Feature Selection and Optimisation with Neural Network for Depression Level classification

Ruimin Chu

Research School of Computer Science, Australian National University

u5924220@anu.edu.au

Abstract. Irrelevant and redundant features can be one of the causes of performance degradation in classification problem. Therefore, feature selection method, which is the process of choosing a subset of features that contribute most to the machine learning algorithm can be implemented to solve the problem. In this paper, a few techniques to determine the significance of the features are combined with a neural network to classify various depression level with the dataset of observers' physiological signals. The implemented feature selection techniques include magnitude measure, distinctiveness analysis, recursive feature elimination, minimum redundancy maximum relevance as well as genetic algorithm. Furthermore, as it has been recognised as a powerful tool for optimisation problem, genetic algorithm will be applied to find the optimal hyperparameters for the neural network model.

Keywords: Neural Network, Feature Selection, Magnitude Measure, Distinctiveness Analysis, Minimum Redundancy Maximum Relevance, Genetic Algorithm

1 Introduction

Depression is one of the major mental disorders and one of the leading causes of disability. At worst case, depression may lead to suicide. Therefore, early detection and accurate diagnosis of depression can help prevent or promote remission from the disease. In recent years, a large amount of research has been conducted on emotion recognition from physiological information and review of Galvanic Skin Response (GSR) [1] and Pupil Dilation (PD) [2] confirmed that these signals could be considered as indicators of depression.

A dataset of physiological signals for observers' responses to videos of people expressing depression of various levels of depression was collected by Zhu et al. [3]. The dataset contains 192 observations and further feature extraction has been conducted on physiological signals, which increased the size of the statistics feature set to 85. A neural network (NN) based classification model is built in the paper and it performs reasonably well on the dataset after training.

However, some irrelevant and redundant features have been found to cause performance degradation in depression level recognition. Feature selection is the process of finding a subset of features that contribute most to building a better predictive model. In this paper, some feature selection techniques are applied with the NN model to remove redundant and insignificant features and hence improve the performance. Magnitude measure [4] is a technique to measure the contribution of input features to outputs and the technique can be used to determine the significance of each feature. Distinctiveness analysis [5] is an approach to examine the functional differences between neurons. The recursive feature elimination (RFE) method is a feature selection method that removes the least important feature at every iteration until the desired number of features is reached. The result of two former techniques can be used as importance score in the RFE method to remove the most insignificant or redundant input neurons. Minimum Redundancy Maximum Relevance (mRMR) is a filter-based feature selection method and should be implemented before building the NN model.

Genetic Algorithm (GA) is a heuristic search algorithm inspired by the theory of evolution by natural evolution that belongs to the family of Evolutionary Algorithms (EA). [6] The algorithm has been recognised as a powerful tool for optimisation problem and therefore it will be applied to feature selection task and hyperparameter optimisation as well. Moreover, as the size of the feature set within the dataset is so large, the result from magnitude measure is incorporated here to create a reasonable initial population.

Finally, the performance of the NN model, when combined with each feature selection technique, is compared in relation to the accuracy of this model on a classification task. This allows the impact of each of these techniques on a classification problem to be measured.

2 Method

2.1 Data set

The dataset consists of the physiological data from 12 participants, who watched stimuli videos of people with various levels of depression [3]. The stimuli videos are from 2014 Audio-Visual Emotion Challenge (AVEC 2014) dataset [7]

and the video set is divided into four depression categories based on the depression scores. The measured data are Galvanic Skin Response (GSR), Skin Temperature (ST) and Pupillary Dilation (PD). Data pre-processing and feature extraction were then performed on the raw data, which produced 85 features, including 23 GSR features, 39 PD features, and 23 ST features [3]. The task is to use those features to train a feed forward neural network as classification model to predict the depression level of the people in the videos. Feature selection will also be applied to the baseline model.

2.2 Classification with Neural Network

The baseline model is a feed forward neural network (FFNN) classification model with one input layer, one hidden layer and one output layer. It takes 85 statistics features as inputs and the output layer has four neurons which represent the four depression levels from None to Severe. Sigmoid and softmax activation functions are both great candidates for multi-class classification problem. Based on the results of some testings, the overall performance of the sigmoid activation function on this task is slightly better than softmax, so the former function will be used for the baseline model as well as the models extended with techniques. Next, the number of hidden neurons is determined by testing the model with different amounts of neurons, ranging from 20 to 200, with an increment of 10. The results show that 40 is the optimal size for the task. FFNN is trained with the Adam optimizer with a learning rate of 0.001.

