Predicting the Level of Hypertension Using Machine Learning

Pham Thu Thuy[1], Nguyen Thanh Tung[2(✉)], and Chu Duc Hoang[3]

[1] Science and Technology Department, Vietnam National University, Hanoi, Vietnam
phamthuthuy@vnu.edu.vn
[2] International School, Vietnam National University, Hanoi, Vietnam
tungnt@isvnu.vn
[3] Ministry of Science and Technology, Hanoi, Viet Nam
hoangcd@gmail.com

Abstract. In recent years, data mining has been put into research and application in many different areas in the world such as economy, education, sports, telecommunications, etc. And the health - health care [1] sector is not out of this trend. If it is possible to successfully analyze the data [2–4] from the huge amount of data of diseases, patients and hospitals every day, it can help a lot of doctors in the process of diagnosis, examination and treatment of diseases for patients. The problem raised here is whether we can accurately diagnose the patient's disease based on the information provided. The information provided may be age, gender, occupation, symptoms, test information, etc. from which it is necessary to achieve the most accurate diagnosis possible to minimize the work pressure for the medical team as well as minimize the time of diagnosis.

Keywords: Machine learning · Data mining · Healthcare · Hypertension

1 Introduction

We has chosen hypertension to test this problem because in the future, hypertension will be very popular because of the aging of the population worldwide. Input data will be the information of patients related to hypertension, and the output information is the result of predicting the level of hypertension patients based on the information given. Details are given in Sect. 3.

There are numerous factors which can cause hypertension [5].

To begin with, aging plays a role and aged people have higher risk of hypertension. Gender is another cause of high blood pressure. Although women of productive age are more likely to suffer from cardiovascular disease than men, the proportion of men under 45 years with high blood pressure is higher than that of women.

In addition, secondary hypertension is also the result of a consequence of several conditions such as: kidney disease, thyroid disease, adrenal adenoma, neurological diseases such as mental disorders, diabetes, and atherosclerosis and so on.

Moreover, hypertension also results from unhealthy lifestyles. First, overweight and obesity increases the risk of hypertension. Obesity is when people's bodies excess their

P. C. Vinh and A. Rakib (Eds.): ICCASA 2020/ICTCC 2020, LNICST 343, pp. 113–122, 2021.
https://doi.org/10.1007/978-3-030-67101-3_9

body weights (BMI in men is more than 25, women are over 30). The formula for calculating BMI is as follows: weight (kg) divides (height × height) (meters). Others factors include: sedentary lifestyle, unhealthy eating, too salty eating (the amount of salt exceeds 5 g/day), too much alcoholic drinking, smoking, dyslipidemia and diabetes, frequent stress and social factors which cause urban people to have higher chances of having urban diseases than the ones living in rural areas due to stressful and urgent life styles.

Research has proved that early symptoms of hypertension can be recognized by some related tests such as blood biochemical tests (blood urea, blood uric acid, blood creatinine, blood electrolytes) to determine the diagnosis of kidney-related diseases - kidney disease increases the likelihood of hypertension. Other tests include Blood glucose test which relates to diabetes. Test index which evaluates blood lipid disorders (Cholesterol, Triglyceride, HDL-C, LDL-C, XN: Electrolyte Na; Potassium) also helps monitor and detects some diseases such as blood lipids, Atherosclerosis and hypertension [6].

2 Related Works

There are more and more researches related to diagnosing patients from the medical records received. We can model the classification problem so each class is a diagnostic result of the disease for the patient. The method of identification given here is to use machine learning [7] method. The algorithms used are specialized algorithms for classification problems such as Naïve Bayes, Association Rule, KNN, Decision Tree. Specifically, with this problem, information about patients will play an important role in classification. So we will have to consider which information is needed for classification, which information to keep and which information should be removed. Thus, we have the machine learning training process [8, 9] to carry out the diagnostic test described in Fig. 1:

Fig. 1. Training process

In which:

- **Data** is the original data set of the disease, including multiple data streams with many information about patients.
- **Feature Filter** is the process of filtering and retrieving features from data. Only those features that are needed will be retained for later training.
- **Training Data** is the data set for training, after removing unnecessary features in the original data set.
- **Training** is the training process based on the training data set using a machine learning algorithm (SVM, Naïve Bayes, etc.).

