A Predictive Model for Heart Disease Detection Using Data Mining Techniques

Jakkrit Premsmith and Hathairat Ketmaneechairat
College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand
Email: {jakkrit.p, hathairat.k}@cit.kmutnb.ac.th

***Abstract*—In this paper, the model is proposed to predict the heart disease detection by using data mining techniques. The data mining algorithm uses the Logistic Regression model and Neural Network model. The dataset of this paper uses the heart disease data at the University of California Irvine (UCI). There are a total of 303 Instances and 75 Attributes in the United States. The evaluation criteria using the confusion matrix table such as accuracy, precision, recall and F-Measure. The results show that the Logistic Regression model is better performance than Neural Network model. The Logistic Regression model has 95.45% precision and 91.65% accuracy. The web application can be support for the user, who wants to diagnose heart disease detection.**

***Index Terms*—heart disease, predictive model, detection, data mining, logistic regression, neural network**

I. Introduction

Heart disease is an affects from the work of the heart. There are a lot of heart disease caused by risk factors in lifestyle habits such as age, sex, smoking or inhaling cigarette smoke, family history, cholesterol, obesity or eating foods that are high in fat, poor diet, blood sugar levels, high blood pressure, physical inactivity, alcohol and body weight. Some risk factors are controllable [1]. The heart disease can be divided into seven types, coronary heart disease, arrhythmia, congestive heart failure, congenital heart disease, cardiomyopathy, angina pectoris and myocarditis [2]. The World Health Organization reports that the heart disease is the number one risk of death in the world, accounting for 31% of the worldwide mortality rate. In Thailand, which the statistic of the Ministry of Public Health in September 2018, there are 432,943 heart disease patients, 20,855 person deaths, equivalent to 2 hours of death per person [3]. The risk of heart disease is divided into several levels. A period at the initial risk level treatment will use lower costs and increase the possibility of saving lives of patients. The heart disease is the biggest cause of death nowadays. The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the mentioned factors are taken into consideration when analyzing and understanding the patients by the doctor through manual check-ups at regular intervals of time. Various data mining techniques formulate due to different research works. These data mining techniques are straightforwardly utilized for developing frameworks or to find crucial inferences and conclusions from the resulted dataset. Various well-known techniques are used to predict the risk of heart disease such as Support Vector Machine (SVM), K-Means, Naïve Bayes, Neural Network, K-Nearest Neighbor (KNN), Decision Trees, Random Forests and Logistic Regression [4].

This research aims to identify the significant features and data mining techniques to predict the heart disease risk. The experiment is conducted to identify the features and data mining techniques. The heart disease datasets are collected from the data source, UCI Machine Learning Repository. Cleveland dataset is selected because it is a commonly used database by machine learning researchers with records that are most complete. Logistic Regression and Neural Network are applied to create prediction models for this experiment using the prepared dataset. The dataset has been divided into a training dataset and a testing dataset. The efficiency of the classifiers is tested with the testing dataset. Additionally, this research also compares the highest accuracy achieved by the best technique identified from this research against the highest accuracy achieved in the existing studies. This research implement sigmoid function in both Neural Network and Logistic Regression due to comparison and obtaining the model which result in best accuracy and less complex for real world application on the web application. The web application is created to predict heart disease detection for user. The remainder of this paper is organized as follows. Section II describes related work. Section III explains the research methodology. The experimental result is presented in Section IV. Finally, the conclusion discussed in the Section V.

II. Related Work

The many researchers have contributed for the development of prediction of heart disease risk. The predication of heart disease risk is based on data mining techniques. Recently, researchers have a lot of papers and research material on this heart disease. In this section, the state of the art work by different authors and researchers are presented. The proposed model and related works are compared as shown in Table I.

Ramalingam, *et al*. [5] presents a survey of various models based on such algorithms and techniques and analyze their performance. Models based on supervised

Manuscript received June 16, 2020; revised November 18, 2020.

doi: 10.12720/jait.12.1.14-20

learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers.

