ICIC Express Letters
Part B: Applications
Volume **14**, Number **8**, August **2023**

ICIC International ©2023 ISSN 2185-2766
pp. 779–786

ELECTROCARDIOGRAM ABNORMAL DETECTION MODEL USING MACHINE LEARNING APPROACH

Ben Rahman[1,*], Boy Subirosa Sabarguna[2]
Harco Leslie Hendric Spits Warnars[3] and Widodo Budiharto[4]

[1]Faculty of Information and Communication Technology
University of Nasional
Jl. Sawo Manila #61, Pejaten, Pasar Minggu, Jakarta 12520, Indonesia
[*]Corresponding author: benrahman@civitas.unas.ac.id

[2]Faculty of Computer Science
University of Indonesia
Pondok Cina, Kecamatan Beji, Kota Depok, Jawa Barat 16424, Indonesia
sabarguna08@ui.ac.id

[3]Computer Science Department, Graduate Program
[4]Computer Science Department
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ shendric; wbudiharto }@binus.edu

Received December 2022; accepted February 2023

Abstract. *Today, datasets can be obtained transparently and freely. It can also be used to categorize and predict diseases with high-risk factors. Moreover, the extracted datasets can generate important information for the entire population if handled accurately. This dataset can predict heart disease using a machine learning approach with explicit calculations. We compared the prediction of abnormal electrocardiograms in this study using machine learning with three algorithms, namely support vector machine (SVM), k-Nearest Neighbors (KNN), and multilayer perceptron (MLP) classifier. We used 14 attributes: (1) age, (2) systolic, (3) heart rate, (4) obesity, (5) smoking, (6) alcohol, (7) exercise, (8) treadmill exercise results, (9) total cholesterol, (10) high-density lipoprotein, (11) low-density lipoprotein, (12) creatinine, (13) serum glutamic oxaloacetic transaminase, and (14) urine protein. The results predict the indicated heart disease and display the accuracy of each algorithm. Furthermore, the results revealed that the machine learning technique employing the KNN algorithm is the most effective, with an accuracy rate of 89.375%.*

Keywords: Electrocardiogram, Prediction electrocardiogram, Machine learning approach

1. **Introduction.** An electrocardiogram (ECG) is a tool that detects and records the heart's electrical activity. An abnormal ECG shows an abnormality in the pattern of electrical activity [1]. Machine learning is the process of finding algorithms that improve the experience and capabilities of the system and are derived automatically from data [2]. Data mining is the process of extracting information from the data itself and is used to find new patterns and generate knowledge. Data can provide valuable information to organizations and individuals [3]. This study employs a combination of data mining and machine learning techniques.

Information mining is utilized to uncover data and recognize information from a dataset. A knowledge discovery database (KDB) is another name for information mining [4]. The

DOI: 10.24507/icicelb.14.08.779

four procedures in information mining are classification, clustering, regression, and association [5,6]. These information mining rules and procedures can rapidly remove a large amount of information from datasets.

This research is different from earlier researches in that the data processing process was carried out utilizing a "scoring function" for each operation [7]. The scoring function aims to assess which attribute has the most influences on the target. The results of the scoring function showed that 11 of the 14 features produced high scores, indicating that they had an effect on the target. This is a new finding compared to previous research [8]. Figure 1 shows a manually performed electrocardiogram.

(a) (b)

FIGURE 1. Examples of normal ECG (a) and ECG device (b)

The main goal of this study is to predict the likelihood of heart disease based on certain risk factors [9,10]. For classification, the support vector machine (SVM) algorithm [11], KNN [2], and MLP [12] were all used. These three models were used to determine the accuracy of the classification techniques. The researchers used datasets taken from one hospital to identify heart disease. With the KNN algorithm, the accuracy of the prediction of heart disease in this study was 89.375%. Machine learning techniques such as SVM, KNN, and MLP algorithms were utilized to predict heart disease [1,13]. These methods were utilized in this study to predict whether a person suffers from heart disease. When using a machine learning classification technique with a scoring function, the results are accurate [7]. The remainder of this paper includes the following sections: part 2 is the literature review, part 3 is the proposed method, part 4 is the result and discussion, and part 5 shows the conclusions.

