Change of performance of Neural Network using Feature selection

QiChang Li
Research School of Computer Science,
Australian National University
Canberra Australia
u6487995@anu.edu.au

Abstract: The task in this paper is to predict whether a participant is First language English (L1) readers or second language English (L2) readers based on the dataset "Lena-distractions-all". Due to the limitation of the size of the dataset, the K-fold cross-validation with 5 fold is used for training the model. Thus the evaluation of the model is based on the average accuracy and average loss. The logistic regression model with backpropagation is being implemented for this binary classification prediction. Also, two feature selection methods are used to improve the performance of Logistic regression. One feature selection method is brute force analysis, which improves the average accuracy of the Logistic regression model from 63.7% to 65.27%. Another feature selection method is the Genetic Algorithm, which improves the average accuracy of the Logistic regression model from 63.7% to 68.24%.

keywords:Neural Network, Logistic Regression, K-fold cross validation, Brute Force analysis, Feature selection, Genetic Algorithm.

1. Introduction

First language English (L1) readers and second language English (L2) readers can perform differently when reading the e-text with easy-to-read text and hard-to-read text[1]. As a Neural network model, Logistic regression with backpropagation has a good performance in predicting the binary classification problem[3]. In this paper, It can be used to predict whether the participant is L1 readers or L2 readers, which is a binary prediction problem.

Moreover, irrelevant features, which have no contribution, even negative contribution to the accuracy of the prediction model can be removed to improve the performance of the prediction model. Two methods below can be implemented to achieve the aim.

- Genetic Algorithm is an excellent method for the feature selection. It is a stochastic method for function optimisation based on the mechanics of natural genetics and biological evolution[6].

- Brute force analysis can also remove the most irrelevant feature by eliminating the inputs of the model and compare the results with predictions[2].

In this paper, the difference between the two methods above will be discussed. Also, we will analysis the advantage and disadvantage of those methods based on the result of them.

2. Method

2.1 Data processing

The data set used in this paper is the leana-distraction data, which is describe the participant's eye movement, their identity and the condition of the environment in the visual distraction experiment.

For the input of the neural network, the data collected from the participant have three problems need to solve:

1. Some features of the data are redundant. To make the neural network work efficiently, the redundant features need to be avoided. Thus, the features such as “Text type”, “Condition” are not being selected since they are generated from the feature “ Condition”.
2. Some of the features are not numerical data, which is not computable for the neural network. For the feature “condition”, it is being converted to the numerical type by converting AE-CH to the number 0-5. Also, for the time-type features like “Time taken”, it is being encoded to the float type as well with converting it to the number with the second unit. For some of the features which are seen as numerical type originally are object type. For all those features, they are being encoded to the float type to facilitate calculation.
3. Some of the features are clearly have no correlation with whether the participant is L1 or L2 reader. Thus the features such as “participant ID” can not be selected as the input features.

Except the features talked above, all the other features exclude the output “L1/L2” are being selected as the input of logistic regression model. As a logistic regression model for solving the binary classification problem, there is only one output which is the feature “L1/L2” with categorical type. Thus, it is being converted to the numerical type with respect to “0” represents “L1” and “1” represent “L2”.

The hypothesis before the experiment is that the L1 and L2 readers are affected differently by the easy-to-read and hard-to-read text, which can be predicted by Logistic regression model. Besides, both two feature selection methods can be useful for improving the accuracy of the prediction model is assumption before the experiment.

2.2 Model design

2.1 Logistic regression

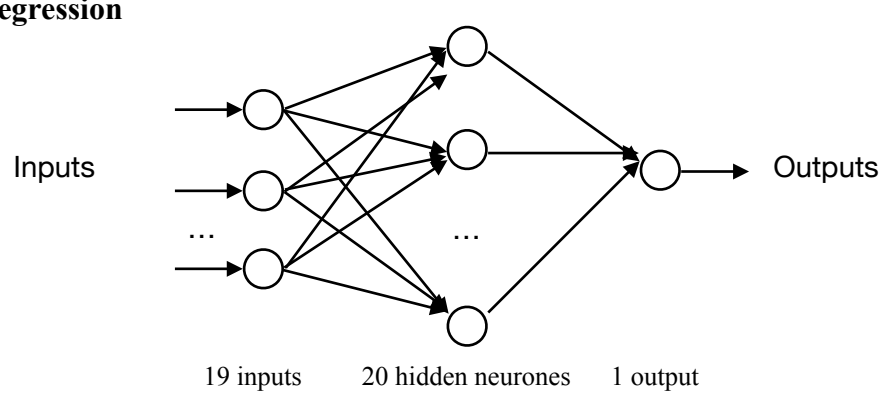


Figure 1 : the architecture of the logistic regression neural network mode

Topology:

The prediction model for predicting whether the participant is L1/L2 readers is Logistic regression Neural Network Model, which have good performance on binary classification problem[3]. In this paper, it designed as a two-layer neural network with one hidden layer with 19 features and 1 output.

