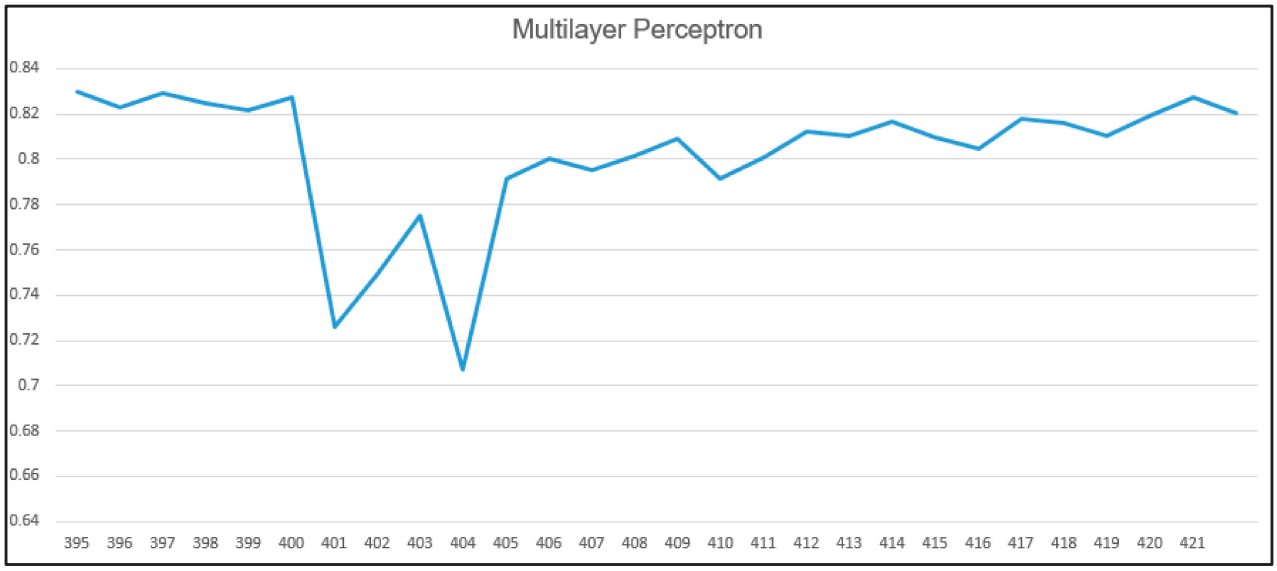


Figure 4. The lowest level stage of the Multilayer Perceptron algorithm (Batch 623-642)

In steps 401-405, the lowest percentage range of the Multilayer Perceptron algorithm fluctuated between 72.61% and 79.17%. However, at step 406, the algorithm quickly returned to its stable performance, achieving 80.03%. In the subsequent steps, although there were fluctuations in the percentage between the steps, the differences were not substantial. Overall, the Multilayer Perceptron algorithm showcases several advantages and relatively consistent performance with an accuracy rate above 80%. This indicates that the Multilayer Perceptron can be a valuable and useful choice in various prediction tasks.

Demo installation:

Each function of the system has been completed and fulfills the initial requirements. The system is divided into distinct sections for regular users and developers. Regular users have access to a form for predicting cardiac diseases, while developers have additional forms to work within the system, including a login form, model list form, data training form, and settings form. The system holds a high potential for practical application due to the increasing prevalence of cardiac issues in Vietnam, driven by unhealthy lifestyles and habits. Moreover, diagnostic standards for cardiac diseases are subject to change over time due to various objective reasons. Consequently, the dataset will undergo significant changes, and the Multilayer Perceptron algorithm is fully equipped to adapt to these modifications. There are two primary functions: Algorithm Setup (admin or developer) and Diagnostics (user).

