**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 21, No. 3, March 2021, pp. 1530~1539
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v21.i3.pp1530-1539

# Predicting heart failure using a wrapper-based feature selection

**Minh Tuan Le[1], Minh Thanh Vo[2], Nhat Tan Pham[3], Son V.T Dao[4]**
[1,2]SEE, International University, Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Viet Nam
[3,4]SIEM, VNU-International University, Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Viet Nam

**Article Info**

***Article history:***

Received Oct 30, 2020
Revised Dec 7, 2020
Accepted Dec 23, 2020

***Keywords:***

Feature selection
Grey wolf optimization
Heart failure
Multilayer perceptron
Neural network

**ABSTRACT**

In the current health system, it is very difficult for medical practitioners/ physicians to diagnose the effectiveness of heart contraction. In this research, we proposed a machine learning model to predict heart contraction using an artificial neural network (ANN). We also proposed a novel wrapper-based feature selection utilizing a grey wolf optimization (GWO) to reduce the number of required input attributes. In this work, we compared the results achieved using our method and several conventional machine learning algorithms approaches such as support vector machine, decision tree, K-nearest neighbor, naïve bayes, random forest, and logistic regression. Computational results show not only that much fewer features are needed, but also higher prediction accuracy can be achieved around 87%. This work has the potential to be applicable to clinical practice and become a supporting tool for doctors/physicians.

***Corresponding Author:***

Son V.T Dao
School of Industrial and Management
International University
Vietnam National University HCMC
Ho Chi Minh City, Viet Nam
Email: dvtson@hcmiu.edu.vn

## 1. INTRODUCTION

The term "heart failure" referring to a condition in which the heart's contraction is not as effective as it should be. The heart is a vital organ in the human body because it pumps blood to every other organ. A patient who is living vegetative states still needs the heart to survive. Heart failure (HF) is a chronic condition in which one of the ventricles or atriums on both sides is not able to pump rich oxygen into the body and poor oxygen into the lungs. There are several common reasons cause heart failure. The majority of (HF) patients are elderly. Cardiac arrest often gradually and deliberately develops after parts of a heart got weaken and makes others such as ventricles and atriums do extra workloads to provide enough blood and oxygen to the body [1-2]. With the ubiquitous application of technology in the medical field, it helps the cost of diagnosis to be inexpensive. Unfortunately, nowadays the number of patients who have been diagnoses with heart failure is gradually increasing in suburbs and dramatically increasing in urban areas. Therefore, the earlier for getting diagnosed, the better off it will be for the patients. Because of the difficulty of diagnosing the process of a heart failure condition, it might cause a postponement in treatment operation. Therefore, it is crucial to develop a heart disease prediction system for heart failure to support whoever works in the medical professional field to diagnose patients with conditions more rapid and accurate. Deep learning and Machine learning algorithms have been successfully applied to various field [3-4], especially medical field to support doctor/physician to diagnose various diseases such as heart failure, diabetes. ANN has also been applied by researchers in the medical field [5-7].

In this study, we will use a multilayer perceptron models (MLP) together with preprocessing methods for predicting heart failure patient. Besides, a metaheuristics-based feature selection algorithm grey wolf optimization (GWO) also applied to MLP models to enhance the performance and reduce training time. The result is a benchmark with other common machine learning/deep learning algorithms in the following such as logistic regression (LR) [8], support vector machine (SVM) [9], k-nearest neighbor classifier (KNN) [10], naive bayesian classifier (NBC) [11-12], decision tree (DT) [13], and random forest classifier (RFC) [14] based on the original set of available medical features.

Several data mining methods have been successfully applied to diagnosing heart failure (HF). In [10], Davide Chicco represents a model to diagnostic the survival rate of patients who have been using clinical record HF data. In the research [15], a list of the machine learning methods was used for the binary classification of survival. A random forest classifier outperformed all other methods compared to the other models, which is very impressive in this age [16]. Guidi *et al.* (2013) represent a clinical decision support system (CDSS) for analyzing HF patients and comparing the performance of neural network, support vector machine, fuzzy-genetic, random forest. In [17], the authors used the detailed clinical data on patients hospitalized with HF in Ontario, Canada. In the machine learning literature, alternate classification schemes have been developed such as bootstrap aggregation (bagging), boosting, random forests, and support vector machines. They also compared the ability of these methods to predict the probability of the presence of heart failure with preserved ejection fraction (HFPEF).