Training process:

Because our dataset is relatively small with only 66-row data, the k-fold cross validation with 5 folds is implemented to ensure that the data in the dataset is fully used. This can ensure all the data in the dataset are fully used, which can improve the accuracy of the prediction[9].

Choice of Hyper-parameter:
Compare to the ReLu function and the sigmoid function, the tanh function has a better performance. For the ReLu activation function, the high learning speed can lead us to a dead ReLu problem in our dataset[8]. Besides, the tanh has a faster learning speed than the sigmoid function[7]. Thus, the activation function that is used between the input layer and the hidden layer is the tanh function. Similarly, the activation function between the hidden layer and the output layer is tanh function as well. Also, the model uses the back-propagation to training the model with the Stochastic gradient descent. The learning rate is set as 0.01 and the number of the epoch is 200 with high prediction accuracy. As for the choice of the number of hidden neurones, 10-20 are being tested, and the result indicates that 20 hidden neutrons can have a good performance.

2.2 Feature selection by Brute force

One method for the feature selection is brute force analysis. It is a method to eliminate two features and using the rest features as the input in the original model to compare the results with the predictions. From this analysis, it can show which features can affect the accuracy of the model most significantly. For the original input, it has 19 features. By dropping two features, it will have 171 different combinations. The weight generated from the original model should being used as the initial weight for each feature in the brute force analysis model(Tamás D. Gedeon, 1997). Whilst, the structure of the model also keeps unchanged exclude the number of inputs. Moreover, the accuracy of the model with the same features trained by different training data can perform differently. Thus, to make the result of the model more convincing, we run each model with the same 17 features 20 times and using the average accuracy as the result for the current model. In those cases, the brute force analysis implements 171 different models with 171 results about the accuracy with 171 different inputs. By ordering those 171 results, it can clearly show which features can affect the result most significantly.

2.3 Feature selection by Genetic Algorithm

Another method for the feature selection is the Genetic Algorithm(GA), the following is the process of the Genetic Algorithm for feature selection.

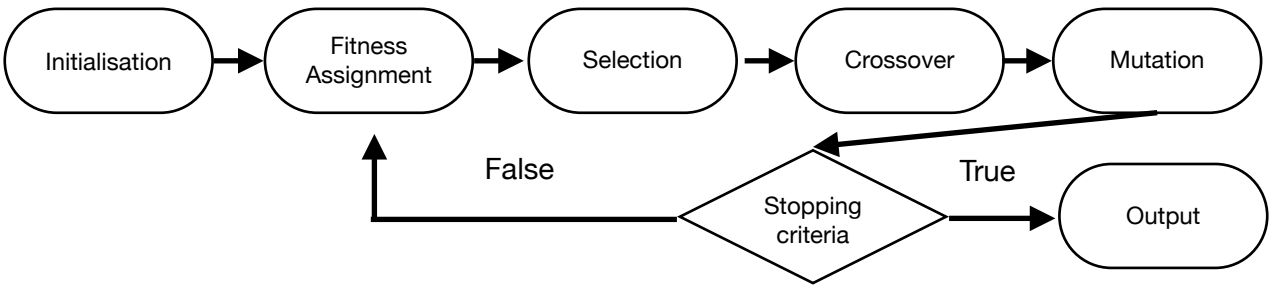


Figure 2: feature selection process with genetic algorithm

Initialisation:

To perform GA, the first step is to initialise the populations. Each chromosome in the population is initialised randomly. Each gene in the chromosome represents an input feature, thus the size of the chromosome is 19 when our initial model has 19 input. Besides, the chromosome is represented by binary, which "1" indicating the corresponding feature is being selected and "0" indicating the corresponding feature is removed. In this task, the population size is designed as1.5 times the inputs number, which is 30. This can ensure the diversity of the offspring[11]. In addition, the size of the population represents how many neural network models it will have.

The following is an example of representation of the chromosome[0,1,1,0,1].