- **Model** is a file that retains the information and training results after using the training machine algorithm on the data set. This file will be used to predict future results.

After the training is complete, we will make predictions for the data we want to label. The prediction process is described as in Fig. 2:

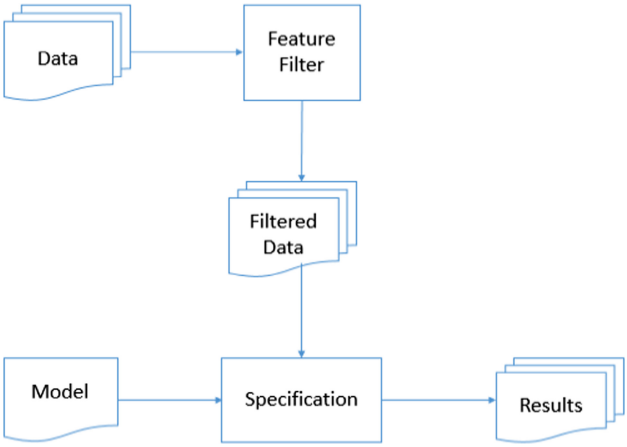


Fig. 2. The process of labeling predictions

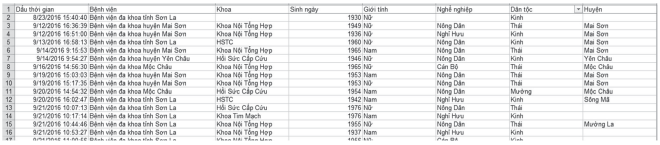
In which:

- **Model** is the model file we get from the training process.
- **Data** is the input data that we want to classify.
- **Feature Filter** is the process of filtering and retrieving features from data. Only those features that are needed will be retained.
- **Filtered Data** is the data set retained after the features that are not necessary for the classification have been removed.
- **Specification** is the process of labeling the data after it has been filtered and given the results.
- **Results** is a set of label results corresponding to the input data set.

3 Select and Build Data to Test the Diagnosis of Hypertension

All the data for the paper is collected from the medical records for hypertension patients from Son La province hospital in the national project "Application and deployment of software system for integrating and connecting biomedical electronic devices and communication networks to support the monitoring of health and epidemiology in the Northwest region" [10]. Initial data for the test is a set of statistical data for hypertension collected by the author from the hospital. All of the data is the hypertension medical records in the paper form. Data was collected, processed and stored in excel

file in tables, with each line containing information for each patient. On each line, each column will represent an information about that patient such as birth date, province, district, information about the patient's health, status at admission, etc. (Fig. 3).