Feature selection (FS) is a process that commonly selects in machine learning to solve the high dimensionality problem. In FS, we choose a small number of features but important and usually ignore the irrelevant and noisy features, in order to make the subsequent analysis easier. According to the redundancy and relevance. Yu *et al.*, [18] have classified those feature subsets into four different types: noisy and irrelevant; redundant and weakly relevant, weakly relevant and non-redundant, and strongly relevant. An irrelevant feature does not require predicting accuracy. Furthermore, many approaches can implement with filter and wrapper methods such as models, search strategies, feature quality measures, and feature evaluation. All features play as key factors for determining the hypothesis of the predicting models. Besides that, the number of features and the size of the hypothesis spaces are directly proportional to each other, and so on. When the number of features increases, the size of the searching space also increased. One such outstanding case is that if there are M features with the binary class label in a dataset, it has $2^{2^M}$ combination in the search space.

There are three types of FS methods, which are defined based on the interaction with the learning model, namely filter, wrapper, and embedded methods. The Filter method selects statistics-based features. It is independent of the learning algorithm and thus requires less computational time. Statistical measures such as information gain, chi-square test [19], Fisher score, correlation coefficient, and variance threshold are used to understand the importance of the features. In contrast, the wrapper method's performance highly depends on the classifier. The best subset of features is selected based on the results of the classifier. Wrapper methods are much more computationally expensive than filter methods since it needs to run simultaneously with the classifier many times. However, these methods are more accurate than the filter method. Some of the wrapper examples are recursive feature elimination [20], Sequential feature selection algorithms [21], and genetic algorithms [22]. Thirdly, the embedded method which utilizes ensemble learning and hybrid learning methods for feature selection. This method has a collective decision; therefore, its performance is better than the previous one. One example is the random forest which is less computationally intensive than wrapper methods. One drawback of the embedded method is that it is specific to a learning model.

Many evolutionary metaheuristics-based feature selection methods are also proposed, many of them are wrapper type since it has been proven that wrapper provides better performance [23]. Too *et al.*, [24] proposed a competitive binary grey wolf optimizer (CBGWO), which is based on the grey wolf optimizer (GWO) proposed by Mirjalili *et al.* [25], for feature selection problem in EMG signal classification. The results showed that CBGWO outranked other algorithms in terms of performance for that case study. Many other wrapper-based feature selection algorithms were also introduced in many previous works to select a subset of features, including binary grey wolf optimization (BGWO) [26], binary particle swarm optimization (BPSO) [27], ant colony optimization (ACO) [28], and binary differential evolution (BDE) [29].

## 2. RESEARCH METHOD

This system includes two steps: data pre-processing which involved outlier detection. Then followed by a multilayer perceptron (MLP). The outlier detection in this research is using interquartile range (IQR) method and then applying grey wolf optimizer to optimize the architecture of multilayer perceptron for classifying the heart failure patients. The detail of this method is described in this section.

### 2.1. Data Collection

Heart failure clinical records data set has been used in this research, which is records heart failure patients from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan). The dataset is available on the UCI repository. The target of this binary classification has two categorial value: 1-yes (patient is sick) and 0-no (patient is healthy). The attribute for predicting is "DEATH_EVENT" which contains two categorical values and is considered as a binary classification problem. Table 1 list the number of instance number of attribute and features of the dataset.

The dataset contains 299 instances and 12 attributes. Each of these attributes is physiological measurements. The patients in this dataset include 194 men and 105 women and the range of their ages between 40 and 95 years old. Features, measurements, and range are listed in Table 1.

Table 1. Features, measurement, meaning, and range of the dataset

| Feature | Explanation | Measurement | Range |
|---|---|---|---|
| Age | Age of the patient | years | [40, …,95] |
| Anaemia | Decrease of red blood cells or hemoglobin | Boolean | 0,1 |
| High blood pressure | If the patient has hypertension | Boolean | 0,1 |
| Creatinine phosphokinase (CPK) | Level of the CPK enzyme in the blood | mcg/L | [23, ....,7861] |
| Diabetes | If the patient has diabetes | Boolean | 0, 1 |
| Ejection fraction | Percentage of blood leaving the heart at each contraction | % | [14, ...,80] |
| Sex | Platelets in the blood | binary | 0, 1 |
| Platelets | woman or man | kiloplatelets/mL | [25.01, ...,850.00] |
| Serum creatinine | Level of serum creatinine in the blood | mg/dL | [0.50, …9.40] |
| Serum sodium | Level of serum sodium in the blood | mEq/L | [114, ...,148] |
| Smoking | If the patient smokes or not | Boolean | 0, 1 |
| Time | Follow-up period | days | [4, ...,285] |
| [target] death event | If the patient deceased during the follow-up period | Boolean | 0, 1 |

### 2.2. Data Preprocessing

For preprocessing data, we use the normalization and IQR method of outlier detection. Before training data, we normalize this dataset since the gradient descent will be effective with normalized (scaled) values. We may get the values in the different scales if we would not normalize the data. To adjust weights, our model would take more time to train on this data. However, if we normalize our data by using normalization techniques, we will have numbers on the same scale which will make our model train much faster and gradient descent will be effective in this case. IQR Method of outlier detection, which is used for pre-processing data IQR (short for "interquartile range") in the middle spread, is also known as the quartile range of the dataset. This concept is used in statistical analysis to help conclude a set of numbers. IQR is used for the range of variation because it excludes most outliers of data.