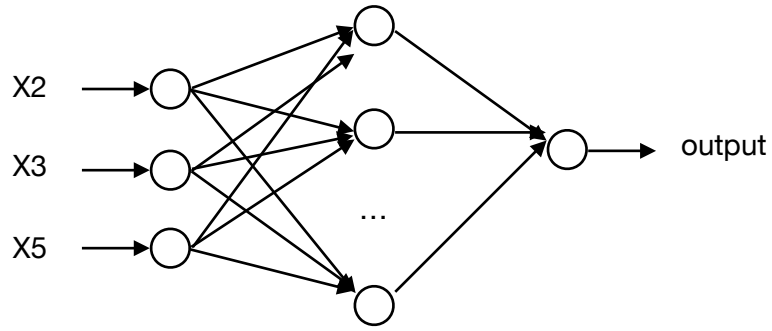


Figure 3: the architecture of the chromosome[0,1,1,0,1]

X1	X2	X3	X4	X5
0	1	1	0	1

Table 1: representation of the chromosome [0,1,1,0,1]

Fitness assignment:
The fitness function that I used for evaluate each individuals in the population is shown below:

$$\text{Fitness}(x) = \text{Accuracy}(x) / \text{loss}(x)$$

Under the limitation of the size, it is hard to improve the accuracy of the neural net model remarkably. In those cases, the evaluation of the performance of the neural net model based on the accuracy and the loss of the result simultaneously. Thus, this fitness function can ensure the Neural network can have high accuracy and low loss.

selection, crossover, and mutation:
After the fitness value is assigned to each individuals, some individuals will be selected for the next generation. Most of selected individuals have high fitness value, some of them are not fitted to the network, which is for the future mutation use. The following step is crossover step, the crossover operator that used for this GA is uniform crossover, which can ensure a high diversity of the offspring[10]. The crossover rate that used for this task is 0.8, which is also to ensure a high diversity of the offspring in each generation. More over, to ensure the high diversity, the mutation rate is set as 1/length of the chromosome, which is 0.05[11].

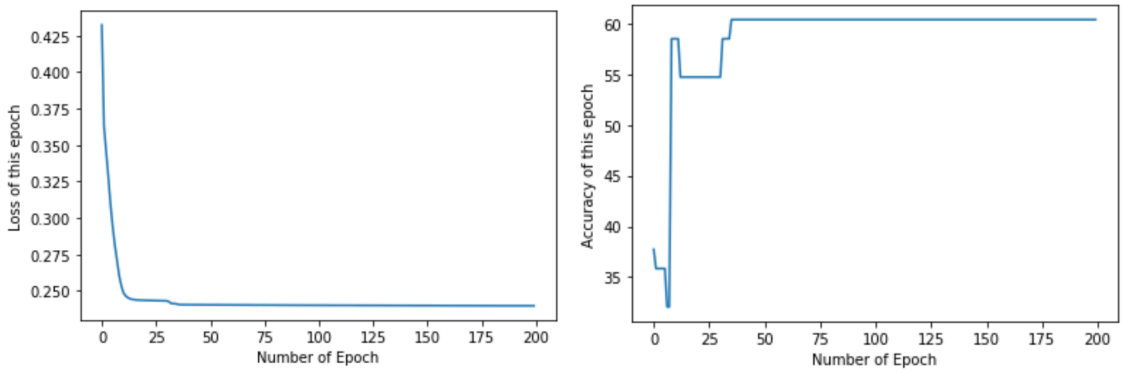
Stopping criteria and output:
Due to the limitation of the size of the data set, the accuracy of the prediction model can only obtain accuracy with 70% and the loss around 0.24 after tried 100 generations. Thus, according to our fitness function, Fitness(x)= Accuracy(x)/ loss(x) which is 70/0.24 =291, when the Fitness(x) can greater than 291, the GA is convergence. In those cases, the stopping criteria should be whether the

Fitness(x) is greater than 291 or not. Besides, the output that satisfies the stopping criteria is a chromosome that represents which features are being selected for the most optimal neural network.

3. Results and Discussion

3.1 Logistic Regression

The Diagram 4 below shows that the accuracy of the most optimal model on the training data in 200 epochs with the 60% accuracy on the training data. And 76% accuracy and 0.2212 loss of the prediction on the testing data.



Accuracy of the model on the training data : 60 %
Accuracy of the model on the test data : 76 %

Figure 4: accuracy and loss on the training data for 200 epochs

Also, the average accuracy of the k fold cross validation is 63.7% with loss of prediction 0.2427. Under a high accuracy, this indicates that we can use the features collected from the dataset (reading-distractions)to predict if a participant is L1 or L2[2]. Moreover, this shows to us the L1/ L2 readers indeed perform differently under different environment when reading the e-text.