	Dấu thời gian	Bệnh viện	Khoa	Sinh ngày	Giới tính	Nghề nghiệp	Dân tộc	Huyện
2	8/23/2016 15:40:40	Bệnh viện đa khoa tỉnh Sơn La			1930 Nữ		Kinh	
3	9/12/2016 16:36:39	Bệnh viện đa khoa huyện Mai Sơn	Khoa Nội Tổng Hợp		1949 Nữ	Nông Dân	Thái	Mai Sơn
4	9/12/2016 16:51:00	Bệnh viện đa khoa huyện Mai Sơn	Khoa Nội Tổng Hợp		1936 Nữ	Nghỉ Hưu	Kinh	Mai Sơn
5	9/13/2016 16:58:13	Bệnh viện đa khoa tỉnh Sơn La	HSTC		1960 Nữ	Nông Dân	Kinh	Mai Sơn
6	9/14/2016 9:15:53	Bệnh viện đa khoa huyện Mai Sơn	Khoa Nội Tổng Hợp		1965 Nam	Nông Dân	Thái	Mai Sơn
7	9/14/2016 9:54:27	Bệnh viện đa khoa huyện Yên Châu	Hồi Sức Cấp Cứu		1946 Nữ	Nông Dân	Kinh	Yên Châu
8	9/16/2016 14:56:30	Bệnh viện đa khoa Mộc Châu	Khoa Nội Tổng Hợp		1965 Nữ	Cán Bộ	Thái	Mộc Châu
9	9/19/2016 15:03:03	Bệnh viện đa khoa huyện Mai Sơn	Khoa Nội Tổng Hợp		1953 Nam	Nông Dân	Thái	Mai Sơn
10	9/19/2016 15:17:35	Bệnh viện đa khoa huyện Mai Sơn	Khoa Nội Tổng Hợp		1953 Nữ	Nông Dân	Thái	Mai Sơn
11	9/20/2016 14:54:32	Bệnh viện đa khoa Mộc Châu	Hồi Sức Cấp Cứu		1954 Nam	Nông Dân	Mường	Mộc Châu
12	9/20/2016 16:02:47	Bệnh viện đa khoa tỉnh Sơn La	HSTC		1942 Nam	Nghỉ Hưu	Sông Mã	
13	9/21/2016 10:07:13	Bệnh viện đa khoa tỉnh Sơn La	Hồi Sức Cấp Cứu		1976 Nữ	Nông Dân	Thái	
14	9/21/2016 10:17:14	Bệnh viện đa khoa tỉnh Sơn La	Khoa Tim Mạch		1976 Nam	Nghỉ Hưu	Kinh	
15	9/21/2016 10:44:46	Bệnh viện đa khoa tỉnh Sơn La	Khoa Nội Tổng Hợp		1955 Nữ	Nông Dân	Thái	Mường La
16	9/21/2016 10:53:27	Bệnh viện đa khoa tỉnh Sơn La	Khoa Nội Tổng Hợp		1937 Nam	Nghỉ Hưu	Kinh	

Fig. 3. Data is stored in excel file

The data of hypertension include a total of 2594 data streams, of which 1868 are labeled for diagnosis. As been classified by World Health Organization, there are 3 types of hypertension patients that we will use as classification labels for hypertension data sets [11]. We also have other labels but the proportion of the data is negligible, so we ignore them in the model because we are only interested in the hypertension patients who already went to the hospital in this paper. Labels are listed in Table 1:

Table 1. Classification labels for hypertension

Label	Quantity	Percentage (%)
Hypertension level III (Systolic $\geq$ 180 and Diastolic $\geq$ 110)	336	17.99
Hypertension level I (Systolic 140–159 and Diastolic 90–99)	700	37.47
Hypertension level II (Systolic 160–179 and Diastolic 100–109)	665	35.60

After having input data of hypertension, we filtered training data to test the diagnosis of this disease. Accordingly, we will build the data in SVMlight form, with each line corresponding to one line in the original data. The line will contain numbers representing the features of the data. For example, with the third feature of value 2, we will represent "3: 2". If the feature has a value of 0, we will not represent it. For features with numerical values we will take that numerical value as a value for the features. As for features with many textual values, we turn that feature into multiple columns corresponding to the value that the feature can receive. If any value is used, that column will be worth one.

For example, "Age" is the 10th feature and has a value of 24. Since "Age" is a number, we always convert the value to "10:24". But "Gender" has 2 values: "Male" and "Female" and being the 11th feature, we need to convert into 2 columns of 11 and 12. When we receive the value "Nam", we save it as "11: 1 12: 0", and when we receive the value "Nữ", we save it as "11: 0 12: 1". But since we will then remove columns with a value of 0, we can save it straight into "11: 1" or "12: 1".

Each training data line will have a classification label and with the SVMlight data structure, we need to put its label value at the beginning of the line. With the data set,

each label has its corresponding value. For example, if there are 6 label values, they will be numbered 1 to 6 respectively.

"Classification" is considered the data label and be converted into the first column. Because the "classification" column has 3 different values as shown in Table 1, it corresponds to numbers 1 to 3. The remaining columns will be numbered the same way as described above.

Since there are many columns with very little data or no value, we only build 2 sets of data based on columns with information greater than or equal to 50% and 70% to conduct the test.

