Jurnal E-Komtek
Vol. 6, No. 2 (2022) pp. 292-302
https://jurnal.politeknik-kebumen.ac.id/index.php/E-KOMTEK
p-ISSN : 2580-3719 e-ISSN : 2622-3066

Application of Artificial Neural Network Algorithms to Heart Disease Prediction Models with Python Programming

Yuda Syahidin, Aditya Pratama Ismail, Fawwaz Nafis Siraj
Department of Information Systems Information, Politeknik Piksi Ganesha, Bandung, Indonesia, 40274

yudasy@gmail.com

https://doi.org/10.37339/e-komtek.v6i2.932

Published by Politeknik Piksi Ganesha Indonesia

Artikel Info
Submitted:
11-06-2022
Revised:
21-12-2022
Accepted:
22-12-2022
Online first :
31-12-2022

Abstract

Heart disease is one of the deadliest diseases in and is the number one killer in the world so many studies are carried out to contribute to predicting a person's heart disease. This study aims to help create an early heart disease prediction model from the UCI Machine Learning Repository dataset. The method proposed in this study is a deep learning technique that applies an artificial neural network algorithm with a hidden layer technique in making a heart disease prediction model. This research stage found problems in improving the accuracy of the datasets used by dealing with problems in pre-processing data, such as missing data and determining the form of data correlation. The model was then tested through a heart disease dataset and yielded 90% accuracy. With the creation of this prediction model with python programming, it is hoped that in addition to helping to make disease predictions, it can also provide further innovations in data science in the health sector.

Keywords: Heart disease, Artificial neural network, Prediction, Python

Abstrak

Penyakit jantung merupakan salah satu penyakit paling mematikan di dunia dan merupakan pembunuh nomor satu di dunia sehingga banyak penelitian dilakukan untuk berkontribusi dalam memprediksi penyakit jantung seseorang. Penelitian ini bertujuan untuk membantu membuat model prediksi penyakit jantung secara dini yang diperoleh dari dataset UCI Machine Learning Repository. Metode yang diusulkan dalam penelitian ini adalah dengan teknik deep learning yang menerapkan algoritma jaringan syaraf tiruan dengan teknik lapisan tersembunyi dalam pembuatan model prediksi penyakit jantung. Tahapan penelitian ini menemukan permasalahan dalam meningkatkan akurasi dataset yang digunakan dengan mengatasi permasalahan pada pre-processing data seperti missing data dan penentuan bentuk korelasi data. Model tersebut, yang kemudian diuji melalui kumpulan data penyakit jantung, menghasilkan akurasi 90%. Dengan terciptanya model prediksi dengan pemrograman python ini, diharapkan selain dapat membantu dalam melakukan prediksi penyakit, juga dapat memberikan inovasi lebih lanjut dalam ilmu data di bidang kesehatan.

***Kata-kata kunci**: Penyakit jantung, Jaringan saraf tiruan, Prediksi, Python*

1. Introduction

Being in life or living in modern big cities impacts our health in various ways. It happens because of: 1) stress associated with urban life, 2) a sedentary lifestyle due to working conditions and lack of time, 3) air pollution that occurs, and 4) the number of people living in poverty; urban populations face an increased risk for the development of chronic health conditions. An increasing percentage of the world's population faces adverse health impacts from urban life. In particular, according to the United Nations **[1],** 54% of the earth's population is in urban areas, which is expected to reach 66% by 2050. Heart disease is still the number one killer in the world. For the latest findings, this disease triggered one-third of all deaths in the world in 2019. The number of deaths continues to grow. China had the highest number of deaths from heart disease in later years **[2].** Cardiovascular diseases such as heart disease, cancer, stroke, and kidney failure continue to increase yearly and rank as the highest cause of death in Indonesia, especially at productive ages. Riskesdas data shows the prevalence of cardiovascular diseases such as hypertension increased from 25.8% (2013) to 34.1% (2018), stroke from 12.1 per mile (2013) to 10.9 per mile (2018), coronary heart disease remains 1 .5% (2013-2018), chronic kidney failure, from 0.2% (2013) to 0.38% (2018). The 2018 Riskesdas data also reported that the prevalence of heart disease based on a doctor's diagnosis in Indonesia reached 1.5%, with the highest prevalence in North Kalimantan Province at 2.2%, DIY at 2%, and Gorontalo at 2%. Apart from these three provinces, there are also 8 other provinces with a higher prevalence when compared to the national prevalence. The eight provinces are Aceh (1.6%), West Sumatra (1.6%), DKI Jakarta (1.9%), West Java (1.6%), Central Java (1.6%), Kalimantan East (1.9%), North Sulawesi (1.8%) and Central Sulawesi (1.9%). Coronary Heart Disease in Indonesia is caused by unhealthy lifestyle changes such as smoking and an unbalanced diet. During the current pandemic, people with comorbidities, especially cardiovascular disease, have a very high risk of being exposed to COVID-19 because they are afraid it can cause worsening and even death **[3].**