Diagnosis User Interface:

The homepage of this system is a heart disease prediction page, constructed with the following input parameters: Age, Sex, ChestPainType, RestingBP, Cholesterol Level, Fasting Blood Sugar Level, Resting Electrocardiographic Results, Max Heart Rate Achieved, Exercise-Induced Angina, ST Depression Induced by Exercise Relative to Rest, and ST Slope. Each parameter is of the floating-point data type. On the diagnosis form designed for users, there will be a button labeled "Start Prediction" and 11 input fields for entering diagnosis data.

Figure 5. Interface for Prediction Users

In this project, the system has been packaged as a ZIP file named "ChanDoanBenhTim-main.zip". After users download and extract the file, there will be a directory named "ChuanDoan". Inside this directory, there are files to run the program. The system requires your computer to be constantly connected to the internet and has the following minimum configuration: Windows 10, 2GB RAM, and 10GB of hard drive space. To run the software, follow these steps:

Python Installation: Navigate to the "SETUP" directory and run the file "python-3.9.9-amd64.exe" to install Python version 3.9.9. Library Installation: Run the file "CaiThuVien.bat" in the same directory to install the required libraries. Start the Server: Run the file "RunServer.bat" to start the program. Access the Application: Open your web browser and access the address 'http://127.0.0.1:8000/'. The program will be running on the default port '8000'. By following these steps, you will be able to install and run the heart disease diagnosis software.

5. CONCLUSION

The completion of the research project as well as the report writing marks a comprehensive achievement. Concerning the software aspect, the Multilayer Perceptron algorithm has been successfully integrated into the program, enhancing diagnostic accuracy and simultaneously addressing dynamic data

processing challenges that other algorithms have yet to tackle. A user interface has been developed, implementing the software through two approaches: command line and graphical interface. In terms of the report, all initially outlined chapters have been satisfactorily completed, with clear content elucidating data, charts, and algorithms in a specific manner. This project has only advanced to the research phase; therefore, there remain numerous opportunities for expansion in the future. These include incorporating automated raw data processing directly within the system, optimizing model training processes, refining the user interface for smoother operation, transitioning the application to practical use, facilitating swift cardiac disease diagnoses, and further diversifying functionality - potentially extending it to mobile platforms.

REFERENCES

[1] Ashrafian, H., Michael, P., Frenneaux, M.D., & Opie, L.H. (2007). *Metabolic Mechanisms in Heart Failure.* https://www.ahajournals.org/doi/epub/10.1161/CIRCULATIONAHA.107.702795

[2] Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., & Carson, A. P. (2020). Heart disease and stroke statistics 2020 update: a report from the American Heart Association. *Circulation, 141*(9), e139-e596.

[3] Trang, T. T. H., Phong, P. Đ., & Hạnh, V. Đ. (2018). Nghiên cứu một số yếu tố thúc đẩy suy tim cấp và biến cố ngắn hạn ở bệnh nhân suy tim mạn tính do bệnh tim thiếu máu cục bộ. *Tạp chí Tim mạch học Việt Nam*, (84+ 85), 138-144.

[4] Vu Thi Thom, Vu Van Nga, Do Thi Quynh, & Vu Thi Mai Anh (2018). Các yếu tố nguy cơ mắc bệnh tim mạch của nhân viên một trường đại học tại Hà Nội. *Tạp chí Khoa học ĐHQGHN: Khoa học Y Dược*, Tập 34, Số 2 (2018) 89-96

[5] Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P., & Jaulent, M. C. (2000). Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. In *Proceedings of the AMIA Symposium* (p. 156). American Medical Informatics Association.

[6] Yan, H., Jiang, Y., Zheng, J., Peng, C., & Li, Q. (2006). A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications*, *30*(2), 272-281.

[7] Fedesoriano (2021). *Heart Failure Prediction Dataset.* https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

[8] Larxel (2022). *Heart Failure Prediction*. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data.

[9] Yasser, H. (2021). *Heart Disease Dataset.* https://www.kaggle.com/datasets/yasserh/heart-disease-dataset.

[10] Kakaparthi, C. (2022). *Heart_Diseases*. https://www.kaggle.com/datasets/charankakaraparthi/heart-dissease.