Specifically, in Table 2, columns with the amount of data greater than or equal to 50% of hypertension are used as features of the training data set as follows:

Table 2. Statistics of columns with data amount of $\geq 50\%$

Name of column	Amount of data (%)	Type of data
Age	100.00	Natural number
Hospital	100.00	Text
Department	99.85	Text
Gender	100.00	Text
Occupation	86.40	Text
Ethnic:	94.95	Text
District	53.81	Text
City	82.63	Text
Province	100.00	Text
Do you have diabetes yourself, blood lipid disorder, coronary artery disease, kidney disease, or smoke?	56.93	Text
Age of high blood pressure	87.60	Natural number
Body weight	97.61	Real number
Body height	97.61	Real number
Temperature	97.50	Real number
Low blood pressure	98.15	Real number
High blood pressure	98.11	Real number
Breathing	95.45	Natural number
Blood glucose	90.83	Real number

(*continued*)

Table 2. (*continued*)

Name of column	Amount of data (%)	Type of data
Urea	91.64	Real number
Uric acid	56.32	Real number
Creatinine	91.14	Real number
Cholesterol	81.39	Text
Triglyceride	77.04	Text
HDL-C	52.54	Text
LDL-C	51.69	Text

Training data after being completed will look like following:

```
1 1:86 2:1 20:1 32:1 43:1 45:1 52:1 54:110 55:36.8 56:120 57:200 58:25 59:13.89 60:20.67 61:336.9 62:1 65:1
1 1:67 3:1 8:1 20:1 22:1 33:1 43:1 45:1 52:1 54:90 55:36.7 56:140 57:200 58:20 59:4.9 60:4.3 61:65 63:1 65:1
1 1:80 3:1 8:1 20:1 23:1 32:1 43:1 45:1 52:1 54:74 55:36.6 56:130 57:180 58:22 59:5.8 60:3.8 61:61 63:1 65:1
1 1:56 2:1 9:1 20:1 22:1 32:1 45:1 52:1 54:85 55:37 56:120 57:220 58:21 59:7.15 60:4.73 61:56.7 62:1 65:1
2 1:51 3:1 8:1 21:1 22:1 33:1 43:1 45:1 52:1 54:59 55:60 56:90 57:140 58:18 59:5.1 60:3.8 61:169
3 1:70 4:1 10:1 20:1 22:1 32:1 43:1 45:1 52:1 54:80 55:37 56:110 57:185 58:20 59:5.6 60:4.9 61:75 63:1
2 1:51 5:1 8:1 20:1 24:1 33:1 45:1 53:1 54:75 55:37 56:100 57:160 58:20 59:5.8 60:6.8 61:91 62:1 65:1
3 1:63 3:1 8:1 21:1 22:1 33:1 43:1 45:1 52:1 54:100 55:36.8 56:100 57:160 58:20 59:15.1 60:5.53 61:94 63:1 66:1
1 1:63 3:1 8:1 20:1 22:1 33:1 43:1 45:1 52:1 54:90 55:36 56:100 57:180 58:20 59:13.5 60:8.1 61:100 63:1 65:1
3 1:62 5:1 10:1 21:1 22:1 34:1 43:1 45:1 52:1 54:95 55:37 56:120 57:180 58:19 59:6.8 60:11.3 61:111
1 1:74 2:1 9:1 21:1 23:1 32:1 45:1 52:1 54:90 55:37 56:100 57:180 58:22 59:8.84 60:9.1 61:116.5 63:1 65:1
2 1:57 3:1 8:1 20:1 25:1 33:1 43:1 45:1 52:1 54:88 55:36.5 56:90 57:150 58:19 59:6.55 60:6.01 61:154 63:1 66:1
3 1:52 3:1 8:1 21:1 25:1 33:1 43:1 45:1 53:1 54:68 55:36.8 56:100 57:170 58:20 59:4.6 60:4.8 61:72 63:1 66:1
4 1:52 3:1 8:1 20:1 25:1 33:1 43:1 45:1 53:1 54:96 55:36.5 56:80 57:130 58:21 59:6.47 60:3.61 61:75 63:1 65:1
4 1:87 3:1 8:1 20:1 25:1 33:1 43:1 45:1 52:1 54:130 55:36.8 56:80 57:130 58:24 59:5.8 60:4 61:80 63:1 66:1
3 1:86 3:1 8:1 20:1 25:1 33:1 43:1 45:1 52:1 54:80 55:36.8 56:100 57:170 58:20 59:5.24 60:4.63 61:86 63:1 65:1
1 1:49 3:1 8:1 20:1 22:1 33:1 45:1 52:1 54:86 55:37 56:110 57:200 58:22 59:5.58 60:4.3 61:73 63:1 65:1
2 1:48 3:1 8:1 21:1 22:1 33:1 43:1 45:1 53:1 54:87 55:37 56:90 57:140 58:20 59:4.41 60:3.6 61:102 63:1 66:1
2 1:76 5:1 8:1 21:1 22:1 33:1 45:1 52:1 54:78 55:37 56:80 57:160 58:20 59:8.5 60:7.1 61:98 62:1 66:1
1 1:53 2:1 9:1 21:1 22:1 32:1 43:1 45:1 52:1 54:100 55:37 56:100 57:180 58:21 59:6.52 60:4.3 61:81 63:1 66:1
2 1:71 2:1 8:1 21:1 24:1 32:1 43:1 45:1 52:1 54:82 55:37 56:90 57:140 58:20 60:5.4
2 1:82 2:1 8:1 21:1 23:1 32:1 43:1 45:1 52:1 54:113 55:37 56:90 57:140 58:22 59:4.8 60:3.6 61:81 63:1 66:1
2 1:74 2:1 9:1 20:1 23:1 32:1 43:1 45:1 54:80 55:32 56:90 57:150 58:20 59:12.42 60:14.76 63:1 65:1
1 1:42 2:1 8:1 21:1 23:1 33:1 43:1 45:1 54:84 55:37 56:80 57:180 58:20 59:4.07 60:4.47 61:79 63:1 66:1
2 1:54 2:1 8:1 20:1 24:1 32:1 43:1 45:1 52:1 54:78 55:37 56:90 57:150 58:18 59:9.23 60:7.39 61:66.7 62:1 66:1
5 1:65 2:1 8:1 20:1 23:1 33:1 43:1 45:1 52:1 54:78 55:37 56:80 57:120 58:20 59:3.3 60:4.7 61:94 63:1 65:1
2 1:74 2:1 8:1 21:1 23:1 32:1 43:1 45:1 52:1 54:37 55:82 56:90 57:150 58:20 59:5.44 60:5.15 61:95.6 63:1 65:1
3 1:51 2:1 8:1 21:1 22:1 33:1 43:1 45:1 54:80 55:37 56:70 57:160 58:20 59:7.5 60:4.8 61:9.5 63:1 66:1
2 1:61 2:1 8:1 20:1 22:1 32:1 43:1 45:1 54:80 55:37 56:80 57:150 58:20 59:5.92 60:5.66 61:77.1 63:1 66:1
```