Previous research with data mining techniques using the Naïve Bayes algorithm by Riani et al. **[4]** in predicting heart disease resulted in an accuracy of 86% for the 303 datasets tested. Derisma **[5]** conducted a heart disease prediction study by comparing 3 (Three) Algortima data mining, namely Naïve Bayes, Random Forest and Neural Network, which resulted in an average accuracy of 83%. Alhamad, Apriyanto et al**. [6]** researched heart disease prediction using machine learning methods. Ensemble-based – Weighted Vote that pays attention to missing value (MV), Data Validation (DV), Unbalanced Class (UC) and Noisy Data (ND) issues resulted in 85.21%

accuracy. The public dataset commonly used by researchers in creating highly accurate Machine Learning models for heart disease predictions is in the UCI Machine Learning Repository.

The contribution that can be made in this study is to propose Deep Learning Techniques by using artificial neural network algorithms to improve accuracy in predicting heart disease using datasets from the UCI Machine Learning Repository. This prediction is an essential first step towards enabling prevention for health systems that target individuals in need of health resources more effectively. Most of the health data is still untapped, for it needs to be done against unstructured data analysis (Unstructured Data Analysis) to be the next innovation in data science in the health field.

2. Method

Deep Learning [7] is a subfield of machine learning in artificial intelligence (Artificial Intelligence) in handling algorithms inspired by biological structures and brain functions to help machines with intelligence.

This realization has led to the distinction of field sequences based on data. The new rules of thumb established that machine learning would not improve performance by improving training data after a certain threshold. In contrast, Deep Learning can utilize surplus data more effectively for performance improvement. The following **Figure 1**. The performance Model illustrates the overall model performance idea with data sizes for the three fields.

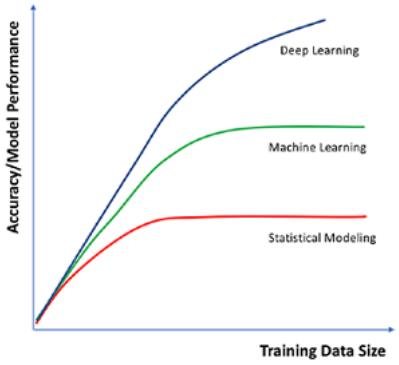


Figure 1. Performance Model [7]

An artificial neural network [8],[7] is a computational network (system of nodes and interconnections between nodes) inspired by biological networks that are neural networks that are complex networks of neurons in the human brain. Nodes created in the Artificial Neural Network are supposed to be programmed to behave like real neurons. **Figure 2** and **Figure 3** show the network of nodes (artificial neurons) that make up an artificial neural network.

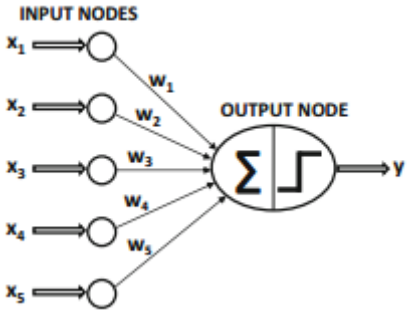


Figure 2. Unbiased perceptron

With the equation formula:

$$\hat{y} = sign\left\{\sum_{j=1}^{d} \quad w_j\, x_j\right\} \qquad (1)$$

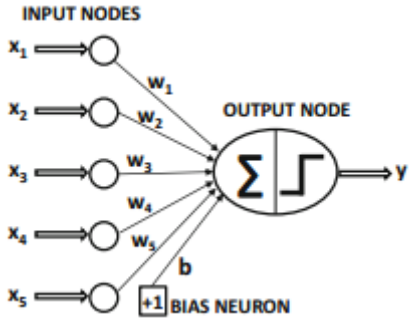


Figure 3. Perceptron with bias

With the equation formula as follows:

$$\hat{y} = sign\left\{\sum_{j=1}^{d} \quad w_j\, x_j + b\right\} \qquad (2)$$

TensorFlow [7] has the unique ability to perform computational partial subgraphs thus enabling distributed training with the help of neural networks later. In other words, TensorFlow allows model parallelism and data parallelism. Heart disease [14] is the formation of a disruption of the balance between the supply and blood needs that occur due to the blockage of blood vessels. Deaths from heart disease reached 959. 227 sufferers, namely 41.4% of all deaths or 2600 residents, died from heart disease daily.

Heart diseases [9] include aortic regurgitation, cardiogenic shock, congenital heart disease, cardiomyopathy, peripartum cardiomyopathy, and tricuspid regurgitation, which is often infected in children's age and is always a significant problem in growing countries. Keras [7], [10] is a compact and easy-to-learn high-level Python library for deep learning that can run on top of TensorFlow. This development focuses more on the main concepts of deep learning, such as creating layers for neural networks. TensorFlow is the back end for Keras.

Keras is used for deep learning applications without interacting with complex TensorFlow. There are two main types of frameworks: sequential APIs and functional APIs. Sequential APIs are based on the idea of a sequence of layers; this is the most common use of Keras and the easiest part of Keras. Sequential models can be considered a stack of linear layers. The research steps can be seen in **Figure 4** Of the Outcome Prediction Model Methodology

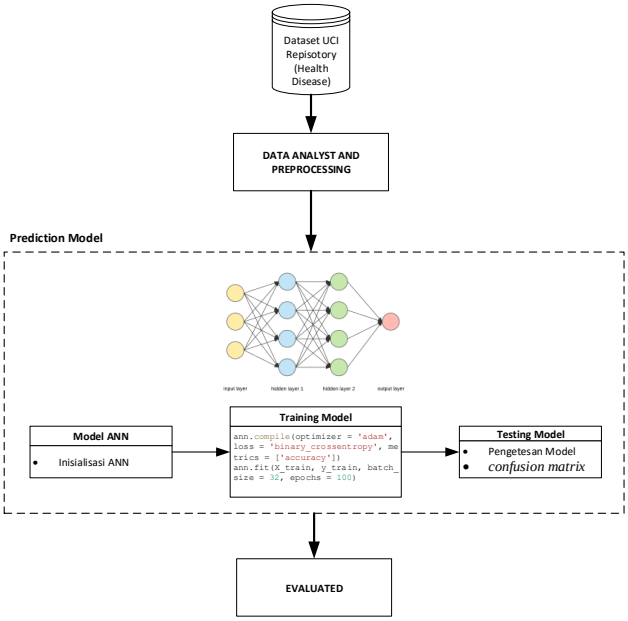


Figure 4. Predictive Model

The data analysis stage analyzes the dataset used to find out the names of features and the correlation between their features, the dataset used is obtained from the UCI Repository. Preprocessing data performs the stages of Splitting datasets into Training sets and Test sets and performs Feature Scaling to check variables that have very varied and random values so that this stage makes numerical data on the dataset have the same value range (scale).

This stage conducts model selection activities in generating outcome prediction using deep learning techniques, namely artificial neural network (ANN). This activity conducts ANN Initialization and Model Training and tests the model that has been made. Scripting the creation of this model assisted by Hard library is a Python library **[11],[8]** intended to develop and evaluate deep learning models.

This stage produces a model form of making predictions by testing models that testing predictions have made fill in the values of dataset features that will produce True or False values that identify the value of absence or the presence of heart disease.

3. Results and Discussion

a. Data Analysis

This stage explores data analysis of the dataset used to find out the names of features and the correlation between their features that will be used to make predictions. Identify the features used in **Table 1** below.