2. **Literature Review.** The dataset contains several rows of data, i.e., 1284 records and 14 attributes. Some attribute values were missing [11]. The missing values were replaced with values according to the mean-mode strategy. This interaction is also known as information pre-processing [14]. The SVM, KNN, and MLP classification algorithms were applied following data pre-processing. A confusion matrix [13] was developed to calculate the accuracy level of classification. Two classes are displayed in a confusion matrix: Class X is TRUE (abnormal/positive) and Class Y is FALSE (normal/negative) as shown in Figure 2.

We propose combining data mining and machine learning techniques to reduce the mortality rate. This proposed system will predict whether a person has heart disease. SVM, KNN, and MLP are the three data mining classification algorithms employed to show the accuracy of heart disease predictions. There were four stages conducted: first, fetch data; second, pre-processing data; third, architecture and model proposed; and four, algorithm tested.

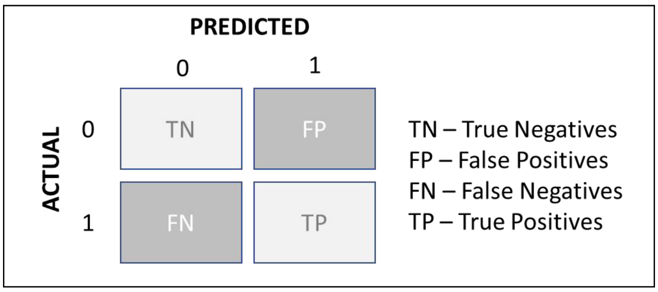

Figure 2. Confusion matrix – Heart disease classification

2.1. **Grabbing data.** The information was collected from patients and stored in a database. The dataset contained 1284 sample data with 14 medical parameters/attributes: (1) age, (2) systolic, (3) heart rate, (4) obesity, (5) smoking, (6) alcohol, (7) exercise, (8) treadmill exercise results, (9) total cholesterol, (10) high-density lipoprotein (HDL), (11) low-density lipoprotein (LDL), (12) creatinine, (13) serum glutamic oxaloacetic transaminase (SGOT), and (14) urine protein. The study's dataset already included attribute datasets for heart disease, including information about the level of blocked pressure, types of chest pain, and electrocardiography result.

2.2. **Pre-processing data.** A few data points were absent from the dataset. These were replaced according to the standard mode strategy. This process is referred to as the pre-handling of information and includes the extraction and selection of a subset of a few fields or related summaries.

Extraction and future choice are also called variable determination and trait choice.

Table 1 contains datasets taken from a normalized heart disease dataset.

Table 1. Datasets of heart disease

Age	Systolic	Heart rate	Obesity	Smoking	Alcohol	Exercise	Treadmill	Total cholesterol	HDL	LDL	Creatinine	SGOT	Urine protein	ECG
50	110	80	1961	150	0	0	0	22500	3600	15600	90	2200	0	0
48	110	64	2388	0	0	0	0	19000	5830	10390	100	2000	0	0
42	120	68	2565	0	0	1	0	20283	5250	13462	122	2668	0	0
34	120	65	3093	0	0	0	0	22900	4000	16700	110	2400	0	0
45	120	60	2160	0	0	0	0	20600	4500	13800	97	2670	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
52	120	80	2392	80	0	1	1	28900	4500	20280	90	4000	0	1
52	130	72	3683	0	56	1	0	25614	4064	16395	98	1974	0	1
49	110	68	2394	396	0	0	0	19937	5185	11048	107	2432	0	1
34	120	60	2180	420	40	1	0	15000	4400	8740	90	2200	0	1
32	120	84	2633	0	0	0	0	20400	4400	13500	100	3400	0	1

3. Proposed Method.

3.1. **System architecture and model proposed.** The features required to detect an abnormal ECG are shown in the model in Figure 3. The identification of the parts was carried out as follows: a) conducting a literature review, b) analyzing the elements of an abnormal ECG based on certain risk factors which are needed in the construction, c) collecting data and normalizing data (data mining), d) formulating the features needed in an abnormal ECG and the required models to classify the data (machine learning), and e) scoring the results of the data using the Python library.

Follow these steps with several scenarios, starting with all attributes (n) and use each of the SVM, KNN, and MLP algorithms. Perform up to $n-1$ up to the value of the feature that is affected (n is the total attributes). Obtain abnormal ECG features from the three algorithms.

