A Parameters-based Heart Disease Prediction Model

Jakub Gurgul[1], Adam Paździerz[1] and Dawid Wydra[1]

[1]*Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska 23, 44-100 Gliwice, Poland*

Abstract

Heart diseases lead to disorders of the entire circulatory system. The risk of their occurrence depends on age. They are caused by poor diet, obesity, excessive stress, smoking, lack of physical activity or genetic predisposition. Heart diseases and circulatory system are the biggest threat to life. The development of technology, the use of appropriate algorithms along with relevant data allows us to increase the chances of detecting more quickly whether a patient has heart disease using given patient parameters. This work concerns the creation of a classification model for the risk of heart disease. It works based on sex, age, blood pressure, cholesterol etc. The model works with data from a data set *Heart Failure Prediction Dataset* [1] downloaded from kaggle. The best model was selected after a series of comparative tests of different classification systems.

Keywords

Heart disease, Prediction model, Classification algorithms, Heart failure prediction, Python

1. Introduction

Artificial intelligence plays an increasingly important role in the operation of information systems [2, 3]. The computing power of modern computers allows to perform large numbers of calculations, thanks to which heuristic algorithms have been developed, biologically inspiring [4, 5], allowing to optimize many processes. Important from the point of view of reducing energy consumption is the use of heuristic algorithms to optimize the operation of [6] transformers. Another group of artificial intelligence algorithms are methods based on the use of fuzzy sets [7]. these methods are widely used, in particular in combination with the IoT [8, 9, 10, 11] infrastructure. Speaking of artificial intelligence, one cannot fail to mention methods based on the use of artificial neural networks [12], which find numerous applications in machine learning [13, 14, 15, 16, 17, 18] in automatic detection of some interesting features of the examined objects [19]. They are also widely used in areas related to the care of health [20, 21].

Heart diseases lead to disorders of the entire circulatory system. The risk of their occurrence depends on age. They are caused by poor diet, obesity, excessive stress, smoking, lack of physical activity or genetic predisposition. Heart diseases and circulatory system are the biggest threat to life. The development of technology, the use of appropriate algorithms along with relevant data allows us to increase the chances of detecting more quickly whether a patient has heart disease using given patient parameters.

2. Assumptions for algorithms

Each of the algorithms should be prepared to meet the following criteria:

1. Prepared according to the mathematical description of the algorithm;
2. Optimized for the performance on our data set;
3. Categorizes the case from the test set, determines whether the case has a heart disease or not;
4. Minimize the number of false negatives cases;
5. Returns an array containing information about:
    - Whether the data supplied to the function was normalized;
    - The number of correct predictions for each of the cases;
    - The number of incorrect predictions for each of the cases;
    - Length of the computed set;
    - Confusion matrix in the following schema: [[TP, FN], [FP, TN]]
6. Based on the selected algorithm, creation a client application for the assessment of the risk of the heart disease.

3. Data set

3.1. Description of the columns

The set consists of 918 rows and 12 columns (11 features), 3 of them are categorical, 3 are factorial and 5 are numerical. A detailed description is provided below:

1. **Age** - Age of the examined person in years;

ICYRIME 2022: International Conference of Yearly Reports on Informatics, Mathematics, and Engineering. Catania, August 26-29, 2022

jakugur@polsl.pl (J. Gurgul); adampaz@polsl.pl (A. Paździerz); dawiwyd801@polsl.pl (D. Wydra)
0000-0002-3159-234X (J. Gurgul); 0000-0003-0853-9695 (D. Wydra)