2.3 Feature selection

Feature selection methods can be roughly categorised into filter, wrapper, embedded and hybrid approaches [8]. Techniques based on the first two approaches will be focused on in this assignment. Filter method selects a subset of features before implementing the learning algorithm [9] while wrapper approach uses a machine learning algorithm to evaluate a subset of features based on the performance of the given algorithm [10]. Generally, filter-based methods take less computational time because the model does not need to be trained at each iteration, but wrapper-based methods can reach higher classification accuracy as the model itself is part of the evaluation criteria.

Magnitude Measure

Magnitude measure introduced by Gedeon [4] is an extension of the measure for the proportional contribution proposed by Garson [11]. It measures the contribution of input features to outputs. The contribution that input neuron i makes to output neuron j can be computed using the formula [4] below:

$$Q_{ij} = \sum_{j=1}^{nh} P_{ij} \times P_{jk}$$

Where P_{ij} is the contribution of an input neuron i to a hidden neuron j and P_{jk} is the contribution of hidden neuron j to an output neuron k. Those two values can be computed with the weights between layers and the formulae are shown below:

$$P_{ij} = \frac{|W_{ij}|}{\sum_{p=1}^{ni} |W_{pj}|}$$

$$P_{jk} = \frac{|W_{jk}|}{\sum_{r=1}^{nj} |W_{rk}|}$$

Since the baseline model has four output neurons, the average value of contribution to each of the four neurons will be used to determine the significance of each input neuron. The advantage of this method is that it avoids the issue of the cancellation of weights with opposite signs. However, it also brings the problem that the sign of the proportion is lost.

Distinctiveness Analysis

Distinctiveness analysis [5] is the technique to examine the functional difference between neurons. The technique was initially used to determine the similarity between hidden neurons and then be extended to find out the functional differences between inputs with the weight matrix between input layer and hidden layer [4]. Thus, each vector represents the weight between an input to all hidden neurons and each vector will be normalized between 0 and 1, then 0.5 will be deducted. Then the vector angle θ between input neuron i and j can be computed using the formula below

$$\theta = \operatorname{acos}\left(\frac{\langle v_i, v_j \rangle}{\|v_i\| \|v_j\|}\right)$$

The similarity between two neurons is inversely proportional to the size of the vector angle. The vector angles will then be ranked based on their values and one of the neurons in the pair of neurons with the smallest vector angle will be removed as they are considered sufficiently similar.

Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a feature selection method that removes the least important feature at every iteration until the desired number of features is reached. The technique starts by constructing a model on the entire feature set and computing an importance score for each feature, which is usually the impact of that feature on the accuracy of the model [12]. For this assignment, the results obtained with magnitude measurement and distinctiveness analysis techniques which have been introduced before will be used as importance scores here. The feature with the lowest contribution to output neurons or most similar to another feature is eliminated at each iteration. Since it is hard to determine an optimal

size of a feature subset without prior knowledge of those physiological signals, one or more inputs will be removed at each iteration until the feature set becomes empty.

Minimum Redundancy Maximum Relevance

Minimum Redundancy Maximum Relevance (mRMR) is a filter-based feature selection method. As its name indicates, the algorithm selects a subset of features which have the highest correlation to the target variable while maintaining the least correlation among themselves. As most of the variables in the dataset are continuous, the F-test Correlation Difference (FCD) and F-test Correlation Quotient (FCQ) schemes are used here to search for the optimal features [13].

$$\text{FCD} : max_{i \epsilon \Omega_S}\{F(i,h) - \frac{1}{|S|}\sum_{j \in S}|c(i,j)|\}$$

$$\text{FCQ} : max_{i \epsilon \Omega_S}\{F(i,h) / \frac{1}{|S|}\sum_{j \in S}|c(i,j)|\}$$

Where $F(i,h)$ is the F-statistic between the feature i and target variable h and $c(i,j)$ is the correlation coefficient between feature i and feature j. S is the set of selected features and $\Omega_S = \Omega$ - S which is the feature set except those already selected

The simple heuristic algorithm introduced in [13] for mRMR optimization problem is applied to obtain the near optimal solution while reducing the number of searches. First, the feature with the highest $F(i,h)$ will be put into the selected feature set. Then, among the remaining features, the one that maximize the chosen criteria will be added into the set.