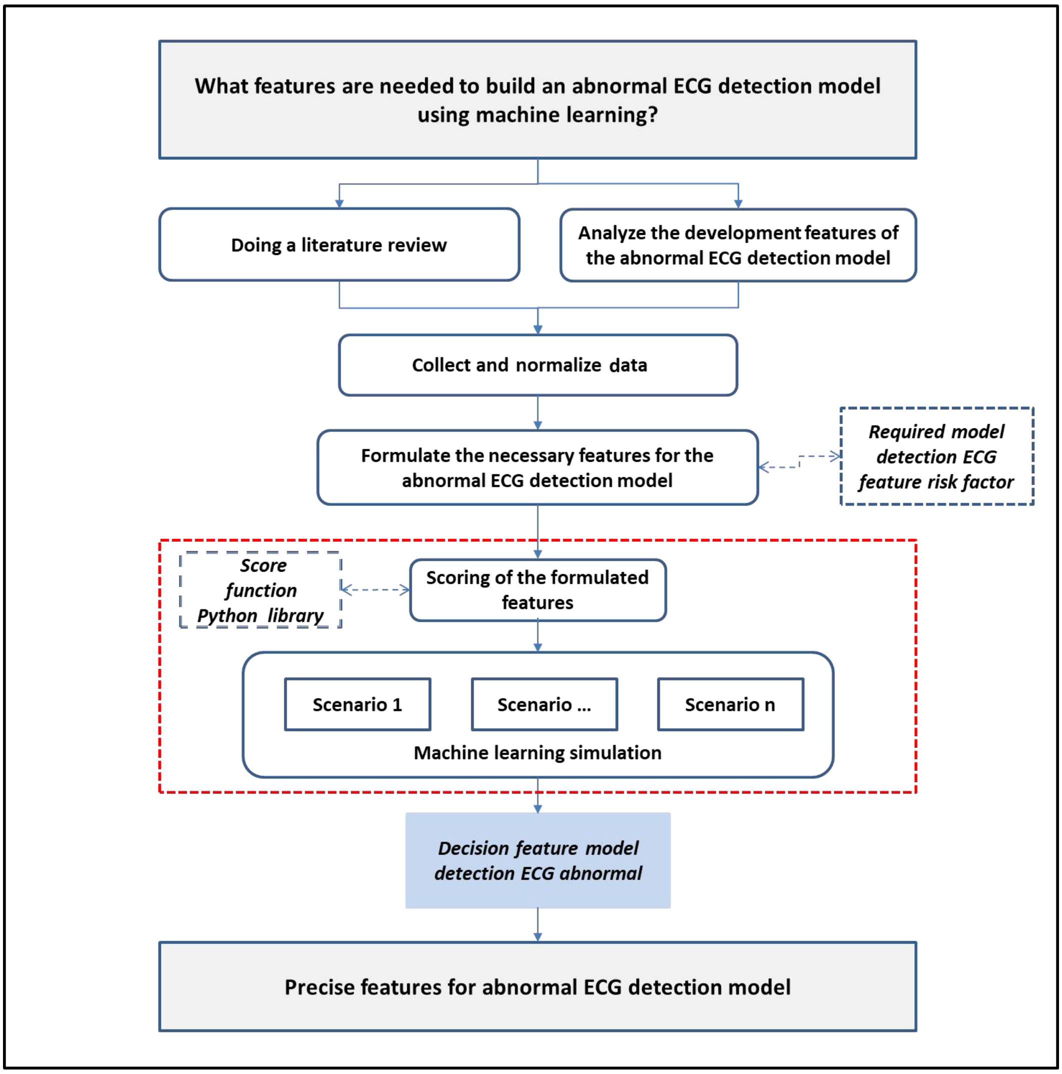

Figure 3. Architecture and model for feature ECG detection proposed

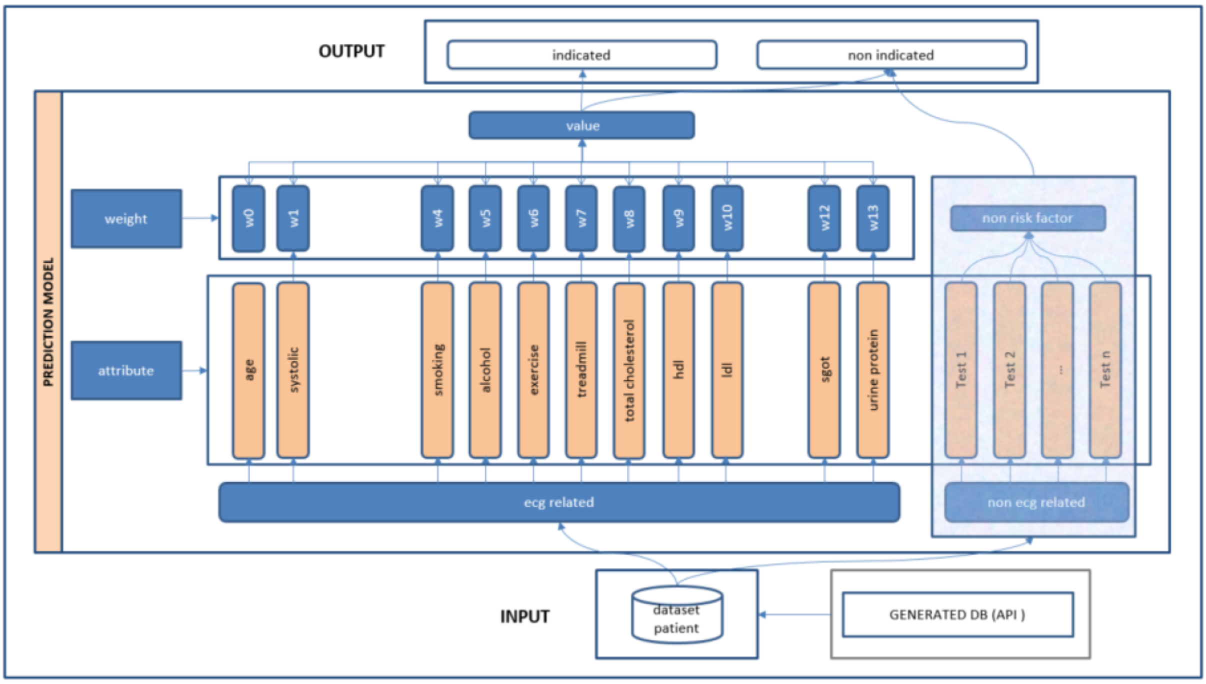

Figure 4. Model detection ECG feature – A risk factor

In the dotted line box, a process is carried out to get the most influential attribute using the scoring function.

Figure 4 shows the model results after running the scoring function for each algorithm; these three attributes have no significant effect on the results of ECG detection. The three attributes that had the most negligible impact were 1) heart rate, 2) obesity, and 3) creatinine. Thus, only 11 attributes have significant impacts on ECG detection.

We discovered a number between 0 and 1 after calculating the weight.

3.2. **Algorithm.** Three algorithms were used in this proposed system. These three were chosen because the researchers wished to determine the accuracy value of the algorithm/pseudocode: a) SVM, b) KNN and c) MLP.

Algorithm 1. Algorithm for Processing SVM Model.
Data: with t_n variables and binary outcome
Output: value of confusion matrix and accuracy
Find the high score values for tuning parameters of the SVM Model
Initialized SVM parameter and structure
Import the library and load the data
Use the function "svm. svc (kernel = 'linear')"
Begin
    Set t to t_0 // Specification for column
    Set n to length(num_column) // Length of column
    For $t_0 <= 1$ to n
        Run scoring(num_column) // Used the scoring function
        Set t_0 to $t_0 + 1$
    End
    Return scoring(num_column) // Number of scores
    Run prediction model
    Call function confusion matrix
    The return value of the result confusion matrix
    Call function accuracy score
    The return value of result accuracy
End

Algorithm 2. Algorithm for Processing KNN Model.
Data: with t_n variables and binary outcome
Output: value of accuracy
Find the high score values for tuning parameters of the KNN Model
Initialized KNN parameter and structure
Import the library and load the data
Use the function "kneighborsclassifier."
Begin
    Set k to a value (choose the value of k) // best value $k = 5$
    For $k = 1$ to n
        Set n to length(num_column) // Length of column
        For $k = 1$ to n
            For $t_0 <= 1$ to n
                Run scoring(num_column) // Used the scoring function
                Set t_0 to $t_0 + 1$
            End
        End
        Return scoring(num_column) // Number of scores
        Run prediction model