2. **Sex** - Sex of the examined person, originally the table contains markings:
   - M (eng. *Male*) - changed to 1;
   - F (eng. *Female*) - changed to 0;
3. **ChestPainType** - Type of chest pain, originally the table contains markings:
   - TA (eng. *Typical Angina*) - changed to 0;
   - ATA (eng. *Atypical Angina*) - changed to 1;
   - NAP (eng. *Non-Anginal Pain*) - changed to 2;
   - ASY (eng. *Asymptomatic*) - changed to 3;
4. **RestingBP** - (eng.*Resting Blood Pressure*) - expressed in mm/Hg;
5. **Cholesterol** - Serum cholesterol expressed in mmol/dl;
6. **FastingBS** - (eng. *Fasting Blood Sugar*) - expressed in mg/dl:
   - 1 - if the test level was above 120 mg/dl;
   - 0 - if it was below the mentioned level;
7. **RestingECG** - (eng. *Resting Electrocardiogram Results*) - original table contains markings:
   - Normal - changed to 0;
   - ST (eng. *ST-T wave abnormality*) - changed to 1;
   - LVH (eng. *Non-Anginal Pain*) - changed to 2;
8. **MaxHR** - (eng. *Maximum Heart Rate*) - highest heart rate recorded, a numerical value between 60 and 202 beats per minute;
9. **Exercise Angina**:
   - Y - changed to 1;
   - N - changed to 0;
10. **Oldpeak** - ST segment depression caused by physical action, numerical value;
11. **ST_Slope** - (eng. *The slope of the ST section*), originally the table contains the indications:
    - Up (eng. *Upsloping* - ST section elevation) - changed to 0;
    - Flat (eng. *Flat* - flat ST section) - changed to 1;
    - Down (eng. *Downsloping* - lowering of the ST section) - changed to 2;
12. **HeartDisease** - Diagnosis of the patient:
    - 1 - Heart disease;
    - 0 - Normal results;

3.2. Statistical data for the primary data set

The data set we worked on was created by combining five other data sets containing data on heart disease. By merging and generalizing them, one of the largest publicly available data sets of its kind was created.

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations
Duplicated: 272 observations
Final data set: 918 observations

There are some inaccuracies in the reported resting blood pressure value, which is 0 [*mm/Hg*]. This may suggest that some data have been misspelled. In the case of blood pressure, the value of 0 occurs in only one case - for a 55-year-old man, according to an article published in the Toronto Metropolitan University Pressbooks the normal average value for his age group is between 110 and 145 [*mm/Hg*], so we omitted this result from our work. A value equal to 0 [*mmol/dl*] is also found for Cholesterol, but here it appears 172 times, which rather rules out confusion for us, it is more likely that its level was indeterminate. Another noticeable fact is the age of the respondents. The average is 53 years old, there are only 80 people under 40 years old (8.72% of the respondents), this shows that the risk of heart disease symptom increases dramatically with increasing age.

Table 1
Table containing statistical data for the input dataset

Column Name	Mean	σ	Min	Max
Age	53.51	9.43	28.0	77.0
RestingBP	132.39	18.51	0.0	200.0
Cholesterol	198.79	109.38	0.0	603.0
MaxHR	136.80	25.46	60.0	202.0
Oldpeak	0.88	1.06	-2.6	6.2

4. Exploratory analysis

The pair plot on figure 1 shows that there are not many outliers (now clearly a 55-year-old man with an outlier for resting blood pressure). Green points indicate cases with confirmed hearing diseases. The pairs plot shows that there are not many outliers (now clearly a 55-year-old man with an outlier for resting blood pressure).

The correlation heat map on figure 2 shows a negative relationship between the presence of heart disease