Training data structure

4 Test Data Set Developed Using Weka

We conducted the test using the data set developed, using four algorithms that are SVM, Naive Bayes, Decision Tree and KNN. The tool used to support the test is Weka, a machine learning software developed by Waikato University, New Zealand in Java. Weka is free software released under the GNU General Public License.

Weka [12] is a synthesis of machine learning algorithms for data mining. Algorithms can be used directly on data sets or can be called from Java code. Weka also includes data pre-processing, classification, regression, clustering, association rules and visualization tools.

In order to use Weka for the test, we follow these steps:

– Step 1: conduct data pre-processing:

We put in the training data set and select the data as shown in Fig. 5 (Fig. 4).

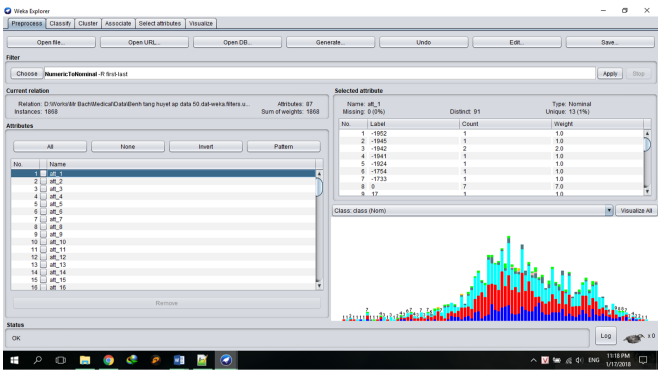


Fig. 4. Data pre-processing.