Sharma, *et al*. [6] describes for wide scope survey in the field of machine learning technique in heart disease. The machine learning algorithm and deep learning opens new door opportunities for precise predication of heart attack. This paper provides slot information about state of art methods in machine learning and deep learning. An analytical comparison has been provided to help new researches' working in this field. An analytical comparison has been done for finding out best available algorithm for medical dataset.

David, *et al*. [7] have proposed heart disease prediction using data mining techniques. A comparative analysis of the three data mining classification algorithms namely Random Forest, Decision Tree and Naïve Bayes are used to develop a prediction system in order to analyze and predict the possibility of heart disease. The experimental setup has been made for the evaluation of the performance of algorithms with the help of heart disease benchmark dataset retrieved from UCI machine learning repository. It is found that Random Forest algorithm performs best with 81% precision when compared to other algorithms for heart disease prediction.

Singh, *et al*. [8] have performed an Effective Heart Disease Prediction System (EHDPS) is developed using neural network for predicting the risk level of heart disease. From ANN, an MLPNN together with BP algorithm is used to develop the system. The MLPNN model proves the better results and assists the domain experts and even the person related to the medical field to plan for a better and early diagnosis for the patient. This system performs realistically well even without retraining. Furthermore, the experimental results show that the system predicts heart disease with ~ 100% accuracy by using neural networks.

Subhadra, *et al*. [9] have proposed a diagnostic system for predicting heart disease using Multilayer Perceptron Neural Network. For diagnosis of heart disease 14 significant attributes are used in proposed system as per the medical literature. For effective prediction, back propagation algorithm was applied to train the data and compare the parameters iteratively. The results tabulated evidently prove that the designed diagnostic system is capable of predicting the risk level of heart disease effectively when compared to other approaches.

Ami, *et al*. [10] is focused on the identify significant features and data mining techniques that can improve the accuracy of predicting cardiovascular disease. The heart disease datasets were collected from the data source, UCI Machine Learning Repository. Cleveland dataset was selected because it is a commonly used database by machine learning researchers with records that are most complete. The prediction models were developed using different combination of features, and seven classification techniques: k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network and Vote (a hybrid technique with Naïve Bayes and Logistic Regression). The experiment results show that the heart disease prediction model developed using the identified significant features and the best-performing data mining technique (i.e.Vote) achieves an accuracy of 87.4% in heart disease prediction.

TABLE I. COMPARISON OF THE PROPOSED MODEL WITH RELATED WORK

Paper No.	Technique	Confusion Matrix	No. of Attributes	Data Set
[1]	Naïve Bayes, Random Forest, Bayes Net, C4.5, Multilayer Perceptron and PART Algorithms.	Accuracy	14	Cleveland Dataset
[7]	Random Forest, Decision trees and Naive Bayes	Precision, Recall, F-Measure, MCC, ROC Area, and PRC Area.	13	StatLog Dataset
[8]	Multilayer Perceptron Neural Network (MLPNN) with Backpropagation (BP)	TP, FP, FN, TN Precision, ROC Area, F-Measure.	15	Standard Dataset
[9]	Decision tree, Logistic Regression, Naïve Bayes, Random forests, Support Vector Machines, Generalized Liner Model, Gradient Boosted Trees, Deep Learning and MLPNN Models	Sensitivity, Specification, Precision and Accuracy.	14	UCI
[10]	k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network and Vote (a Hybrid Technique with Naïve Bayes and Logistic Regression)	Accuracy, F-Measure and Precision.	14	Cleveland Dataset
Proposed Model	Logistic Regression, Neural Network	Accuracy, Precision, Recall, and F-Measure.	14	Cleveland Dataset

III. RESEARCH METHODOLOGY

In this section, the methodology for prediction of heart disease detection using data mining techniques is explained. The research methodology is divided into seven parts such as dataset, data mining tool, data preprocessing, feature selection, create predictive model, evaluation criteria and deployment methods.