3.2 Pearson correaltion

Before the analysis of the contribution of each feature, the correlation between each input feature and the output "L1/L2" can get from the data mining tool "rattle". It uses the Pearson correlation to demonstrate the correlation. And the result is shown in Table 1. This can be used as a reference for our analysis[12].

Input features/Pearson correlation	L1/L2	Input features/Pearson correlation	L1/L2
Scan.ratio	-0.111	fixation duration out of text area	-0.032
Skim.ratio	-0.472	Reading ratio	0.333
Total score	-0.260	Time taken	0.494
number of distractions	-0.273	Num fixations in text area	0.104
number of distractions(DB)	0.194	Total num fixations	0.097
ratio of fixation duration in text are to out of text area	-0.120	Num fixations in text area	0.187
Num fixations in text area/out of text area	-0.120	fix duration in text area	0.182
Num fixations out of text area	-0.092	Total fixation dur (s)	-0.008
Do you find that you are distracted by these technologies during study or work time?	-0.007	Do you often use social media, email and/or instant message while you are reading course materials or work materials?	-0.003

Longest reading sequence	0.313		

Table 2: Pearson correlation between the features and the output "L1/L2"

3.3 Brute force analysis

Diagram 1 shows the relationship between the accuracy and different combinations of elimination of pairs of inputs

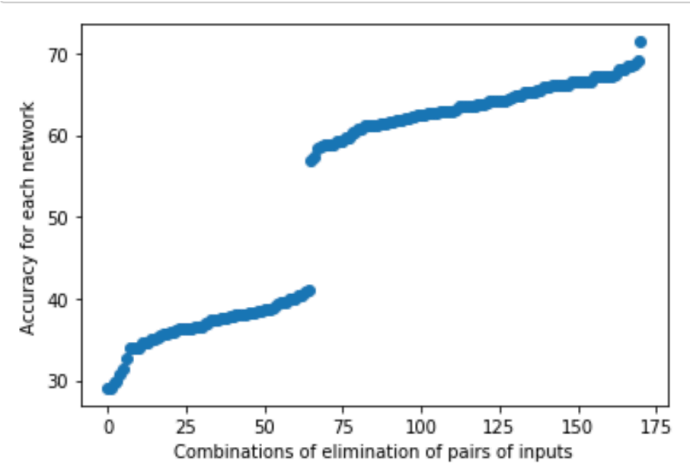


Figure 5: accuracy of 171 different Logistic regression model by brute force

Form Diagram 2, we can see there is a remarkable increase in the accuracy between 45% and 55%. To evaluate which features can affect the output of the result most significantly. I count the number of times of appearance for each feature below the accuracy of 45% and the number of times of appearance for each feature above the accuracy of 55%.

For all the combinations of elimination of pairs of inputs that have the accuracy above 55%, the feature "Do you find that you are distracted by these technologies during study or work time?" appears most frequently with the16 times. Following with the feature "Reading ratio" with 15 times. This shows to us that eliminate the feature "Do you find that you are distracted by these technologies during study or work time?" and "Reading ratio" can only slightly affect our accuracy. Whilst, this indicates the contribution of the feature "Do you find that you are distracted by these technologies during study or work time?" and "Reading ratio" to the accuracy of the output is the smallest among all the features.

For all the combinations of elimination of pairs of inputs that have the accuracy below 45%, the feature "Num fixations out of text area" appears most frequently with the 10 times. Following the feature "Time Taken" with 9 times. This indicates the correlation between those two features and the accuracy of the output is the most significant.

Compare the result from the data mining software "Rattle" to the result of getting from the brute force analysis. We can found that the feature "Do you find that you are distracted by these technologies during study or work time?" matches the result getting from the rattle, with the smallest Pearson correlation -0.00736. However, for the feature "Reading ratio", it has a relatively large Pearson correlation 0.334 to the output, which doesn't match our result generated by the brute force analysis. This may be caused by the size of the dataset is not big enough. Thus the Pearson relation generates by the "Rattle" and the result from the brute force analysis model may not

accurate enough. Also, for the feature “Num fixations out of text area” which should have a significant correlation with the output only have a Pearson correlation -0.032 from “Rattle”. However, the feature “Time Taken” match the result from “Rattle”, with a large Pearson correlation 0.49.