In Figure 1, the minimum, maximum is the minimum and maximum value in the dataset. The median is also called the second quartile of the data. Q1 is the first quartile of the data, it means that 25% of the data is lies between minimum and Q1. And Q3 is the third quartile of the data, it says that 75% of the data lies between maximum and Q3. The equation below is the Inter-Quartile Range or IQR, which is the difference between Q3 and Q1.

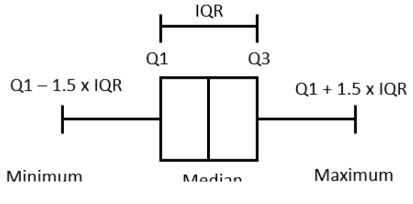


Figure 1. A box-whisker plot

$$IQR = Q3 - Q1 \qquad (1)$$

For detecting the outliers with this technique, we have to define a new range, which is called a decision range. If any point lies outside this range, it is considered an outlier. This range is (2), (3):

$$Lower\ Bound: (Q1\ - 1.5 * IQR) \quad (2)$$

$$Upper\ Bound: (Q3 + 1.5 * IQR) \quad (3)$$

## 2.3. Research Methodology

### 2.3.1. Grey wolf optimizer (GWO)

Swarm intelligence is the way of communication between an individual and a group. The application of herd intelligence in the fields of industry, science, and commerce has many unique and diverse applications. Research in herd intelligence can help people manage complex systems. GWO simulates the way that the wolves look for food and survive by avoiding their enemies (Figure 2). GWO was firstly introduced by Mirjalili *et al*., 2014 [25]. Alpha means that the leader gives the decision for a sleeping place, hunting grey, time to wake up. The second level of gray wolves is beta. The betas are the wolves in herds under alpha but also commanded another low-level wolf. The lowest rank among the gray wolves is Omega. They are weak wolves and have to rely on other wolves in the pack. Delta ones are dependent on alphas and betas, but they are more effective than omega. They are responsible for monitoring territorial boundaries and warning inside in case of danger, protect and ensure safety for herds, take care of the weak, and illness wolves in the pack.

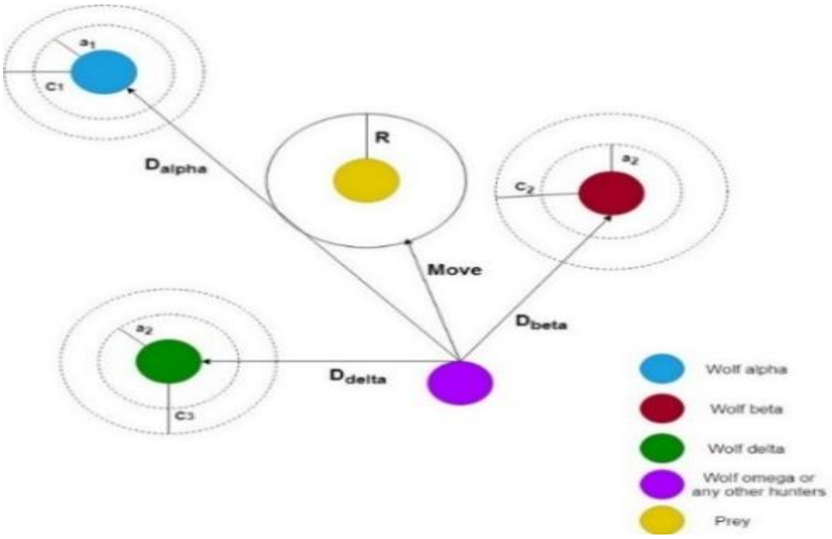


Figure 2. Position updating in GWO

To develop the mathematical model, the best solution is considered as alpha. Beta and delta are the second and the third solution, respectively. The step of GWO is encircling prey is shown (4), (5):

$$\vec{D} = \left|\vec{C}.\overrightarrow{X_p}(t) - \vec{X}(t)\right|) \quad (4)$$

$$\vec{X}(t+1) = \overrightarrow{X_p}(t) - \vec{A}.\vec{D} \quad (5)$$

where t shows the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors, $X$ is the position vector of a grey wolf and $\overrightarrow{X_p}$ is the position vector of the prey. The coefficient is indicated in the (6), (7):

$$\vec{A} = 2\vec{a}\,.(\vec{r}_1 - \vec{a}) \quad (6)$$

$$\vec{C} = 2\vec{r}_2 \quad (7)$$

where $\vec{a}$ is linearly decreased from 2 to 0, $\vec{r}_1$ and $\vec{r}_2$ are random vector in [0, 1].