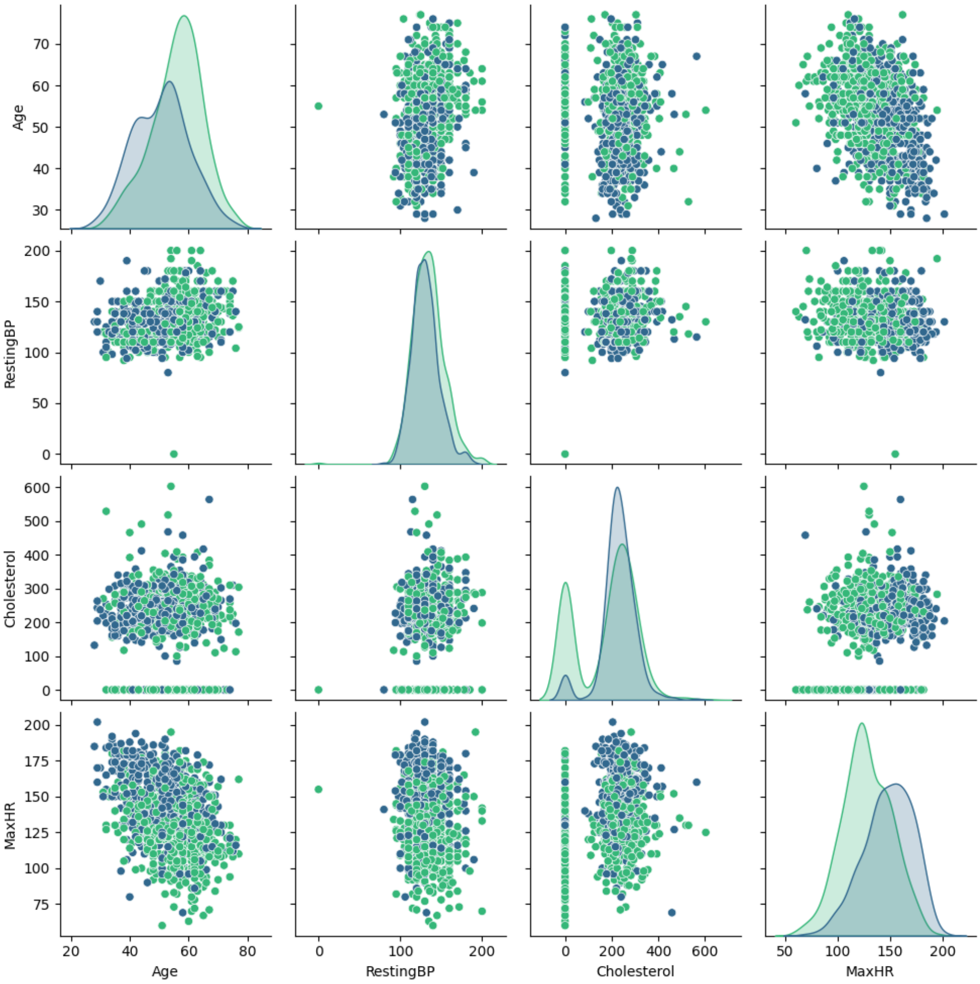


Figure 1: Pair plot of numeric data in our set

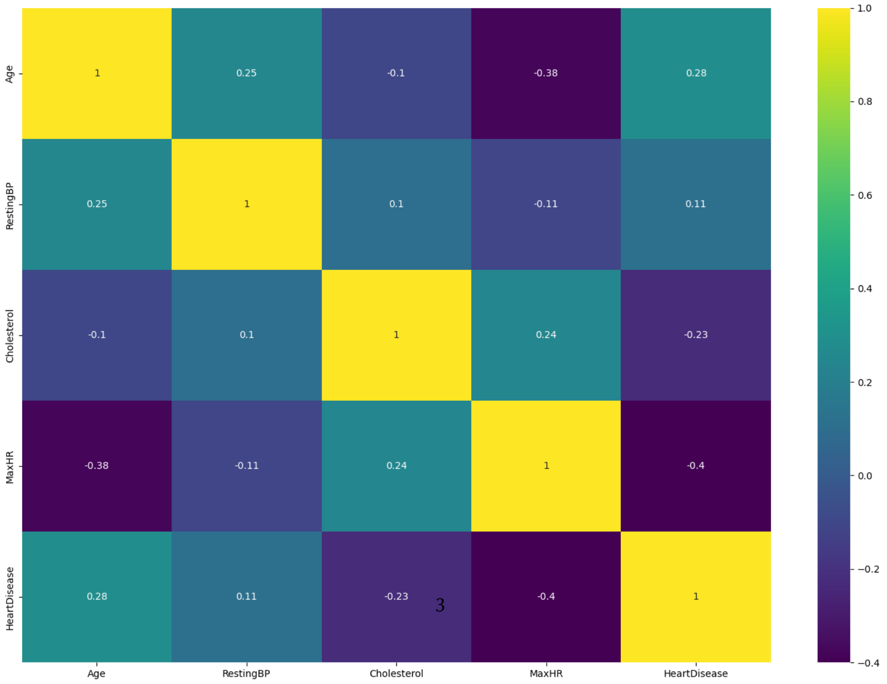


Figure 2: Correlation heat map for the set

and maximum heart rate, and a similar relationship is found between heart rate and age. On the other hand, we can observe a positive correlation between resting blood pressure and age, between age and the occurrence of heart disease (we have already noticed this feature during the statistical analysis of the data), the last clear positive relationship is shown by the value of cholesterol with the value of maximum heart rate.