3.4 Genetic Algorithm

The 3 tables below are the fitness value, accuracy and loss for the 30 generations in the GA. From tables below, we can see that there is a significantly increasing of the fitness value from 15-18 generation, which indicates to us that the features being removed in those generations can affect negatively to the result.

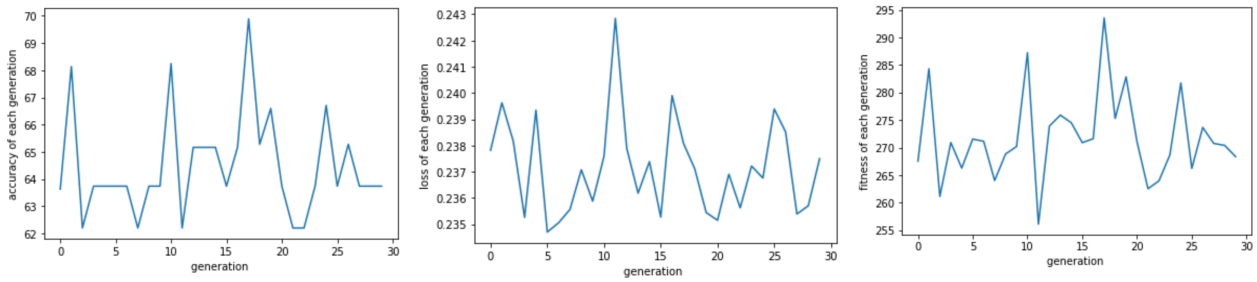


Figure 6: Average Accuracy/Average Loss/Fitness value of each generation

The following features are being selected with the most optimal performance. The fitness result of the Neural net model with those attributes is 294.5, with the 69.67% accuracy and 0.24 loss of the result on the testing set.

Selected feature	Pearson correlation	Selected feature	Pearson correlation
Num fixations in text area	-0.12	Num fixations out of text area	-0.092
Num fixations in text area/out of text area'	0.104	fixation duration out of text area'	-0.032
ratio of fixation duration in text are to out of text area'	-0.12	Reading ratio'	0.333
Longest reading sequence'	0.313	Time Taken'	0.494
Number of Distractions (from DB)'	0.194		

Table 3: the Pearson correlation for the features selected by the Genetic Algorithm

From the table above, we can see that the Pearson correlation of features selected by the GA mostly has an absolute value greater than 0.1, which indicates that they are correlated to the output. Some of the features that have a strong correlation to the output are also being selected with a Pearson correlation greater than 0.3. Moreover, from table 3, we can find all the irrelevant with Pearson’s correlation below 0.05 are removed after GA. Overall, the genetic algorithm can help us remove irrelevant features and find the most relevant features.

3.5 Comparison of the result:

The table and diagram below show the compression of the average accuracy and average loss with the k-fold cross validation after using feature selection by brute force analysis and GA on the training data. Also, the accuracy of the most optimal model from three different models on the testing data is shown as well.

Model	Average accuracy	Average loss of the result	Accuracy by most optimal model
Initial logistic regression model	63.74%	0.2427	76%
Feature selection by Brute Force	65.27%	0.2417	76%
Feature selection by Genetic Algorithm	68.24%	0.2375	84%

Table 4: Comparison of three different models

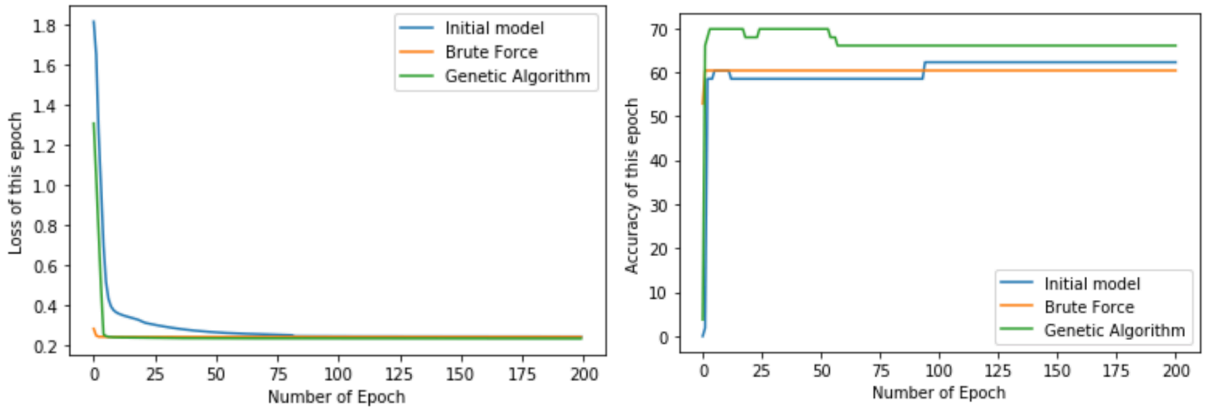


Figure 7: Comparison of three different models on the accuracy and loss on the training data