These equations below define the final position of the wolf $\vec{X}(t+1)$:

$$\vec{D}_\alpha = |\vec{C}_1.\overrightarrow{X_\alpha} - \vec{X}| \tag{8}$$

$$\vec{D}_\beta = |\vec{C}_2.\overrightarrow{X_\beta} - \vec{X}| \tag{9}$$

$$\vec{D}_\delta = |\vec{C}_3.\overrightarrow{X_\delta} - \vec{X}| \tag{10}$$

$$\vec{X}_1 = \overrightarrow{X_\alpha} - \overrightarrow{A_1}.\overrightarrow{D_\alpha} \tag{11}$$

$$\vec{X}_2 = \overrightarrow{X_\beta} - \overrightarrow{A_2}.\overrightarrow{D_\beta} \tag{12}$$

$$\vec{X}_3 = \overrightarrow{X_\delta} - \overrightarrow{A_3}.\overrightarrow{D_\delta} \tag{13}$$

$$\vec{X}(t+1) = \frac{\vec{X}_1+\vec{X}_2+\vec{X}_3}{3} \tag{14}$$

### 2.3.2. Multilayer perceptron (MLP)

The single-layer perceptron solves only linearly separable problems, but some complex problems are not linearly separable. Therefore, in order to solve some complex problems, one or more layers are added in a single layer perceptron, so it is known as a multilayer perceptron (MLP) [30-33]. The MLP network is also known as a feed-forward neural network having one or more hidden layers as can be seen in Figure 3.

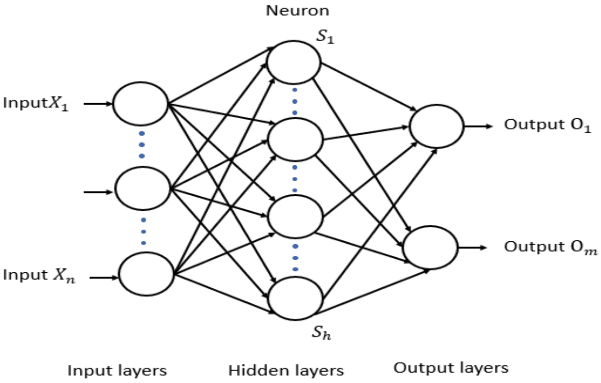


Figure 3. The Architecture of a multilayer perceptron

In Figure 3, the neural network has an input layer with n neurons, one hidden layer with n neurons, and an output layer

- Input layer: call input variable ($x_1$, …, $x_n$), also called the visible layer
- Hidden layer: the layer of the node lies between the input and output layer.
- Output layer: this layer produces the output variables

The following steps below show the calculation of the MLP output after giving the weights, inputs, and biases:

- The weighted sums of inputs are calculated as follow:

$$s_j = \sum_{i=1}^{n}(W_{ij}.X_i) - \theta_j, j = 1,2, \dots h \tag{15}$$

where $X_i$ shows the $i$th input, $n$ represent the number of nodes, $W_{ij}$ is the connection weight from the $i$th node to the $j$th node and $\theta_j$ is the threshold of the hidden node.

- The calculation of the output of each hidden node:

$$S_j = sigmoid\ (s_j) = \frac{1}{(1+\exp(-s_j))}, j = 1,2, \dots h \tag{16}$$

– The final outputs are based on the calculation of the output of hidden nodes:

$$o_k = \sum_{i=1}^{n}(w_{jk}.S_j) - \theta'_k, k = 1,2, \dots m \tag{17}$$

$$O_k = sigmoid\ (o_k) = \frac{1}{(1+\exp(-o_k))}, k = 1,2, \dots, m \tag{18}$$

where $w_{jk}$ is the connection weight from $j^{th}$ to $k^{th}$ and $\theta'_k$ is the threshold of the $k^{th}$ output node.

For the definition of the final output, the weights and biases are used. We find the values for weights and biases to achieve a relationship between the inputs and outputs. In this algorithm, weights and biases have been adjusted repeatedly for minimizing the actual output vector of the network and output vector.