Call function accuracy score
The return value(k) of result accuracy
Set k to $k+1$
End
Take the highest score
End

Algorithm 3. Algorithm for Processing MLP Model.
Data: with t_n variables and binary outcome
Output: value of accuracy
Find the high score values for tuning parameters of the MLP Model
Initialized SVM parameter and structure
Import the library and load the data
Use the function "MLPClassifier()."
Begin
Set t to t_0 // Specification for column
Set n to length(num_column) // Length of column
For $t_0 <= 1$ to n
Run scoring(num_column) // Used the scoring function
Set t_0 to $t_0 + 1$
End
The return scoring(num_column) // Number of scores
Run prediction model (fix the best data)
Call function accuracy score
The return value of result accuracy
End

The scoring function analyzes the dataset using the number of influential characteristics starting at 14, 13, 12, and 11 and concludes that 11 should be chosen because they are compelling. The findings were quite significant: 88.285% accuracy for SVM, 89.375% accuracy for KNN, and 88.285% accuracy for MLP.

4. **Result and Discussion.** A dataset investigation was conducted to determine which attributes can be used to predict abnormal ECG heart disease. The dataset contained 1284 records. The records in the dataset were isolated to prepare and test the information. Following information handling, a machine learning approach was used to detect abnormal electrocardiograms.

The existing dataset was analyzed using Python programming, with 80% utilized as training data and 20% used as test data. Table 2 and Figure 5 demonstrate the accuracy of the output result of model, which shows the possibility of suffering from heart disease.

Table 2. Result of model detection ECG feature – A risk factor

	Accuracy	**Recall**	**Precision**
SVM	88.28571	100.00000	100.00000
KNN	89.37514	97.24919	98.36633
MLP	88.28571	100.00000	100.00000

Table 2 shows the data, while Figure 5 shows the graph resulting from the third process algorithm in visualization.

5. **Conclusions.** This study proposed a coronary illness expectation model that utilized machine learning, specifically SVM, KNN, and MLP. Three algorithms are being tested in this study. The three algorithms are processed with the same dataset using a "scoring

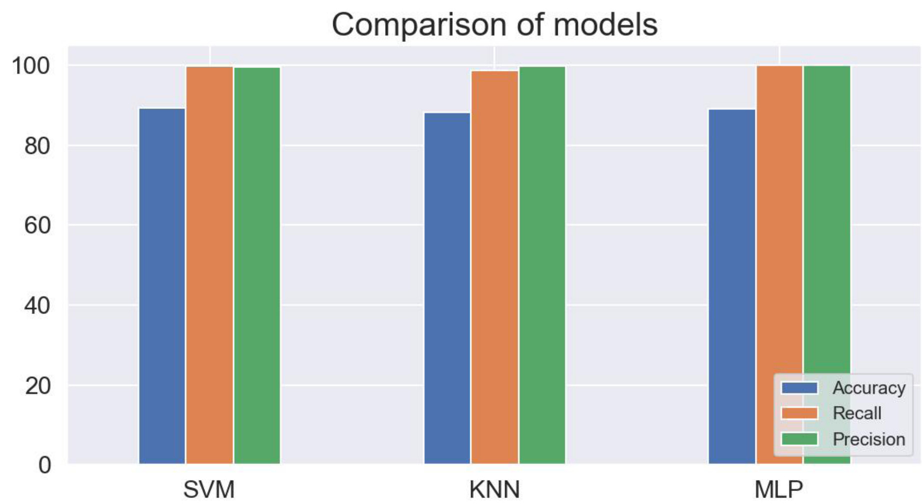

FIGURE 5. Graph of result of model detection ECG feature – A risk factor

function" to determine the most accurate algorithm. Researchers obtained high accuracy values for the three algorithms, namely the KNN algorithm. The findings of this study can be used to make preliminary recommendations for diagnosing heart disease (ECG abnormal). The novelty of this study is that the researcher was able to reduce attributes by using the scoring function, which used 11 features: age, systolic, smoking, alcohol, exercise, treadmill, total cholesterol, HDL, LDL, SGOT, and urine protein. KNN was the most accurate, with an accuracy of 89.375%. Among the datasets available, the KKN algorithm has been shown to have the best accuracy during testing. In the future, we will propose various approaches or data sets for more precise and accurate predictions.