The table and diagram above show that the feature selection for the input in this task can improve the performance of the Logistic regression model. Although the improvement of the accuracy of the Neural network model is not significantly due to the limitation of the size of the dataset. Compare to the model after using GA, the model after using brute force analysis for feature selection has less improvement. This is because the brute force analysis can only help us find 2 of the most irrelevant feature and 2 features with the strongest correlation with the output. Thus, removing two irrelevant features can not improve the performance of the model significantly. Overall, the accuracy improved from 63.7% to 68.35% indicating that the neural net model after feature by GA has the most significant improvement.

4. Conclusion and Future Work

In this task, the result of the Logistic regression model shows that it can predict whether the reader is L1 readers or L2 readers based on the dataset “Lena-distractions-all”. Moreover, results from two feature selection methods indicate that the feature selection is useful for improving the performance o the Logistic regression Neural network model.

The result shows that the accuracy of the logistic regression model indicate that using the data from the dataset(reading- distractions) can predict whether the participant is L1 readers or L2 readers. The brute force analysis can present the correlation between each feature and the output to some extent, which, however, is not accurate enough as a consequence of the limited size of the dataset. The lack number of the data lead the data is not representative enough to show the correlation

between each feature and the output. However, the implementation of the brute force analysis can improve the performance of the prediction model to some extent.
The genetic algorithm satisfies the expectation, features that being selected all have a correlation to the output. Under the limitation of the size of the data, it still improves the performance of the prediction model remarkably.

Furthermore, the performance of the Logistic regression model still has room for improvement. The limitation for improvement is caused by the size of the dataset. Having a much larger dataset with sufficient data may make the contribution of each feature to the output more clearly. And then, the feature selection method can be more useful for improving the performance of the prediction model. In conclusion, the dataset having a larger number of inputs can benefit more from the feature selection model. Because the feature selection method makes the model fitting faster with fewer features but more accuracy without interference from irrelevant features.

References

[1] Copeland, Leana, and Tom Gedeon. "Visual Distractions Effects on Reading in Digital Environments: A Comparison of First and Second English Language Readers." *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. 2015.
[2] Gedeon, Tamás D. "Data mining of inputs: analysing magnitude and functional measures." *International Journal of Neural Systems* 8.02 (1997): 209-218.
[3] Bonney, George Ebow. "Logistic regression for dependent binary observations." *Biometrics* (1987): 951-973.
[4] Han, Jun, and Claudio Moraga. "The influence of the sigmoid function parameters on the speed of backpropagation learning." *International Workshop on Artificial Neural Networks*. Springer, Berlin, Heidelberg, 1995.
[5] Li, Yuanzhi, and Yang Yuan. "Convergence analysis of two-layer neural networks with relu activation." *Advances in neural information processing systems*. 2017.
[6] Gomez, F, Quesada, A. & Artelnics, R.L, Genetic algorithms for feature selection. Available at: https://www.neuraldesigner.com/blog*genetic_algorithms_for_feature_selection* [Accessed May 30, 2020].
[7] Nwankpa, C.E. et al., 2018. Activation Functions: Comparison of Trends in Practice and Research for Deep Learning,
[8] Lu, Lu & Shin, Yeonjong & Su, Yanhui & Karniadakis, George. (2019). Dying ReLU and Initialization: Theory and Numerical Examples.
[9] Jung, Y. & Hu, J. 2015, "A K-fold averaging cross-validation procedure", *Journal of Nonparametric Statistics,* vol. 27, no. 2, pp. 167-179.
[10] Sato, H. & Tanaka, K. 2011, "Genetic Diversity and Effective Crossover in Evolutionary Many-objective Optimization" in Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 91-105.
[11] Haupt, R.L. 2000, "Optimum population size and mutation rate for a simple real genetic algorithm that optimizes array factors", IEEE, , pp. 1034.

[12]Adler, J., Parmryd, I., Stockholms universitet, Naturvetenskapliga fakulteten & Wenner-Grens institut 2010, "Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient", *Cytometry Part A,* vol. 77A, no. 8, pp. 733-742.