The pie chart on figure 3 shows the overall distribution of heart disease. It can be seen here that as many as one in five subjects was eventually diagnosed. When creating the test sets, we tried to get as close as possible to the percentage distribution of cases so that our test set would match the total as closely as possible, we tried to keep a difference of 1-2 percentage points from the total when generating the test set.

The box plot on figure 4 shows us the distribution of heart disease incidence by gender and age, here we see that many more early positive cases have been recorded among men, in ladies the threshold at which the chance increases is much higher.

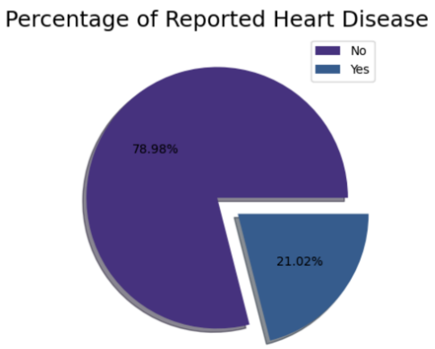
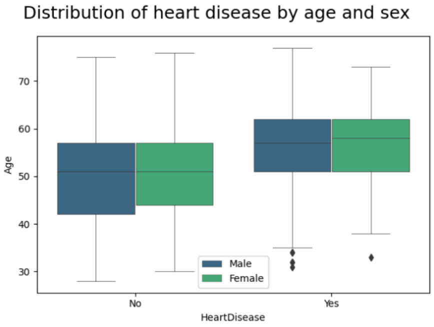


Figure 3: Pie chart showing percentage of positive and negative diagnoses for heart disease distribution



https://www.overleaf.com/project/62e0506668ca31480b86be3e

Figure 4: Box plot showing the distribution of heart disease by sex and age of study subjects

5. Preparing data for use

The first action taken to process the data was to check the completeness of the collection. It contained 918 rows, each of which contained complete data, which was checked with the appropriate function.

The next step was to convert the factored values Y/N, M/F to 1 and 0, respectively. Then categorical values were converted to numeric using our own function *category_to_numeric()*. In order to conduct a comparison of how the algorithms perform on the normalized and non-normalized data set, a copy of the data set was written (non-standardised - *heart_nnorm.csv*). The columns that can be normalized are Age, RestingBP, Cholesterol, MaxHR and Oldpeak. The operation was performed using a custom function - *normalize()* using the min-max formula. After the data was normalized, it was written to a file *heart_norm.csv* for further use.

6. Classification algorithms

It was decided to create three algorithms from scratch:

- **KNN** - an algorithm that checks the nearest neighbors of a case and determines what class its neighbors are. On this basis, the algorithm deduces what class is the searched case. The experimental method was used to determine the hyperparameter k, its value was set to the 5 nearest neighbors.
- **Classification using soft sets** - the data set was adapted for use, then using our own functions we checked which category best fits the processed case.
- **Naive Bayes classifier** - it is a simple probabilistic model based on the assumption of mutual independence of predictors. Vectors of mean values and standard deviations were calculated. On this basis a function was built to predict the class of cases.

7. Quality assessment of classifiers and results

The classifiers were evaluated based on the execution time of the algorithms on the test sample (184 instances - 20 percent of the original set, prepared to best reflect the gender distributions during partitioning and the percentage of positive instances in the original set), as well as on the ordered confusion matrix shown in table 2

In addition, to make the results obtained more readable, a standard set of derivatives from the confusion matrix was used:

Table 2
Confusion matrix structure

Population		Predicted condition	
		Positive classification	Negative classification
Actual condition	Positive	True Positive TP	False Negative FN
	Negative	False Positive FP	True Negative TN

1. Accuracy (ACC);

$$ACC = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

2. Sensitivity recall (SEN):

$$SEN = \frac{TP}{(TP + FN)} \quad (2)$$

3. Specificity (SPE):

$$SPE = \frac{TN}{(TN + FP)} \quad (3)$$

4. Precision (PRE):

$$PRE = \frac{TP}{(TP + FP)} \quad (4)$$

5. F-score (F1), a measure of the accuracy of the test;

$$F1 = \frac{(2 * PRE * SEN)}{(PRE + SEN)} \quad (5)$$

To make the results obtained on the normalized and unnormalized set using a single algorithm as good as possible for comparison we presented them in pairs on one page.