Table 1. Dataset Heart UCI Repository

Feature	Description
Age	age in years
Sex	sex (1 = male; 0 = female)
cp	The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
trestbps	The person's resting blood pressure (mm Hg on admission to the hospital)
Chol	The person's cholesterol measurement in mg/dl
fbs	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
restecg	Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalach	The person's maximum heart rate achieved
exang	Exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
ca	The number of major vessels (0-3)
thal	A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
target (the predicted attribute)	Heart disease (0 = no, 1 = yes)

Missing data in data sets derived from EHR can be caused by lack of collection or lack of documentation, so it needs to be checked before making predictions. Data cleaning processes ensure that the data is correct, consistent, and usable in cleaning data by identifying errors or damages, correcting or deleting them, or processing data manually to prevent errors in determining prediction variables. The dataset used in this study used the open dataset from the UCI Repository **[12],** using the data.shape() function from Python can determine the number of records and features, namely 303 Records and 14 features in the heart attack dataset. In checking the missing data using the data.info () function, the result is as **Figure 5** below:

```
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Age                  303 non-null    int64
 1   Sex                  303 non-null    int64
 2   Chest Pain Type      303 non-null    int64
 3   Rest BP              303 non-null    int64
 4   Cholestrol           303 non-null    int64
 5   FBS                  303 non-null    int64
 6   RestECG              303 non-null    int64
 7   Max Heart Rate       303 non-null    int64
 8   Exer Angina          303 non-null    int64
 9   Prev Peak            303 non-null    float64
 10  Slope                303 non-null    int64
 11  No of Major Vessels  303 non-null    int64
 12  Thal Rate            303 non-null    int64
 13  Target               303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Figure 5. Missing data checking

From the results of the picture above, it can be concluded that there is no blank or missing data by knowing that each feature contains a record number of 303. This stage searches for features that affect the target variable. The results of determining the relationship between the feature and the target can be seen in **Figure 6** below.

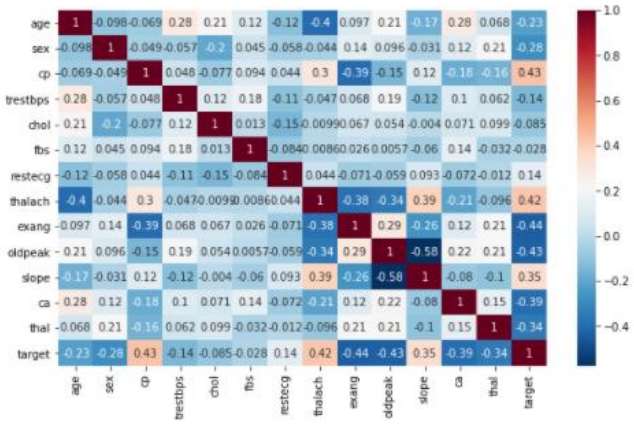


Figure 6. Correlation Matrix

Then chooses a feature with a more significant correlation because this feature will give more information. The minimum threshold is 0.2, which aims to make it easier to choose features that have a high correlation with the target variable of the diagnostic result. **Figure 7** below is the result of the threshold > 0.2 with the correlations[abs(correlations) function > 0.2.