A. Dataset

The heart disease dataset used in this research is the Cleveland Heart Disease dataset taken from the UCI machine learning repository [11]. UCI heart disease dataset consists of four separate databases collected from four various medical hospitals. The dataset consists of 303 records and 75 attributes. There are 14 attributes in the dataset, which are described as following. Table II describes the description and type of attributes. There are 13 attributes that feature in heart disease prediction and one attribute serves as the output or the predicted attribute for the presence of heart disease in a patient.

TABLE II. HEART DISEASE DATASET

Attribute Name	Attribute Description	Type
Age	Age of the patient in years.	Numeric
Sex	Gender of the patient. Represented as a binary number. 1 = Male, 0 = Female	Nominal
Cp	Chest pain type. Values range from 1 to 4. Value 1 = typical angina, Value 2 = atypical angina, Value 3 = non-angina pain, Value 4 = asymptomatic	Nominal
trestbps	Resting blood pressure measured in mm/Hg on admission to the hospital.	Numeric
Chol	Serum cholesterol of the patient measured in mg/dl.	Numeric
Fbs	Fasting blood sugar of the patient. If greater than 120 mg/dl the attribute value is 1 (true), else the attribute value is 0 (false). Value 1 = true, Value 0 = false	Nominal
Restecg	Resting electrocardiographic results for the patient. This attribute can take 3 integer values 0, 1, or 2. Value 0 = normal, Value 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2 = showing probable or definite left ventricular hypertrophy by Ester's criteria.	Nominal
Thalach	Maximum heart rate achieved of the patient.	Numeric
Exang	Exercise induced angina. Values can be 0 or 1. Value 1 = yes, Value 0 = no	Nominal
Oldpeak	ST depression induced by exercise relative to set.	Numeric
Slope	Measure of slope for peak exercise. Value can be 1, 2, or 3. Value 1 = up sloping, Value 2 = flat, Value 3 = down sloping	Nominal
CA	Number of major vessels (0-3) colored by fluoroscopy. Attribute values can be 0 to 3.	Numeric
Thal	Represents heart rate of the patient. It can take values 3, 6, or 7. Value 3 = normal, Value 6 = fixed defect, Value 7 = reversible defect	Nominal
Target	Represents the diagnosis of heart disease which have 2 nominal values Value 0: diagnosis of non-heart disease (false) Value 1: diagnosis of heart disease (true)	Nominal

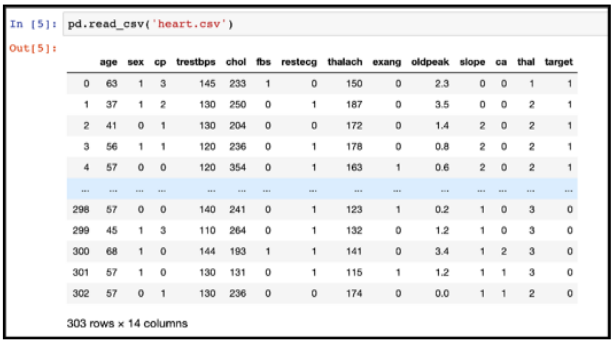
B. Data Mining Tool

The tools provide ready applications to be used for data mining algorithms. The tools have an easy to use interface and researchers can easily use data mining tools because of free open-source software. The use of Python in the area of data science has reached unprecedented levels, especially in the area of freely available tools and libraries [12]. The popular programming languages for data mining include Python, R, and MATLAB. This research implemented Python Programming with scientific machine learning, Scikit-Learn for all processes. Python is utilized to conduct experiments because of powerful computation, compatibility and various library. There are numerous libraries for all state of data mining such as Pandas for import data, Scikit-Learn for features selection, model building and evaluation, Matplotlib for visualization and Jupyter Notebook for workspace. The visual representation of the workflow is one of the efficient features for beginners. In the experiment, the UCI Cleveland heart disease dataset is imported into Python Jupyter Notebook. The main step of data mining process composes of five phases, the data preprocessing phase, the feature selection phase, and create classification modeling phase, evaluation phase and deployment phase. The deployment is used Python programming and ML Studio which is Azure Cloud service for deploy model as Application Programming Interface (API).