Genetic Algorithm

Genetic Algorithm (GA) is a heuristic search algorithm inspired by the theory of evolution by natural evolution that belongs to the family of Evolutionary Algorithms (EA). [6] GA generally outperforms traditional feature selection techniques for its large search space and the ability to manage parallelism. However, GA also has drawbacks that it is computationally expensive and takes a longer time to converge. [14] For the feature selection problems, GA will have a population of chromosomes that represent the selected features using binary encoding as candidate solutions. Since the ultimate goal here is to increase the performance of the models, the classification accuracy on the test set is used as the fitness score.

The elitist roulette wheel selection method is a combination of the roulette wheel selection and the elitism and is adapted here to choose the fittest parents for the next generation [15]. In roulette wheel selection method, the probability of an individual is chosen is related to the ratio of the fitness to the sum of fitness values for all the members in the population. Since there is no guarantee that the best solution will be kept using the roulette wheel selection method, elitism will be added that the best two solutions will be kept and the rest of parents will be selected based on the fitness values to preserve the diversity of the population. After that, the crossover operation and mutation operation will be applied to generate the children, forming the rest of the generation.

The detailed process can be found in the flowchart and parameters table below. During some experiments, it was observed that the generation would sometimes proceed without any improvements on the classification accuracy. Therefore, it was considered reasonable to cut off the process if no improvement was found on the testing accuracy for five successive generations.

Figure 1. Genetic Algorithm Process

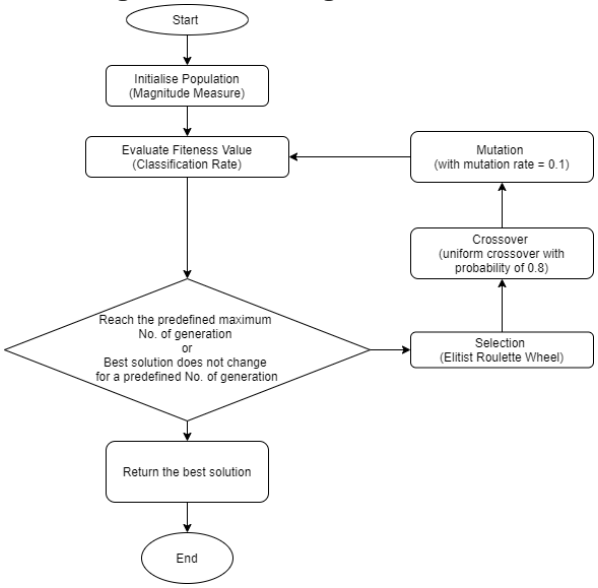


Table 1. GA parameters

GA Parameter	Value
Population Size	80
Parent Size	12
Selection method	Elitist Roulette wheel
Crossover method	Uniform crossover
Crossover probability	0.8
Mutation rate	0.1

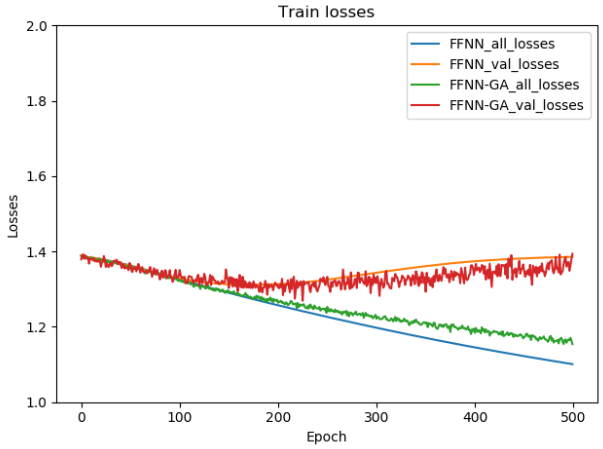
Since there are 85 features in this problem, which means the total number of possible solutions is 2^{85} and the quality of the initial population will largely affect the result of feature selection. If a randomly generated population excludes some significant features, then the crossover operator and mutation operator will have less power in introducing them back to the population. Therefore, the result obtained from the magnitude measure will be used here to select the initial population.

First, the model will be trained with all the features in the sets so the contribution of each input neuron to output neuron can be obtained then the probability of a feature being selected will depend on the contribution of the features.