## 3. RESULTS AND ANALYSIS

### 3.1. Performance Evaluation

In this system, the performance of these algorithms is studied based on performance metrics such as accuracy, precision, recall, F1 score, which are given in these (19-22):

$$Accuracy\ = \frac{TP+TN}{TP\ +TN+FP+FN} \tag{19}$$

$$Precision\ = \frac{\mathrm{TP}}{TP\ +FP} \tag{20}$$

$$Recall\ = \frac{\mathrm{TP}}{TP\ +FN} \tag{21}$$

$$F_1\ = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{22}$$

Where: a true positive (TP): the samples are classified as true (T) while they are (T); a true negative (TN): the samples are classified as false (F) while they are (F); a false positive (FP): the samples are classified as (T) while they are (F); A false negative (FN): the samples are classified as (F) while they are (T).

### 3.2. Analysis Results Using Various Machine Learning Models

In this research, six classifier models LR, KNN, SVM, NB, DT, RFC are used. The dataset is divided into 80:20, which is 80% of data for training the models and 20% is used for testing the accuracy of the models. In this research, we apply the removal of the outlier dataset for training.

The bar chart in the Figure 4 indicates the accuracy of machine learning algorithms. As can be seen from the figure that the random forest classifier is having the highest accuracy with 85% compared to the other algorithms. LR also achieves good accuracy as compared to SVM.

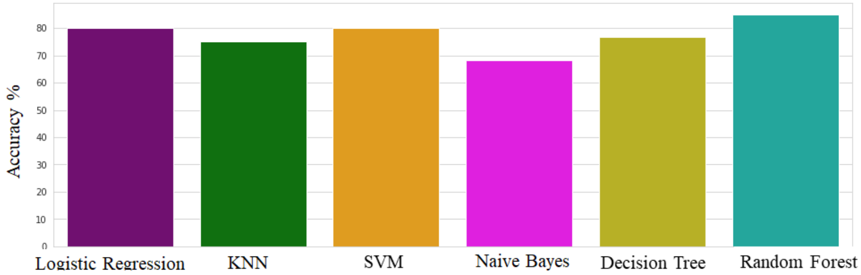


Figure 4. Classification accuracy with different models

Table 2 indicates the classification results of the six models based on accuracy, recall, precision, $F_1$ score on the heart failure dataset. It can be seen on the table that, among the six methods, the RFC method performs the best performance, with the highest accuracy 85%, 65.21% recall, 93.75% precision, and 76% $F_1$ score than the other machine learning models.

Table 2. The prediction results

| ID | Method | Accuracy | Recall | Precision | $F_1$ score |
|---|---|---|---|---|---|
| 1 | LR | 0.80 | 0.4782 | 1 | 0.6470 |
| 2 | KNN | 0.75 | 0.4782 | 0.7857 | 0.5945 |
| 3 | SVM | 0.80 | 0.4782 | 1 | 0.6470 |
| 4 | NBC | 0.6833 | 0.3043 | 0.7 | 0.4242 |
| 5 | DT | 0.7667 | 0.6086 | 0.7368 | 0.6666 |
| 6 | RFC | 0.85 | 0.6521 | 0.9375 | 0.7692 |

Figure 5 shows the confusion matrix of all algorithms. The diagonal elements of the matrix are the correctly classified number of points for each data layer. From here, the accuracy can be inferred by the sum of the elements on the diagonal divided by the sum of the elements of the entire matrix. A good model will give a confusion matrix with the elements on the main diagonal having a big value and the remaining elements having a small value. It can be seen from the confusion matrix that the main diagonal elements of the matrix of LR, SVM, and RFC have a higher value.

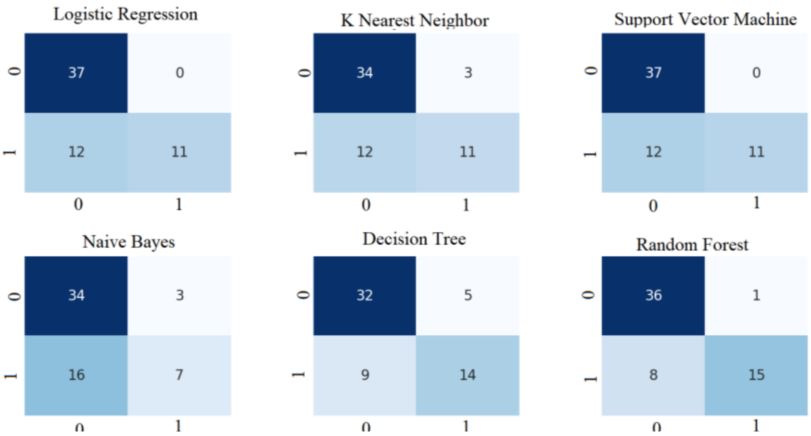


Figure 5. Confusion matrix of all algorithms

The receiver operating characteristic (ROC) plot is a measurement for evaluating the classifier performance of each algorithm, as shown in Figure 6.