C. Data Preprocessing

After the UCI Cleveland heart disease dataset is imported into Python Jupyter Notebook as shown in Fig. 1. The data preprocessing will be start. There are six records that have missing values in the UCI Cleveland dataset. All the records with missing values are removed

from the dataset, thus reducing the number of records from 303 to 297. The values of predicted attribute for the presence of heart disease in the data set is transformed from multiclass values (0 for absence and 1, 2, 3, 4 for presence) to the binary values (0 for absence; 1 for presence of heart disease). The data preprocessing task is performed by converting all the diagnosis values from 2 to 4 into 1. The resulting data set contains only 0 and 1 as the diagnosis value, where 0 is the absence and 1 is the presence of heart disease. After the reduction and transformation, the number of record is 297 records, there are 160 records as '0' and 137 records as '1' [9], [10].



```
In [5]: pd.read_csv('heart.csv')
Out[5]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Figure 1. Import dataset into Python Jupyter Notebook.

The missing value is handled by utilized basic principle and algorithms to each specific data type. For example, the age attribute is the numerical data type, it is proceeded by mean as shown in equation (1) [13].

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

On the other type, nominal data type is unable to apply mean because of bias occurrence. Instead, the utilization of distance base algorithm, K-Nearest Neighbor is capable of handling nominal data type as shown in equation (2) [13].

$$d(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_i-q_i)^2 + \cdots + (p_n-q_n)^2} \tag{2}$$

The algorithm considers other factors in row then apply Euclidean Distance and KNN to predict potential result of missing value as shown in equation (3) [13].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

D. Feature Selection

There are 14 attributes in the dataset, 13 attributes that feature in heart disease prediction and 1 attribute serves as a class label or the output or the predicted attribute for the presence of heart disease in a patient. For the "age" and "sex" attributes refer to the personal information of each patient. The remaining 11 attributes are clinical attributes. The select attributes are selected from independent variable, which has more than 30% relationship with control variable. The eight attributes are selected to use in this research. The feature selection is conduct in the coding formatting of Python Jupyter Notebook as shown in Fig. 2.

```
1  correlation_feature = outlier_elimi . corr() # declare
2  variable "correlation_feautres" to stored correlation of
3  each features.
4  cor_target = abs(outlier_elimi.corr()['target']) # declare
5  "cor_target" to stored target or dependent variable
6  relavant_features = cor_target[cor_target > 0.3] #
   proceeding features selection by eliminating features less
   30% correlate to dependent variable
```

Figure 2. Feature selection in Python programming.

TABLE III. FEATURE SELECTION

Attribute Name	Attribute Correlation (> 30%)
Cp	(40%)
Thalach	(43%)
Exang	(42%)
Oldpeak	(42%)
Thal	(35%)
Slope	(32%)
Ca	(46%)
Sex	(30%)

According to features analysis, this work applies Correlation Metrix to screen features which has correlated to target more than 30% resulting as shown in Table III. The eight features (except for target) is remaining; Chest pain, Maximum heart rate, Exercise induced angina, ST depression, Represents heart rate of the patient, Measure of slope for peak exercise, Sex and Number of major vessels. The observation points out all of highly correlating factors are only internal body and coming from behavior of patients significantly.

E. Predictive Model

After feature selection, the Logistic Regression model and Neural Network models are created. The 10-folds cross validation technique is used to validate the performance of the models. The dataset is divided into 10 subsets and then processed 10-times. There are three subsets for testing data and seven subsets for training data.

1) Logistic Regression [14]: Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable (a dependent variable that can take a limited number of values) from a set of predictor or independent variables. In logistic regression the dependent variable is always binary (with two categories). Logistic regression is mainly used to for prediction and also calculating the probability of success. The logistic regression can be calculated using the equation (4). The logistic regression model is created in Python as shown in Fig. 3 [15].

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{4}$$

2) *Neural Network [16]:* Neural Network (NN) is a parallel, distributed information processing structure consisting of multiple numbers of processing elements called node, they are interconnected via unidirectional signal channels called connections. Each processing element has a single output connection that branches into many connections; each carries the same signal i.e. the processing element output signal. The NN can be classified in two main groups according to the way they learn. supervised learning and unsupervised learning. The

neural network can be calculated using the equation (5) [15]. The neural network model is created in Python as shown in Fig. 4.

$$z = \sum_i w_i x_i \quad (5)$$

The Feed-Forward Neural Network called Multilayer perception is trained and weight adjustment by back-propagation algorithm. It learns how to transform input data into a desired response, so they are widely used for pattern classification. With one or more hidden layers, Network can map input-output for optimal result in difficult classification problem. The multilayer is trained with error correction learning. The error correction learning works in the following way from the system response at neuron j at iteration t, $y_j(t)$, and the desired response $d_j(t)$ for given input pattern an instantaneous error $e_j(t)$ is defined by equation (6) [15].

$$e_j(t) = d_j(t) - y_j(t) \quad (6)$$