2.4 Neural Network Optimisation

Through some experiments, it can be observed that the model is very easy to become overfitting that the validation loss starts increasing while the training loss is still dropping. The accuracy result on training set can grow much higher with bigger number of epochs, but the accuracy result on test set may decrease below 25%. Dropout is a regularisation technique that prevent overfitting in the neural network models. The technique randomly drops neurons from the NN with a predetermined probability during the training phase in each iteration to prevent neurons from being too dependent on others. [16] As shown in the figure 1, the model with a dropout layer shows more resistant to the overfitting than the baseline model. Having the same number of epochs, the model with dropout layer get smaller differences between the training loss and the validation loss. As adding a dropout layer will increase the number of hyperparameters, the genetic algorithm will also be utilised to find the optimal hyperparameter sets for the modified model. The parameter to be determined are the number of hidden neurons, learning rate, number of epochs and the dropout rate. In comparison to the feature selection task, there are much less combinations for the hyperparameters, therefore, the elitism selection method and random initial population generalisation are adapted here to find the optimal set for the parameters.

Figure 2. Train/Validation losses



The model with a dropout layer and hyperparameter optimisation using GA will be called FFNN-GA in the later section. The table 2 below summarises all combinations of models and techniques and the structure of the models. These methods will be evaluated in the next section.

Table 2. Summary of models and techniques

Model	Techniques	Category of feature selection
FFNN	-	-
FFNN + MM FFNN-GA + MM	Magnitude Measure + RFE	Wrapper
FFNN+ DA FFNN-GA + DA	Distinctiveness Analysis + RFE	Wrapper
mRMR + FFNN mRMR + FFNN-GA	Minimum Redundancy Maximum Relevance	Filter
FFNN + GA FFNN-GA + GA	Genetic Algorithm	Wrapper

3 Results and Discussion

3.1 Evaluation Criteria and Table of Results

A Leave-One-Participant-Out Cross-Validation (LOPO-CV) scheme, which is adapted from the dataset paper, is used when training and testing the models. The session of a participant is removed and will be used as the test set and then the model is trained on the data of the remaining participants. The process will be repeated for all and the median result of the trained models will be used to determine the performance of the model. Thus, it will lead to 12-fold cross validation as there are 12 participants in the dataset. Accuracy result on the test set will be used to evaluate the performances of the models as well as the models with techniques.

The Table 2 below summaries the best accuracy results of all the models and the number of inputs with best accuracy result for each model.

Table 3. The accuracy results with models and techniques

FFNN model without GA Hyperparameter optimisation			FFNN + GA Hyperparameter optimisation		
Feature selection criteria	No. of inputs with best accuracy result	Accuracy Result on test set	Feature selection criteria	No. of inputs with best accuracy result	Accuracy Result on test set
-	85	34.375 %	-	85	34.375 %
MM	25, 22	43.75 %	MM	21	46.25 %
DA	55, 50	37.50%	DA	62	37.50%
FCD	39	40.63%	FCD	56	37.50%
FCQ	5	37.50 %	FCQ	67	37.50 %
GA	Multiple values	40.63%	GA	Multiple values	43.75%
Dataset paper [3]	-	88%	Dataset paper [3]	-	92%

Note that the FFNN model and FFNN-GA model without implementation of feature selection achieve the same result. This is because that the FFNN model has already been through some hyperparameter tuning and also, as the FFNN-GA model has a dropout layer that randomly drops the neurons, it will have a different result but we only take the result at a fixed random state for the purpose of comparsion.

3.2 Comparison of models

Magnitude Measure & Distinctiveness Analysis

As shown in Table 2, for both the FFNN model and FFNN-GA model, selecting features based on magnitude measures achieve the highest accuracy results among all the techniques. As for distinctiveness analysis, the best results are only slightly better than the result with no feature removed. The figures show the accuracy results with a different number of input neurons removed. It can be seen from the figure that it is hard to determine a desired number for the size of feature subsets beforehand as the accuracy result is unstable. The highest accuracy result is achieved when around 60 neurons are eliminated from the model, which indicates that many features within this dataset offer little useful information. When considering both techniques, the FFNN-GA model generates a result with larger variance.