```
target     1.000000
cp         0.433798
thalach    0.421741
slope      0.345877
age       -0.225439
sex       -0.280937
thal      -0.344029
ca        -0.391724
oldpeak   -0.430696
exang     -0.436757
Name: target, dtype: float64
```

Figure 7. Correlation Feature With 0.2 threshold

Based on the results of the selection of these features will be seen the effect of the parameters max heart rate (thalach), cholesterol (chol), resting blood pressure (trestbps) and ST depression (oldpeak) on age (Age). Below is **Figure 8**, which shows the effect of feature parameters on age.

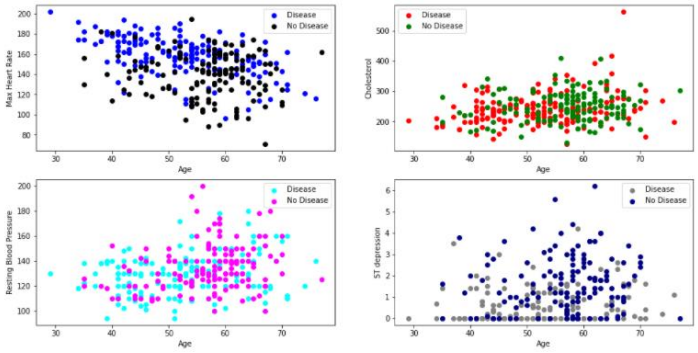


Figure 8. The effect of feature parameters on age

b. Data Pre-processing

Based on the results of the data analysis, it is concluded that the dataset features are interdepedataset features.

1) Splitting datasets into Training sets and Test sets: This stage divides the dataset into training data and Test data into 80% training data and 20% test data.

2) Feature Scaling: Some variables have very variable and random values in a raw dataset, so it is crucial to feature scaling the feature. It is a way to make numerical data in the dataset have the same range of values (scale). No more one data variable dominates the range of the other data variables. In carrying out this process by using the function in the sklearn library **[11],** namely with the function from sklearn.preprocessing import StandardScaler.

3) Model

The selection of models to produce outcome predictions using deep learning techniques, namely artificial neural networks (ANN). Keras is a Python library intended for devdataset's features deep learning models. It stores a numeric computing library and TensorFlow that can train neural network models when they are built.

a) Initialize Artificial Neural Networks **[13]:** This network has two layers, a hidden layer and an output layer. Hidden Layer will use the sigmoid function for activation. The Output Layer has only one node and is used for regression; the output node is the same as the input node. That is, its activation is f(x)=x. The activation function is A function that takes an input signal and generates an output signal but considers the threshold (Threshold).

b) Training Model: epochs are used for the number of times the dataset will pass through the network and each time it updates the weight. As the number of epochs increases, the network gets better and better at predicting targets in a set training set. In the selection of epochs, it must be adjusted to train the network properly, and it is hoped that there will not be a manageable mountable amount because it will result in overfitting. In this study, epochs= 30, 50, 80, and 100 were selected to predict the target in the training set. In compiling the ANN Model by conducting several experiments for epochs values and **Table 2** and **Table 3** below explains the values of the epochs filled in and the results.

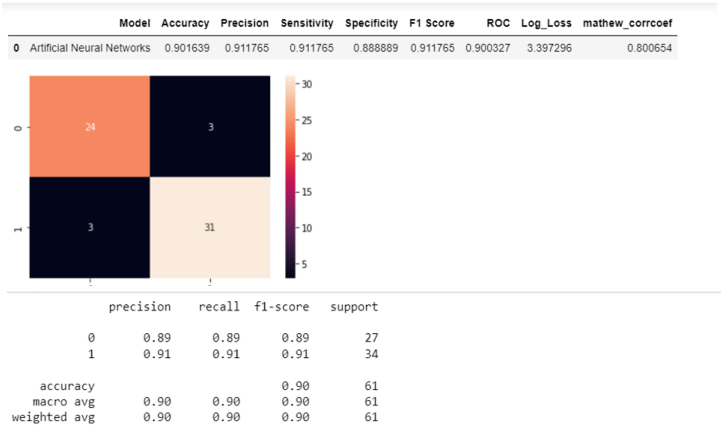
Table 2. Epoch Value Testing

Nilai Epochs	Confusion Matrix	Accuracy
30	[[21 6] [4 30]]	0.84
50	[[21 6] [3 31]]	0.85
80	[[22 5] [6 28]]	0.82
100	[[24 3] [3 31]]	0.90

Table 3. Testing hidden layer values with epoch = 100

Hidden Layer 1	Hidden layer 2	Hidden Layer 3	Accuracy
ReLu	ReLu	Tanh	0.79
Sigmoid	Sigmoid	Sigmoid	0.85
ReLu	ReLu	Sigmoid	0.87
ReLu	Sigmoid	Sigmoid	0.90

From the results of the study, it illustrates that the Artificial Neural Network model in predicting heart disease using a free dataset from the UCI Machine Learning Repository has an accuracy of 90%. In the compile stage of this neural network model, epoch = 100 is given and added as an attribute of validation data parameters. The data validation information resulted in a loss of: 0.2159 - accuracy: 0.9174 - val_loss: 0.3219 - val_accuracy: 0.901. The features of the heart disease dataset that have contributed from the results of data analysis are st_slope_upsloping, st_slope_flat, excercise_induced_angina, sex and cholesterol influence the target variables. Here is a **Figure 9** Confusion Matrix.