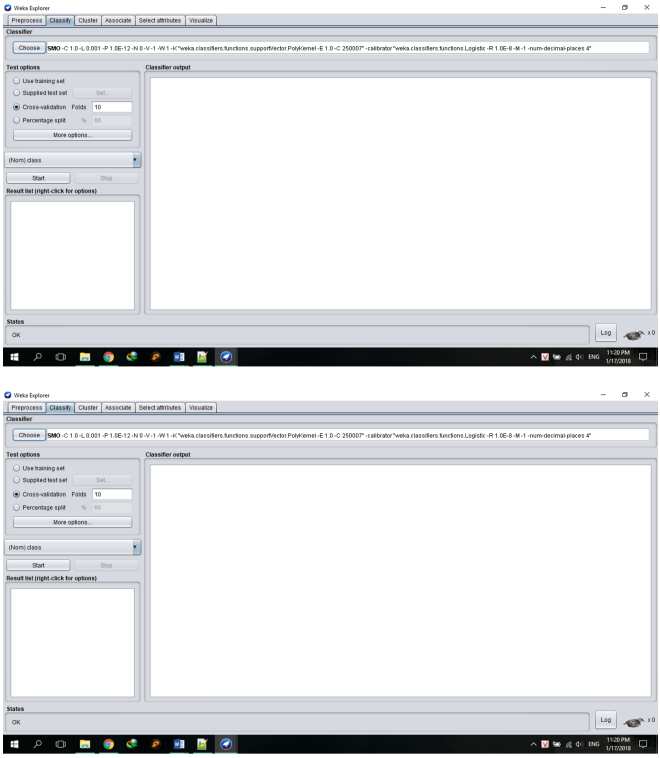


Fig. 5. Test configuration.

– Step 2: Configure the test settings:

Click on the Classify tab and set Cross-validation to 10 and other algorithms to conduct the test as shown in Fig. 6.

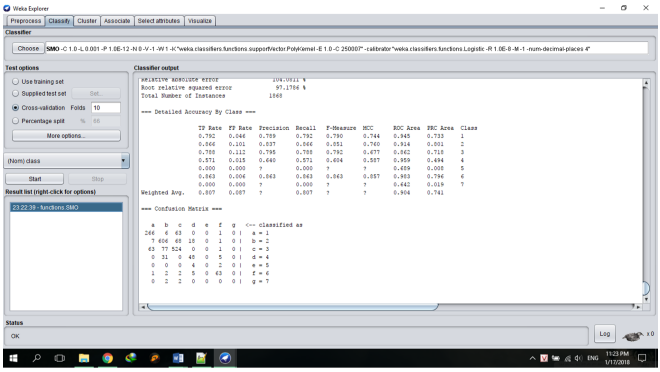


Fig. 6. Test configuration.

– Step 3: run the test

After configuration is complete, just click the Start button and the Weka tool will perform the test and the results will be displayed as shown in Fig. 7.

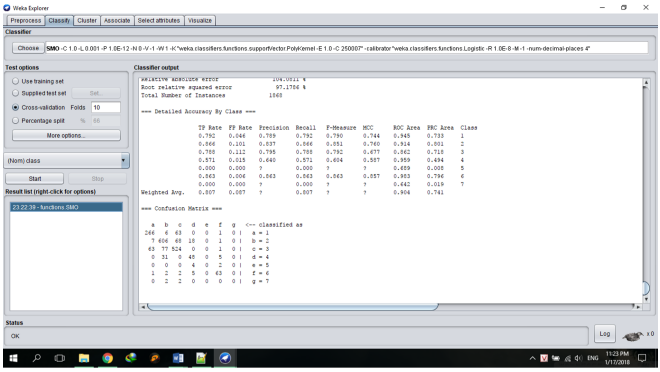


Fig. 7. Test running screen.

In running the test, we will use 2 models. One model uses features with data amount of $\geq 70\%$ and a model with data amount of $\geq 50\%$. Based on the test results of the two models, we assess the influence of the features on the classification of labels.