```
1  from sklearn.linear_model import LogisticRegression
2  logisticReg = LogisticRegression(
3                          solver='lbfgs'
4                          random_state=5)
5  logisticReg.fit(X_train, y_train)
6  logistic_y_pred = logisticReg.predict(x_test)
```

Figure 3. Logistic regression model.

```
1  from sklearn.neural_network import MLPClassifier
2  neuralNet = MPLClassifier(solver='lbfgs',
3                          hidden_layer_sizes=(8,8),
4                          random_state=5,
5                          activation='logistic')
6  neuralNet.fit(X_train, y_train)
7  neural_y_pred = neuralNet.predict(x_test)
```

Figure 4. Neural network model.

F. Evaluation Criteria

The performance of Logistic Regression model and Neural Network model are often evaluated using the confusion matrix such as accuracy, precision, recall and F-Measure [7]-[10]. Thus, in order to evaluate the diagnostic performances of Logistic Regression model and Neural Network model presented in this research which are described in the Table IV, will be used based on both the training and testing datasets.

TABLE IV. EVALUATION METHODS AND EQUATION FOR PERFORMANCE OF THE MODEL

Evaluation Methods	Equations
Accuracy	(TP + TN) / (TP + FN + FP + TN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F-Measure	2 * [(Precision * Recall) / (Precision + Recall)]

where TP = True Positive, FP = False Positive, TN = True Negative and FN = False Negative, Respectively.

G. Deployment Methods

The process of deployment methods composed of flow-design and API integrations with web application. This research was used Azure ML for deployment. Azure ML addresses more than 100 techniques such as regression, anomaly detection, binary and multiclass classification, text analysis, etc., however, it allows as well the use of python and R languages for user models customization [17]. The utilization of ML Studio on Azure Cloud is implemented by relying on data mining flow, import data, data preprocessing, model building, logistic regression which is the best accuracy model. The ML Studio on Azure Cloud is shown in Fig. 5. Another part is building Application Programming Interface (API) for the calling model to manipulate real data. We developed web application to interact with the user and display the probability of diagnose heart disease detection. The deployment on cloud computing in the real world has limitation in term of machine performance and business logic. The model modification has been adopted to suit of deployment by reducing some model parameters (result in lesser accuracy but being suitable for the real world usage).

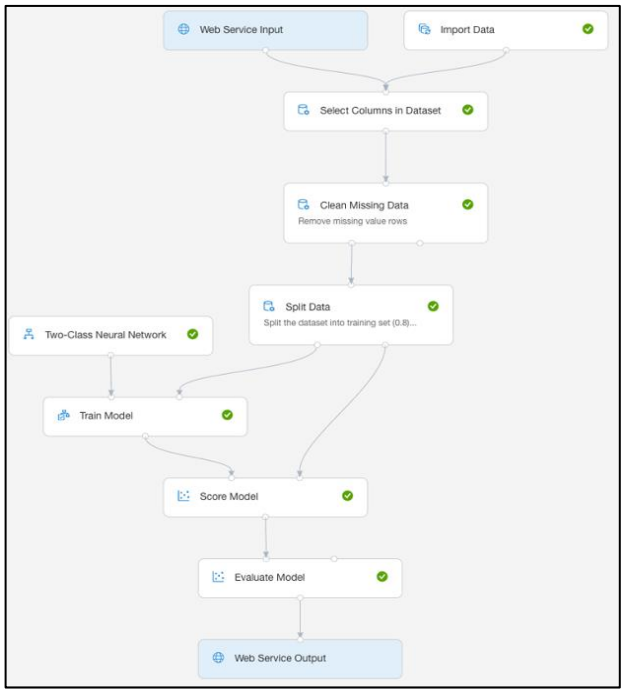


Figure 5. Data mining flow of ML studio on azure cloud.

IV. EXPERIMENTAL RESULT