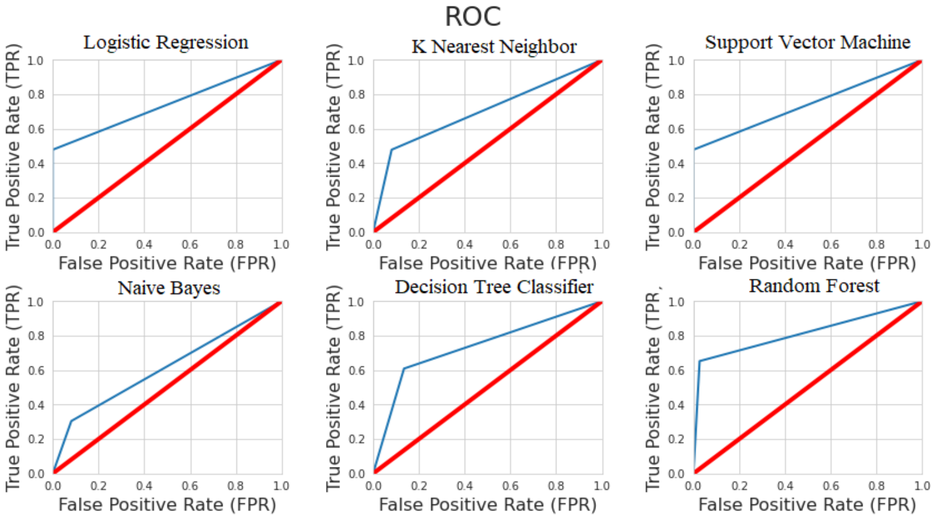


Figure 6. Receiver operating characteristic (ROC) of all algorithm

### 3.3. Experimental Results of GWO-MLP

In this work, we show the experimental result when grey wolf optimization is applied to multilayer perceptron (GWO-MLP). The Table 3 shows the samples of the selected feature using grey wolf optimization and Figure 7 shows Convergence curve of the fitness function.

Table 3. Samples of feature selection

| ID | Feature | Feature selection |
|---|---|---|
| 1 | age | Selected |
| 2 | anaemia | Selected |
| 3 | creatinine_phosphokinase | x |
| 4 | diabetes | x |
| 5 | ejection_fraction | Selected |
| 6 | high_blood_pressure | Selected |
| 7 | platelets | x |
| 8 | serum_creatinine | x |
| 9 | serum_sodium | x |
| 10 | sex | Selected |
| 11 | smoking | x |
| 12 | time | Selected |

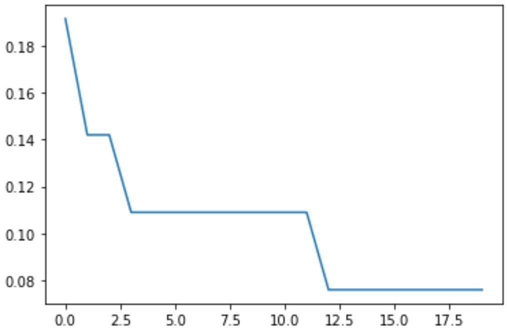
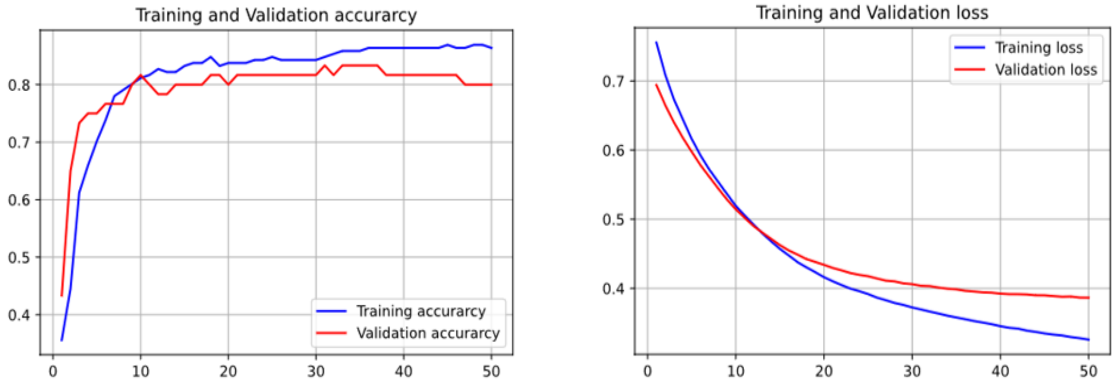