Test model 1 performed using features with data amount of $\geq 70\%$ and we have results in the following Table 3:

Table 3. Results of test model 1.

Algorithm	Accuracy (%)
SVM	80.19
Naïve Bayes	74.18
Decision Tree	82.49
KNN	65.36

Test model 2 performed using features with data amount of $\geq 50\%$ and we have results in the following Table 4:

Table 4. Results of test model 2.

Algorithm	Accuracy (%)
SVM	80.67
Naïve Bayes	66.54
Decision Tree	83.29
KNN	61.29

Some detailed results of predictive algorithms with hypertension data with Model 1 (Tables 5 and 6):

Table 5. Detailed results for each predictive label of Decision Tree algorithm for hypertension data with model 1

Label	Precision	Recall	F1
Hypertension level III (Systolic $\geq$ 180 and Diastolic $\geq$ 110)	0,855	0,842	0,849
Hypertension level I (Systolic 140–159 and Diastolic 90–99)	0,835	0,873	0,853
Hypertension level II (Systolic 160–179 and Diastolic 100–109)	0,819	0,823	0,821

Table 6. Detailed results for each predictive label of Decision Tree algorithm for hypertension data with model 2

Label	Precision	Recall	F1
Hypertension level III (Systolic $\geq$ 180 and Diastolic $\geq$ 110)	0,791	0,789	0,790
Hypertension level I (Systolic 140–159 and Diastolic 90–99)	0,836	0,851	0,844
Hypertension level II (Systolic 160–179 and Diastolic 100–109)	0,785	0,791	0,788

So, based on the results we can see that using both models, Decision Tree always gives the highest results with accuracy of 82.49% and 83.29% respectively for the data of hypertension in 2 models. And using more features gives higher results. SVM algorithm also gave quite good results, with accuracy of 80.19% in models 1 and 80.67 in model 2 on the same set of data of hypertension. The remaining two algorithms Naïve Bayes and KNN gave poor results, with KNN giving the worst results for both models, with an accuracy of 70% the highest.

5 Conclusion

In this paper, we has explored and analyze data on hypertension disease using machine learning methods with the test supporting tool of Weka. Based on the results of this test, we can more or less assess the ability to predict hypertension using the information obtained from patients, thereby helping doctors in the process of medical examination and treatment. In the future, we can continue to test using with more accurate and complete data sets in order to be able to improve the accuracy rate of disease prediction.

References

1. Machine Learning in Healthcare. https://emerj.com/ai-market-research/machine-learning-in-healthcare-executive-consensus/
2. Nguyen, L.L., Su, S.: Neural network approach for non-invasive detection of hyperglycemia using electrocardiographic signals (2014)
3. Lyon, A., Mincholé, A., Martínez, J.P., Laguna, P., Rodriguez, B.: Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. J. R. Soc. Interface **15**, 20170821 (2018)
4. Kachuee, M., Fazeli, S., Sarrafzadeh, M.: ECG heartbeat classification: a deep transferable representation (2018)
5. Causes of hypertension. https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/suc-khoe-tong-quat/cac-xet-nghiem-sinh-hoa-mau-chan-doan-benh-cao-huyet-ap/
6. Diagnose of hypertension. https://www.dieutri.vn/sinhhoalamsang/xet-nghiem-sinh-hoa-trong-tang-huyet-ap
7. Machine Learning for Medical Diagnostics. https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/
8. Zhu, Q.-Y., Qin, A.K., Suganthan, P.N.: Evolutionary extreme learning machine. Pattern Recogn. **38**(10), 1759–1763 (2005)
9. Gondra, I.: Applying machine learning to software fault-proneness prediction. J. Syst. Softw. **81**(2), 186–195 (2008)
10. Thanh Tung, N.: Application and deployment of software system for integrating and connecting biomedical electronic devices and communication networks to support the monitoring of health and epidemiology in the Northwest region. National research project funded by Ministry of Science and Technology
11. WHO Information about hypertension. https://www.who.int/health-topics/hypertension/#tab=tab_1
12. Weka. https://www.cs.waikato.ac.nz/ml/weka/