This section presents the results achieved in the experiments. The attributes feature in heart disease prediction and performance data mining techniques, logistic regression and neural network is identified based on the analysis of the experimental results. The performance analysis is shown in Table V. The logistic regression has accuracy 91.65%, precision 95.45%, recall 84% and F-Measure 89.36%. The neural network has accuracy 89.65% precision 82.61%, recall 90.47% and F-Measure 86.36%. Therefore, based on the testing results, the logistic regression has the high value of performance evaluation demonstrate that the development of logistic regression model is able to predict a high accuracy in heart disease diagnosis in the patients. The web application for heart disease detection is shown in Fig. 6.

TABLE V. PERFORMANCE ANALYSIS

Evaluation	Logistic Regression	Neural Network
Accuracy	91.65%	89.65%
Precision	95.45%	82.61%
Recall	84.00%	90.47%
F-Measure	89.36%	86.36%

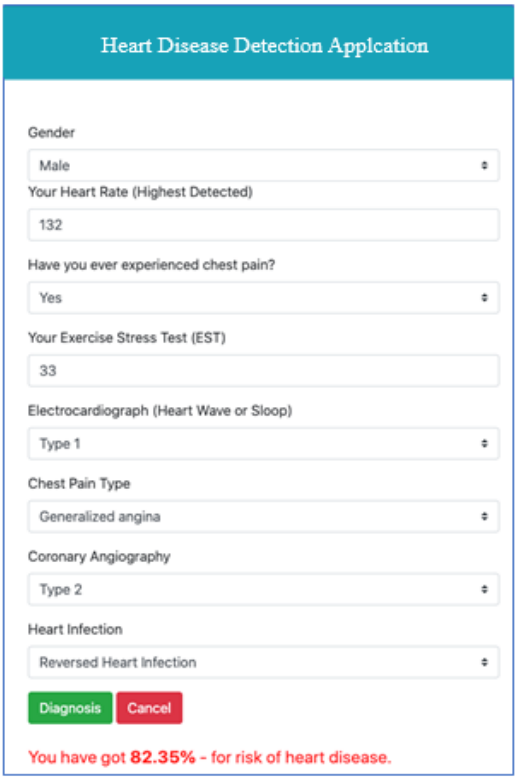


Figure 6. Web application for heart disease detection.

V. CONCLUSION

In this paper, the logistic regression model and neural network model are developed and evaluated based on diagnostic performance of heart disease in patients using accuracy, precision, recall and F-Measure. The heart disease data are from 303 patients and 75 Attributes at the Cleveland Clinic Foundation (CCF) located in Cleveland, Ohio in the United States. The dataset is obtained from the Heart Disease Database made available in the UCI Machine Learning Repository. The dataset is divided into a training dataset and a testing dataset. The Python Programming is used to implement the experiment. The experimental result shows that using logistic regression model, the system predicts heart disease is better performance more than neural network. The web application can predict heart disease detection for user. In the future research, the researcher will be conducted to test different combination of data mining techniques in heart disease prediction by using python. Additionally, new feature selection methods can be applied to improve the accuracy in prediction. The dataset will be collected from the patients in Thailand hospital.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Jakkrit Premsmith conducted the research; Hathairat Ketmaneechairat analyzed the data, implement and development web application. All authors wrote the paper and had approved the final version.

ACKNOWLEDGMENT

This research is supported by the College of Industrial Technology Faculty (CIT) of the King Mongkut's University of Technology North Bangkok (KMUTNB).

REFERENCES

[1] B. C. Latha and C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Journal of Informatics in Medicine Unlocked*, vol. 16, pp. 1-9, 2019.
[2] S. Kumar, "A survey on data mining techniques for prediction of heart diseases," *IOSR Journal of Engineering*, pp. 22-27, 2018.
[3] World Health Organization and Ministry of Public Health. [Online]. Available: https://www.naewna.com/lady/446144
[4] A. Jain, M. Ahirwar, and R. Pandey, "A review on intutive prediction of heart disease using data mining techniques," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 7, pp. 109-113, July 2019.
[5] V. V. Ramalingam, A. Dandapath, and K. Raja, "Heart disease prediction using machine learning techniques: A survey," *International Journal of Engineering & Technology*, vol. 7, pp. 684-687, 2018.