7.1. KNN – normalized data

Algorithm execution time: **18 sec, 10.10 elem/s**

Table 3
Confusion matrix for KNN, normalized data

184	Predicted condition	
Actual condition	89 TP	16 FN
	9 FP	70 TN

In each of the measured measures, we see quite high percentages, none of the determinants differs significantly from the others. A high PRE value means that the classifier has done well to identify true positive cases. Additionally, which was extremely important for us, the algorithm classified only 16 cases as false negative. In the case of KNN, we used the Minkowski metric with m = 2.

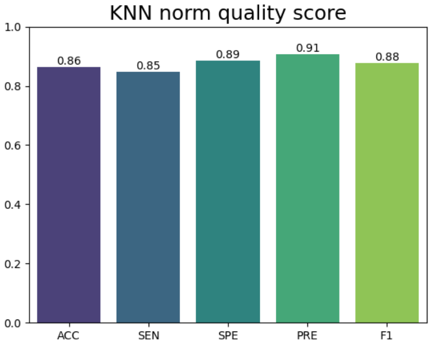


Figure 5: Plot of quality measures for KNN on normalized data

Table 4
Confusion matrix for KNN, non-normalized data

184	Predicted condition	
Actual condition	79 TP	30 FN
	28 FP	47 TN

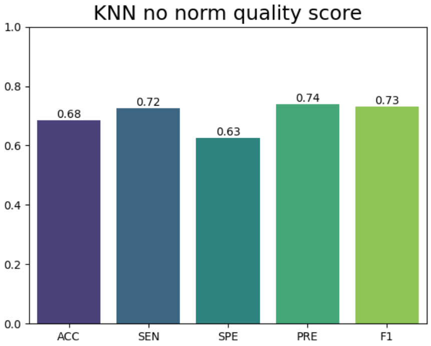


Figure 6: Plot of quality measures for KNN, non-normalized data

7.2. KNN - non-normalized data

Algorithm execution time: **19 sec, 9.37 elem/s**
Compared to the version running on normalized data, we can see much larger differences in quality measures. The overall performance is much worse. Such a difference is caused by the lack of normalization, to which this algorithm is very sensitive. This is because features with larger values will completely cover those with small values. For example, the feature *Age* (operating on the range 28.0 - 77.0) will have a much higher weight than *FastingBS*, which ranges from 0.0 to 1.0.

7.3. Soft sets, normalized data

Algorithm execution time: **23 sec, 7.76 elem/s**

Table 5
Confusion matrix for soft sets on normalized data

184	Predicted condition	
Actual condition	96 TP	41 FN
	5 FP	42 TN

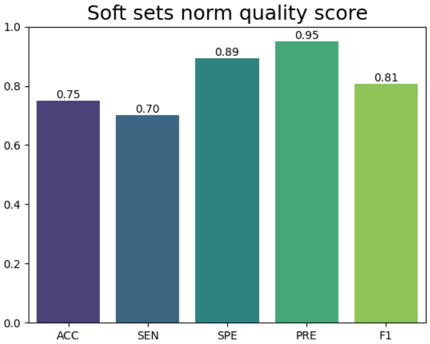


Figure 7: Plot of quality measures for soft sets on normalized data

The time is much longer than the KNN algorithm, the results are worse than KNN. The SPE and PRE indices stand out the most. The PRE index is at the highest level among all the algorithms, which tells us that this version of the algorithm did very well with true positive cases, as it classified as many as 96 of them. However, the number of 41 false negative cases is unacceptable.

7.4. Soft sets, non-normalized data

Algorithm execution time: **18 sec, 9.72 elem/s**

Table 6
Confusion matrix for soft sets on non-normalized data

184	Predicted condition	
Actual condition	84 TP	20 FN
	23 FP	57 TN

Compared to the normalized version, the results again show large differences between each other. The overall trend is worse in all indicators. Of note is the rather low execution time of the algorithm.

7.5. Naive Bayes classifier, normalized data

Both Bayes classifiers do not have a given time because their execution is almost instantaneous. This is due to the lack of need to calculate the mean value and standard deviation each time, which we can calculate only once and very quickly using the built-in functions of the *pandas* package. For the normalized data we see very similar results in each of the indicators but they are slightly lower than those in KNN. None of them deviate from the rest.