REFERENCES

[1] N. Absar, E. K. Das, S. N. Shoma, M. U. Khandaker, M. H. Miraz, M. R. I. Faruque and R. K. Pathan, The efficacy of machine-learning-supported smart system for heart disease prediction, *Healthcare*, vol.10, no.6, 1137, DOI: 10.3390/healthcare10061137, 2022.

[2] M. Ashtiyani, S. N. Lavasani, A. A. Alvar and M. R. Deevband, Heart rate variability classification using support vector machine and genetic algorithm, *Journal of Biomedical Physics and Engineering*, vol.8, no.4, pp.423-434, DOI: 10.31661/jbpe.v0i0.614, 2018.

[3] E. Faraggi and A. Kloczkowski, A global machine learning based scoring function for protein structure prediction, *Biophysical Journal*, vol.106, no.2, pp.656a-657a, DOI: 10.1016/j.bpj.2013.11.3634, 2014.

[4] J. Ha, S.-G. Kim, D. Paek and J. Park, The magnitude of mortality from ischemic heart disease attributed to occupational factors in Korea – Attributable fraction estimation using meta-analysis, *Safety and Health at Work*, vol.2, no.1, pp.70-82, DOI: 10.5491/SHAW.2011.2.1.70, 2011.

[5] R. Hagan, C. J. Gillan and F. Mallett, Comparison of machine learning methods for the classification of cardiovascular disease, *Informatics in Medicine Unlocked*, vol.24, 100606, DOI: 10.1016/j.imu.2021.100606, 2021.

[6] Y. Li, H. Liu, K. Zhou, H. Qin, W. Yu and Y. Liu, Machine learning approach for delamination detection with feature missing and noise-polluted vibration characteristics, *Composite Structures*, vol.287, 115335, DOI: 10.1016/j.compstruct.2022.115335, 2022.

[7] A. Mincholé, J. Camps, A. Lyon and B. Rodríguez, Machine learning in the electrocardiogram, *Journal of Electrocardiology*, vol.57, pp.S61-S64, DOI: 10.1016/j.jelectrocard.2019.08.008, 2019.

[8] T. Smole, B. Žunkovič, M. Pičulin, E. Kokalj, M. Robnik-Šikonja, M. Kukar and Z. Bosnić, A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy, *Computers in Biology and Medicine*, vol.135, 104648, DOI: 10.1016/j.compbiomed.2021.104648, 2021.

[9] P. Svec, L. Benko, M. Kadlecik, J. Kratochvil and M. Munk, Web usage mining: Data pre-processing impact on found knowledge in predictive modelling, *Procedia Computer Science*, vol.171, pp.168-178, DOI: 10.1016/j.procs.2020.04.018, 2020.

[10] M. Swathy and K. Saruladha, A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using machine learning and deep learning techniques, *ICT Express*, vol.8, no.1, pp.109-116, DOI: 10.1016/j.icte.2021.08.021, 2022.

[11] M. C. Thomas, W. Zhu and J. A. Romagnoli, Data mining and clustering in chemical process databases for monitoring and knowledge discovery, *Journal of Process Control*, vol.67, pp.160-175, DOI: 10.1016/j.jprocont.2017.02.006, 2018.

[12] G. Grewal, T. Polisetty, D. Cannon, A. Ardeljan, R. M. Vakharia, H. C. Rodriguez and J. C. Levy, Alcohol abuse, morbid obesity, depression, congestive heart failure, and chronic pulmonary disease are risk factors for 90-day readmission after arthroscopic rotator cuff repair, *Arthroscopy, Sports Medicine, and Rehabilitation*, vol.4, no.5, pp.e1683-e1691, DOI: 10.1016/j.asmr.2022.06.015, 2022.

[13] X. Wang, Y. Hu, L. Q. Qin and J. Y. Dong, Combined association of central obesity and depressive symptoms with risk of heart disease: A prospective cohort study, *Journal of Affective Disorders*, vol.297, pp.360-365, DOI: 10.1016/j.jad.2021.10.096, 2022.

[14] Y. Wang, Y. Jia, Y. Tian and J. Xiao, Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring, *Expert Systems with Applications*, vol.200, 117013, DOI: 10.1016/j.eswa.2022.117013, 2022.