Figure 7. Convergence curve of the fitness function

We can see that 6 of the 12 features are selected. After that, this subset of features is trained on the MLP with 100 epochs, which yields the following result in Figure 8. In this approach, the performance of GWO-MLP achieved 87% of accuracy, 74% recall, 77% precision, and 76% F1 score. With this approach, only six out of the original twelve features were selected, the accuracy of the GWO-MLP model was higher than the other methods, indicating there were unuseful features in the data. Furthermore, the training time of this data is lower than the other models.



(a) (b)

Figure 8. (a) Training and Validation accuracy result, (b) Training and Validation loss result

## 4. CONCLUSION

In this paper, we propose a machine learning model to predict heart failure using an ANN. At first, a wrapper-based feature selection approach using a metaheuristic called GWO to select 6 features out of the original 12 features. These features are used as inputs for the MLP for the prediction task. Our proposed results achieve an accuracy of 87%, which shows that our approach outperformed other machine learning models such as SVM, LR, KNN, NBC, DT, RFC. Furthermore, with fewer features, our machine learning model is much simpler and requires much less computational effort. Potential future works are listed as follows: fine-tuning the MLP architecture, i.e the number of hidden layers and hidden nodes, as well as the activation functions; or optimizing the parameters of the feature selection algorithm for achieving a better performance.

## REFERENCES

[1] T. Ee, P. Tg, K. Gs, N. Kk, and F. Di, “Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques.,” *Comput Struct Biotechnol J*, vol. 15, pp. 26–47, Nov. 2016, doi: 10.1016/j.csbj.2016.11.001.

[2] L. Hussain, I. A. Awan, W. Aziz, S. Saeed, and A. Ali, “Detecting Congestive Heart Failure by Extracting Multimodal Features and Employing Machine Learning Techniques,” *BioMed Research International*, vol. 2020, pp. 1–19, 2020, doi: https://doi.org/10.1155/2020/4281243.

[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[4] M. Rungruanganukul and T. Siriborvornratanakul, “Deep Learning Based Gesture Classification for Hand Physical Therapy Interactive Program,” in Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Posture, Motion and Health, Jul. 2020, pp. 349–358, doi: 10.1007/978-3-030-49904-4_26.

[5] C. S. Dangare and S. S. Apte, “A Data Mining Approach for Prediction of Heart Disease Using Neural Networks”, *International Journal of Computer Engineering and Technology(IJCET)*, vol. 3, no. 3, pp. 30–40, Oct. 2012.

[6] S. Smiley, “Diagnostic for Heart Disease with Machine Learning,” *Medium*, Jan. 12, 2020. https://towardsdatascience.com/diagnostic-for-heart-disease-with-machine-learning-81b064a3c1dd (accessed Sep. 19, 2020).

[7] D. Shen, G. Wu, and H.-I. Suk, “Deep Learning in Medical Image Analysis,” *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[8] R. E. Wright, “Logistic regression,” in *Reading and understanding multivariate statistics*, Washington, DC, US: American Psychological Association, 1995, pp. 217–244.

[9] Vladimir N. Vapnik, "Statistical Learning Theory", Canada, A Wiley-Interscience Publication, Sep. 1998

[10] N. S. Altman, “An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression: The American Statistician" *The American Statistician*: Vol 46, No 3, pp. 175-185, Aug. 1992, [Online]. Available: https://doi.org/10.2307/2685209.

[11] K. M. Ting and Z. Zheng, “Improving the Performance of Boosting for Naive Bayesian Classification,” in *Methodologies for Knowledge Discovery and Data Mining*, Berlin, Heidelberg, 1999, pp. 296–305, [Online]. Available: https://doi.org/10.1007/3-540-48912-6_41.

[12] C. Kerdvibulvech, “Human Hand Motion Recognition Using an Extended Particle Filter,” in Articulated Motion and Deformable Objects, Cham, 2014, pp. 71–80, doi: 10.1007/978-3-319-08849-5_8.

[13] J. R. Quinlan, “Induction of decision trees” *Mach Learn 1*, vol. 1, no. 1, pp. 81–106, Mar. 1986, [Online]. Available: https://doi.org/10.1007/BF00116251.