[6] H. Sharma and M. A. Rizvi, "Prediction of heart disease using machine learning algorithms: A survey," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 8, pp. 99-104, 2017.
[7] B. David and A. Belcy, "Heart disease prediction using data mining techniques," *Journal on Soft Computing*, vol. 9, no. 1, pp. 1824-1830, October 2018.
[8] P. Singh, S. Singh and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques," *International Journal of Nanomedicine*, pp. 121-124, 2018.
[9] K. Subhadra and B. Vikas, "Neural network based intelligent system for predicting heart disease," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 5, pp. 484-487, March 2019.
[10] M. Amin, Y. K. Chiam, and K. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Journal of Telematics and Informatics*, vol. 36, pp. 82-93, 2019.
[11] UCI Machine Learning Repository. Heart disease data set. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease
[12] I. Stančin and A. Jović, "An overview and comparison of free Python libraries for data mining and big data analysis," in *Proc. 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics*, May 2019, pp. 997-998.
[13] J. Han, M. Kamber, and J. Pei, *Data Mining*: *Concepts and Techniques*, third edition, Elsevier, 2012.
[14] T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, "Heart disease prediction model using SVM-Decision Trees-Logistic regression (SDL)," *International Journal of Computer Applications*, vol. 68, no. 16, pp. 11-15, April 2013.
[15] I. H. Witten and E. Frank, *Data Mining Practical Machine Learning Tools and Techniques*, second edition, Elsevier, 2005.
[16] C. Dangare and S. Apte, "A data mining approach for prediction of disease using neural network," *International Journal of Computer Engineering and Technology*, vol. 3, no. 3, pp. 30-40, October-December 2012.
[17] M. Prist, *et al.*, "Machine learning-as-a-service for consumer electronics fault diagnosis: A comparison between Matlab and Azure ML," in *Proc. IEEE International Conference on Consumer Electronics*, January 2020.

Jakkrit Premsmith received PhD in Information and Communication Technology for Education with the thesis title "Challenge-Based Learning Management System on Ubiquitous Cloud Environment to Enhance Real-World Problem-Solving Skills for Undergraduate Students" from King Mongkut's University of Technology North Bangkok **(**KMUTNB**)** Bangkok, Thailand**.** Currently, he is lecturer at the Department of Information and Production Technology Management **(**IPTM**)**, College of Industrial Technology at King Mongkut's University of Technology North Bangkok. He works in the field of Information and Production Technology Management**.**

Hathairat Ketmaneechairat received PhD in Electrical Engineering with the thesis title "Smart Buffer Management for Different Start Video Broadcasting" from the King Mongkut's University of Technology North Bangkok, Thailand. Currently, she is a lecturer at the College of Industrial Technology, King Mongkut's University of Technology North Bangkok. Her research areas are Natural Language Processing and Data Mining, Machine Learning and Artificial Intelligence.