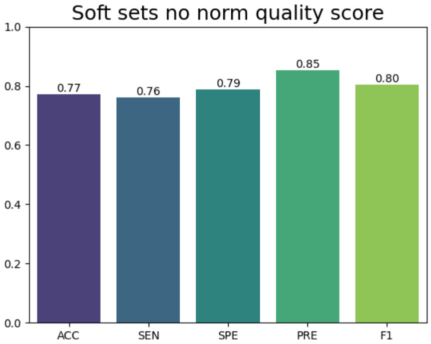


Figure 8: Plot of quality measures for soft sets on non-normalized data

Table 7
Confusion matrix for naive Bayes classifier on normalized data

184	Predicted condition	
Actual condition	79 TP	16 FN
	19 FP	70 TN

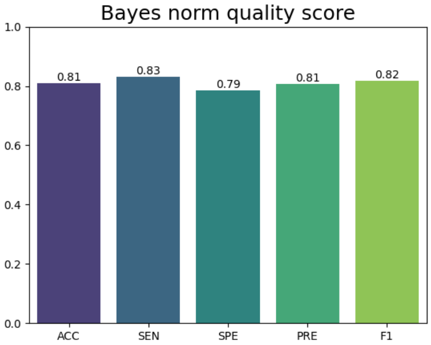


Figure 9: Quality measure plot for naive Bayes classifier on normalized data

7.6. Naive Bayes classifier, non-normalized data

For non-normalized data, the scores obtained by this algorithm are very low. This is similarly to the KNN case due to incorrectly set weights on the learning set. Due to

Table 8
Confusion matrix for naive Bayes classifier on non-normalized data

184	Predicted condition	
Actual condition	16 TP	62 FN
	98 FP	8 TN

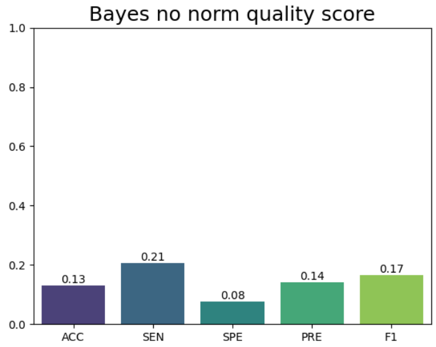


Figure 10: Quality measure plot for naive Bayes classifier on non-normalized data

such low scores, this version of the naive algorithm will not be considered further.

8. Selecting the best classifier

Each of the classifiers was used 50 times (the same normalized and non-normalized test set was always used during a single trial), so that the results we obtained were as close as possible to their actual accuracy. In our opinion, the best among the tested algorithms was KNN on normalized data. It features very high quality scores and very high closeness. It found slightly fewer TP cases than the algorithm operating on soft sets, but the number of FN cases is significantly lower which is very important for disease classification. I would say it is better to make a false positive diagnosis and refer for further testing than to make a false negative diagnosis. Ultimately this argument convinced us to choose KNN algorithm as the best algorithm to predict the occurrence of heart disease in our database. It is also very important that the data on which the model would work in the future should be normalized, without this the effectiveness of the algorithm will fall dramatically.

9. Client app

The application was written based on the *flask* library, due to its easy and light-to-use structure. The application is created on the basis of the website using the *Bootstrap* package in order to facilitate the management of the appearance. It has a number of fields for the output that correspond to a column from the data set. After their completion, the classification for a given unknown case is carried out with the use of the KNN algorithm on the normalized data. After the program determines the class, it is returned to the user in the form of a text block with the confidence of the result, calculated from the n nearest neighbors determined for the case.

10. Conclusion and future work

In summary, the design assumptions were successfully met, the proven algorithms were successful on the dataset. To improve their performance, further tests should be performed with more algorithms, or with two or more algorithms simultaneously. An additional factor to increase the usefulness of our work would be to expand the set with more features and cases. With more variety and quantity, the algorithm could learn better to identify classes and calculate the risk of heart disease.

References

[1] FEDESORIANO, Heart failure prediction dataset, 2021. URL: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

[2] M. A. Sanchez, O. Castillo, J. R. Castro, Generalized type-2 fuzzy systems for controlling a mobile robot and a performance comparison with interval type-2 and type-1 fuzzy systems, Expert Systems with Applications 42 (2015) 5904–5914.

[3] Q.-b. Zhang, P. Wang, Z.-h. Chen, An improved particle filter for mobile robot localization based on particle swarm optimization, Expert Systems with Applications 135 (2019) 181–193.