[14] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324

[15] S. Angraal *et al.*, “Machine Learning Prediction of Mortality and Hospitalization in Heart Failure with Preserved Ejection Fraction,” *JACC: Heart Failure*, vol. 8, no. 1, pp. 12–21, Jan. 2020, [Online]. Available: https://doi.org/10.1016/j.jchf.2019.06.013.

[16] D. Chicco and G. Jurman, “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 16, Feb. 2020, [Online]. Available: https://doi.org/10.1186/s12911-020-1023-5.

[17] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, “Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes,” Journal of Clinical Epidemiology, vol. 66, no. 4, pp. 398–407, Apr. 2013, doi: 10.1016/j.jclinepi.2012.11.008.

[18] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, Dec. 2004.

[19] Y. Yang and J. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” 1997.

[20] K. Yan and D. Zhang, “Feature selection and analysis on correlated gas sensor data with recursive feature elimination,” *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, Jun. 2015, [Online]. Available: https://doi.org/10.1016/j.snb.2015.02.025.

[21] A. Jain and D. Zongker, “Feature selection: evaluation, application, and small sample performance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, Feb. 1997, [Online]. Available: https://doi.org/10.1109/34.574797.

[22] C. De Stefano, F. Fontanella, C. Marrocco, and A. Scotto di Freca, “A GA-based feature selection approach with an application to handwritten character recognition,” *Pattern Recognition Letters*, vol. 35, pp. 130–141, Jan. 2014, [Online]. Available: https://doi.org/10.1016/j.patrec.2013.01.026.

[23] E. Zorarpacı and S. A. Özel, “A hybrid approach of differential evolution and artificial bee colony for feature selection,” *Expert Systems with Applications*, vol. 62, pp. 91–103, Nov. 2016, [Online]. Available: https://doi.org/10.1016/j.eswa.2016.06.004.

[24] J. Too, A. R. Abdullah, N. Mohd Saad, N. Mohd Ali, and W. Tee, “A New Competitive Binary Grey Wolf Optimizer to Solve the Feature Selection Problem in EMG Signals Classification,” *Computers*, vol. 7, no. 4, Art. no. 4, Dec. 2018, [Online]. Available: https://doi.org/10.3390/computers7040058.

[25] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” *Advances in Engineering Software*, vol. 69, pp. 46–61, Mar. 2014, [Online]. Available: https://doi.org/10.1016/j.advengsoft.2013.12.007.

[26] E. Emary, Hossam M. Zawbaa, “Binary Grey Wolf Optimization Approaches for Feature Selection” *Neurocomputing*, vol. 172, pp. 371-381, Jan. 2016, [Online]. Available: https://doi.org/10.1016/j.neucom.2015.06.083.

[27] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, and C.-H. Yang, “Improved binary PSO for feature selection using gene expression data” *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 29–38, Feb. 2008, [Online]. Available: https://doi.org/10.1016/j.compbiolchem.2007.09.005.

[28] M. H. Aghdam, N. G. Aghaee, M. EhsanBasiri “Text feature selection using ant colony optimization" *Expert systems with applications*, vol. 36, pp. 6843-6853, April. 2009, [Online]. Available: https://doi.org/10.1016/j.eswa.2008.08.022.

[29] X. He, Q. Zhang, N. Sun and Y. Dong, "Feature Selection with Discrete Binary Differential Evolution," *2009 International Conference on Artificial Intelligence and Computational Intelligence*, pp. 327-330, 2009, [Online]. Available: https://doi.org/10.1109/AICI.2009.438.

[30] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron," *2017 Intelligent Systems Conference (IntelliSys)*, pp. 722-728, 2017, [Online]. Available: https://doi.org/10.1109/IntelliSys.2017.8324209.

[31] Jahangir, M., Afzal, H., Ahmed, M. *et al*, "Auto-MeDiSine: an auto-tunable medical decision support engine using an automated class outlier detection method and AutoMLP," *Neural Computing and Applications*, pp. 2621–2633 (2020), [Online]. Available: https://doi.org/10.1007/s00521-019-04137-5.

[32] Chaudhuri, B. B., and U. Bhattacharya. “Efficient Training and Improved Performance of Multilayer Perceptron in Pattern Classification.” *Neurocomputing*, 34, no. 1 (September 1, 2000): 11–27. https://doi.org/10.1016/S0925-2312(00)00305-2.

[33] Orhan, Umut, Mahmut Hekim, and Mahmut Ozer. “EEG Signals Classification Using the K-Means Clustering and a Multilayer Perceptron Neural Network Model.” *Expert Systems with Applications*, 38, no. 10 (September 15, 2011): 13475–81. https://doi.org/10.1016/j.eswa.2011.04.149.