Fig. 3 Accuracy result with MM

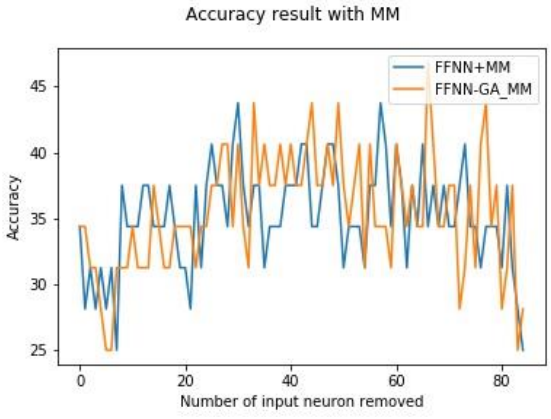
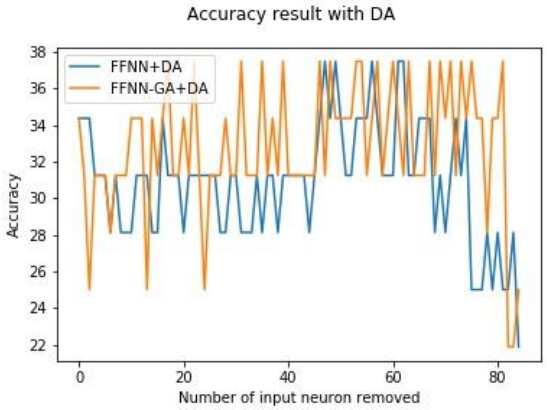


Fig. 4 Accuracy result with DA



mRMR

As show in the table 2, the best accuracy result that the model with mRMR technique using FCD criteria is 40.63%, which is 6.25% higher than the result with baseline model. Fig. 5 shows that the accuracy results for FCD fluctuate between 40.63% and 25%. The FFNN model performs better than the FFNN-GA alternative, and this drop in accuracy can most likely be attributed to the inclusion of the dropout layer.

Fig. 5 Accuracy result with mRMR-FCD

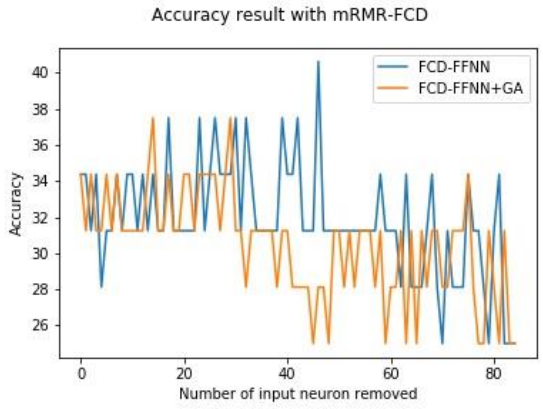
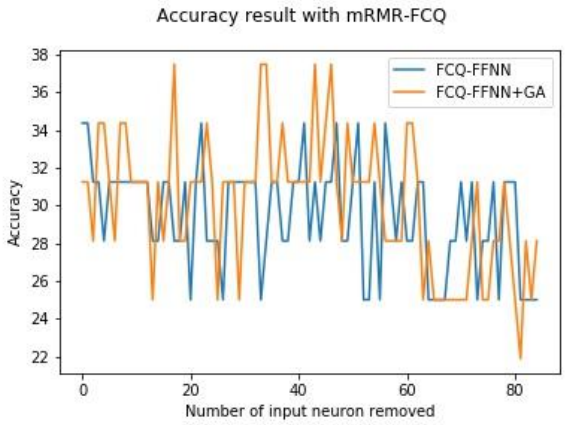


Fig. 6 Accuracy result with mRMR-FCQ



Genetic Algorithm

As show in the table 2, the best accuracy result that the FFNN model with GA technique is 40.63%, which is the same as mRMR-FCD and the best result for FFNN-GA model is 43.75%. The performance of Genetic algorithm is comparably great but not as good as magnitude measure. The initialisation using magnitude measure as criterion does not promise the optimality. Moreover, the method is very inefficient compared to other methods for this depression level classification problem. Processing through one generation with a population of 80 generally takes over one hour as there are 80*12 models (80 sets of features and 12-fold cross validation) to be trained in one generation. Although it has more chance to find the global optimal solution, it is still very likely to get stuck in the local maxima if the initial population is terrible. For example, based on the results we've obtained with FFNN-GA model extended with feature selection techniques, the magnitude measure manages to find the best result out of all the models with 46.25% while the best performance that the GA gets is 43.75%.