[4] T. Qiu, B. Li, X. Zhou, H. Song, I. Lee, J. Lloret, A novel shortcut addition algorithm with particle swarm for multisink internet of things, IEEE Transactions on Industrial Informatics 16 (2019) 3566–3577.

[5] Y. Zhang, S. Cheng, Y. Shi, D.-w. Gong, X. Zhao, Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm, Expert Systems with Applications 137 (2019) 46–58.

[6] M. Woźniak, A. Sikora, A. Zielonka, K. Kaur, M. S. Hossain, M. Shorfuzzaman, Heuristic optimization of multipulse rectifier for reduced energy consumption, IEEE Transactions on Industrial Informatics 18 (2021) 5515–5526.

[7] Y. Sun, H. Qiang, J. Xu, G. Lin, Internet of things-based online condition monitor and improved adaptive fuzzy control for a medium-low-speed maglev

train system, IEEE Transactions on Industrial Informatics 16 (2020) 2629–2639. doi:10.1109/TII.2019.2938145.

[8] M. Woźniak, A. Zielonka, A. Sikora, M. J. Piran, A. Alamri, 6g-enabled iot home environment control using fuzzy rules, IEEE Internet of Things Journal 8 (2020) 5442–5452.

[9] V. Ponzi, S. Russo, V. Bianco, C. Napoli, A. Wajda, Psychoeducative social robots for an healthier lifestyle using artificial intelligence: a case-study, in: CEUR Workshop Proceedings, volume 3118, 2021, p. 26 – 33.

[10] M. Woźniak, A. Zielonka, A. Sikora, Driving support by type-2 fuzzy logic control model, Expert Systems with Applications 207 (2022) 117798.

[11] G. Capizzi, C. Napoli, S. Russo, M. Woźniak, Lessening stress and anxiety-related behaviors by means of ai-driven drones for aromatherapy, in: CEUR Workshop Proceedings, volume 2594, 2020, p. 7 – 12.

[12] V. S. Dhaka, S. V. Meena, G. Rani, D. Sinwar, M. F. Ijaz, M. Woźniak, A survey of deep convolutional neural networks applied for prediction of plant leaf diseases, Sensors 21 (2021) 4749.

[13] A. T. Özdemir, B. Barshan, Detecting falls with wearable sensors using machine learning techniques, Sensors 14 (2014) 10691–10708.

[14] R. Aureli, N. Brandizzi, G. D. Magistris, R. Brociek, A customized approach to anomalies detection by using autoencoders, in: CEUR Workshop Proceedings, volume 3092, 2021, p. 53 – 59.

[15] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, Machine learning in agriculture: A review, Sensors 18 (2018) 2674.

[16] G. Lo Sciuto, G. Susi, G. Cammarata, G. Capizzi, A spiking neural network-based model for anaerobic digestion process, in: 2016 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), IEEE, 2016, pp. 996–1003.

[17] F. Bonanno, G. Capizzi, G. Lo Sciuto, C. Napoli, G. Pappalardo, E. Tramontana, A cascade neural network architecture investigating surface plasmon polaritons propagation for thin metals in openmp, in: Artificial Intelligence and Soft Computing: 13th International Conference, ICAISC 2014, Zakopane, Poland, June 1-5, 2014, Proceedings, Part I 13, Springer, 2014, pp. 22–33.

[18] F. Bonanno, G. Capizzi, G. Lo Sciuto, C. Napoli, Wavelet recurrent neural network with semiparametric input data preprocessing for micro-wind power forecasting in integrated generation systems, in: 2015 International Conference on Clean Electrical Power (ICCEP), IEEE, 2015, pp. 602–609.

[19] O. Dehzangi, M. Taherisadr, R. ChangalVala, Imu-based gait recognition using convolutional neural networks and multi-sensor fusion, Sensors 17 (2017) 2735.

[20] M. Woźniak, M. Wieczorek, J. Siłka, D. Połap, Body pose prediction based on motion sensor data and recurrent neural network, IEEE Transactions on Industrial Informatics 17 (2020) 2101–2111.

[21] S. I. Illari, S. Russo, R. Avanzato, C. Napoli, A cloud-oriented architecture for the remote assessment and follow-up of hospitalized patients, in: CEUR Workshop Proceedings, volume 2694, 2020, p. 29 – 35.