	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Artificial Neural Networks	0.901639	0.911765	0.911765	0.888889	0.911765	0.900327	3.397296	0.800654

	precision	recall	f1-score	support
0	0.89	0.89	0.89	27
1	0.91	0.91	0.91	34
accuracy			0.90	61
macro avg	0.90	0.90	0.90	61
weighted avg	0.90	0.90	0.90	61

Figure 9. Confusion Matrix

Based on the results of machine learning, it can be implemented using a flask application **[14], [15]** that is integrated with a python program so that an application web interface can be created for simulating heart disease detection such as **Figure 10** of the heart disease detection web interfaces.

Figure 10. Interface Web *Deteksi Penyakit Jantung*

4. Conclusion

Based on the results of research that has been carried out to produce outcome predictions in predicting heart disease using the Artificial Neural Network algorithm, it produces accuracy = 90% and is a comparison material for accuracy results compared to the Random Forest algorithm, namely accuracy = 85%. There are 5 features that contribute to influencing the target variable, namely st_slope_upsloping, st_slope_flat, excercise_induced_angina, sex and cholesterol. Testing of the model made by filling in the feature values and resulting in a

prediction of 1 = hearth disease 0 = No Hearth disease. This neural network technique uses 3 (three) hidden layers with the value of epochs = 100.

Future research will predict more than one dataset regarding heart disease (Heart failure disease) apart from the UCI Machine Learning Repository in order to have a comparison to accuracy values through the selection of certain algorithms and the existence of exploration activities for data analysis (Exploration Data Analyst) which needs to be improved to find out the correlation of the features in the dataset and further handling of preprocessing data and activities The research that has been done can make a good contribution in the field of health data science and as material for subsequent studies.

References

[1] United Nations Department of Economic and Social Affairs, "World's population increasingly urban with more than half living in urban areas," Report on World Urbanization Prospects, Jul. 2014, 2014.
[2] G. Perkasa, "Penyakit Jantung Penyebab Kematian Utama di Dunia," Kompas.com, 2020.
[3] https://sehatnegeriku.kemkes.go.id
[4] A. Riani, Y. Susianto, N. Rahman, dan U. D. Ali, "Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Mengunakan Metode Naive Bayes Data Mining Implementation to Predict Heart Disease using Naive Bayes Method," vol. 1, no. 01, hal. 25–34, 2019, doi: 10.35970/jinita.v1i01.64.
[5] S. Komputer dkk, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," vol. 4, no. 1, hal. 84–88, 2020.
[6] A. Alhamad, A. I. S. Azis, B. Santoso, dan S. Taliki, "Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote," vol. 5, no. 3, hal. 352–360, 2019.
[7] J. Moolayil, Learn Keras for Deep Neural Networks: A Fast-Track Approach to Modern Deep Learning with Python. 2019.
[8] J. Eckroth, Python Artificial Intelligence Projects for Beginners: Get up and Running with Artificial Intelligence Using 8 Smart and Exciting AI Applications. Packt Publishing, 2018.
[9] H. F. Hananta, I. Y., & Muhammad, Dietisien Deteksi Dini & Pencegahan 7 Penyakit Penyebab Mati Muda. 2011.
[10] N. K. Manaswi, Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition with TensorFlow and Keras. 2018.
[11] L. W. Scikit-learn dan M. L. W. Scikit-, Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media, 2017.
[12] U. of C. I. M. L. Repository, "Heart Disease Dataset." available: https://archive.ics.uci.edu/ml/datasets/heart+Disease.
[13] A. Çelik dkk, FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS, vol. 1, no. 1. 2018.
[14] M. Grinberg, Flask Web Development: Developing Web Applications with Python. 2014.
[15] I. Maia, Building Web Applications with Flask. 2014.