The result shown in the research paper with a NN classification model can achieve an average of 88% accuracy. However, the best result I can achieve with a FFNN model with similar structure and the same cross-validation scheme is only 34.38%. My result is only slightly better than a classification model guessing at random, which would have an accuracy of 25%. With the implementation of the feature selection methods, the accuracy results get improved but is still much lower than 88%. Adding a dropout layer to prevent overfitting and apply genetic algorithm for hyperparameter optimisation may find the neural network model with better performance but the good performance is not guaranteed for each run as the neurons are dropped randomly. The reason for not running for multiples times and taking the average value as the final result is that it generally takes 30 minutes with each technique to get to the state where all the neurons are removed.

Although going from 34.38% to 43.75% is a great improvement, the size of the testing set is not large, which means that the results will have a high variance. If the size of the dataset is larger, more credible results can be observed.

4 Conclusion and Future Work

In this report, various feature selection methods are implemented with the FFNN model and their performances are compared. Some techniques can offer a much smaller subset of features while improving the performance of the FFNN model. When trained on the observers' physiological signals to videos of various depression level dataset, the FFNN model extended with the magnitude measure technique to eliminate features achieves the highest accuracy result on the test set. This is different to the conclusion in the technique paper [4] which determines that the use of distinctiveness analysis works better than the magnitude-based technique. This could be due to the discrepancy between the size of the dataset and between types of variables.

Using the result from magnitude measure as a criterion in constructing the initial population for the genetic algorithm does not show much improvement from the random initial population method, therefore a more advanced initialisation approach can be developed to accelerate the convergence speed as well as improve the performance of the solutions. Furthermore, a more complexed method can be developed to check through the quality of the members in the population and replace the bad ones with some better individuals from last generation.

References

[1] Vahey, R., Becerra, R.: Galvanic skin response in mood disorders: A critical review. In: International Journal of Psychology & Psychological Therapy, vol. 15(2), pp. 275--304. (2015)

[2] Li, M., Cao, L., Zhai, Q., Li, P., Liu, S., Li, R., et al.: Method of depression classification based on behavioral and physiological signals of eye movement. In: Complexity, vol. 2020, pp. 9. (2020)

[3] Zhu, X., Gedeon, T., Caldwell, S. & Jones, R.: Detecting emotional reactions to videos of depression. In: IEEE International Conference on Intelligent Engineering Systems. (2019)

[4] Gedeon, T.D.: Data Mining of Inputs: Analysing Magnitude and Functional Measures. In: International journal of neural systems, vol. 8(2), pp. 209-18. (1997).

[5] Gedeon, T.D., Harris, D: Network Reduction Techniques. In: Proceedings International Conference on Neural Networks Methodologies and Applications. AMSE, vol. 1, pp. 119-126, San Diego. (1991)

[6] Nowé A. Genetic Algorithms. In: Gargaud M. et al. (eds) Encyclopedia of Astrobiology. Springer, Berlin, Heidelberg. (2011)

[7] Valstar, M., et al.: Avec 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 3–10. (2014)

[8] Hoque, N., Bhattacharyya, D. K., & Kalita, J. K.: MIFS-ND: A mutual information-based feature selection method. In: Expert Systems with Applications, vol. 41(14), pp. 6371-6385. (2014)

[9] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. In: The Journal of Machine Learning Research, vol. 3, pp. 1157-1182. (2003)

[10] Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. In: Artificial Intelligence, vol. 97 (1), pp. 245-271. (1997)

[11] Garson, G.D.: Interpreting Neural Network Connection Weights. In: AI Expert, pp. 47-51. (1991)

[12] Kuhn, M., Johnson, K.: Feature engineering and selection: a practical approach for predictive models. Chapman and Hall/CRC. 2020.

[13] Ding, C., Peng, H.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: Journal of Bioinformatics and Computational Biology, vol. 03, No. 02, pp. 185-205. (2005)

[14] Genetic algorithms for feature selection. (n.d.). Retrieved from https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection

[15] Alkhateeb, F., Al Maghayreh, E., & Doush, I. A. (Eds.).: Multi-Agent Systems: Modeling, Interactions, Simulations and Case Studies. (2011).

[16] Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. In: J. Mach. Learn. Res., 15, 1929-1958. (2014).