

# Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction

Adi Sakti Almajid<sup>1,\*</sup>, Riza Arifudin<sup>1</sup>

<sup>a</sup> Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia  
\*Corresponding author: adisaktijrs@students.unnes.ac.id

## ARTICLE INFO

## ABSTRACT

### Article history

Received 11 August 2021  
Revised 11 September 2021  
Accepted 11 October 2021

### Keywords

Classification  
Imbalanced Data  
MLP  
OHE  
SVM-SMOTE

Credit risk assessment analysis by classifying potential users is an important process to reduce the occurrence of default users. The problems faced from the classification process using real-world datasets are imbalanced data that cause bias-to-majority in model training outcomes. These problems cause the algorithm to only focus on the majority class and ignore the minority class, even though both classes have the same important role. To overcome this problem, a combination of One-hot encoding (OHE) and SVM-Synthetic minority oversampling technique (SVM-SMOTE) techniques are used for the optimization process of the MLP classification algorithm. OHE is used to encode values categorical nominal and SVM-SMOTE for the oversampling. The results of the measurement of the ability of the model generated from the optimized MLP are then compared with the baseline using the AUC score. The data used is the default of credit card client dataset from Taiwan which has 30000 instances. The result of the highest AUC score of the MLP that has gone through optimization is 0.7184, an increase of 0.2179 compared to the baseline.

This is an open access article under the [CC-BY-SA](#) license.



## 1 Introduction

The development of technology has provided many benefits in various fields of life, including in the economic field. One of the development products in the economic sector is non-cash payments using credit cards. A credit card is a card issued to users as a means of payment for the purchase of goods and services (Raj and Portia, 2011). Credit cards are one of the most popular payment methods. In addition to having a positive impact, the growing number of credit card users also creates problems for banks as issuers. One of the problems that often occurs is credit risk, namely events when there is a failure to fulfill obligations by credit card users to banks (Pertiwi et al., 2020). This credit risk occurs because banks use credit lending standards that are too loose without analyzing the credibility of potential users (Yüksel, 2018).

Chen et al. explained that risk management on credit risk is important for banks to pay attention to in order to reduce the risk of losing loan money due to credit card users who default (failed to pay) (Chen et al., 2020). To overcome this problem, the bank can conduct a credit assessment by creating a model that can be used for the process of classifying prospective customers into certain classes (Louzada et al., 2016).

There are many algorithms that can be used for the modeling process for classification. One of the algorithms that can be used is artificial neural networks (ANN) which tend to produce better performance on large data (Osisanwo et al., 2017). One type of ANN that is commonly used is multilayer perceptron (MLP), as in the research conducted by Yildirim (2017), Faris et al. (2016), and Neagoe et al. (2018) who used MLP in case classification.

There are several previous studies related to classification cases using MLP which is applied to credit card default prediction. The research was conducted by Pasha et al. (2017) and Koklu (2016) who tested the Taiwan default of credit card client dataset using multilayer perceptron and produced

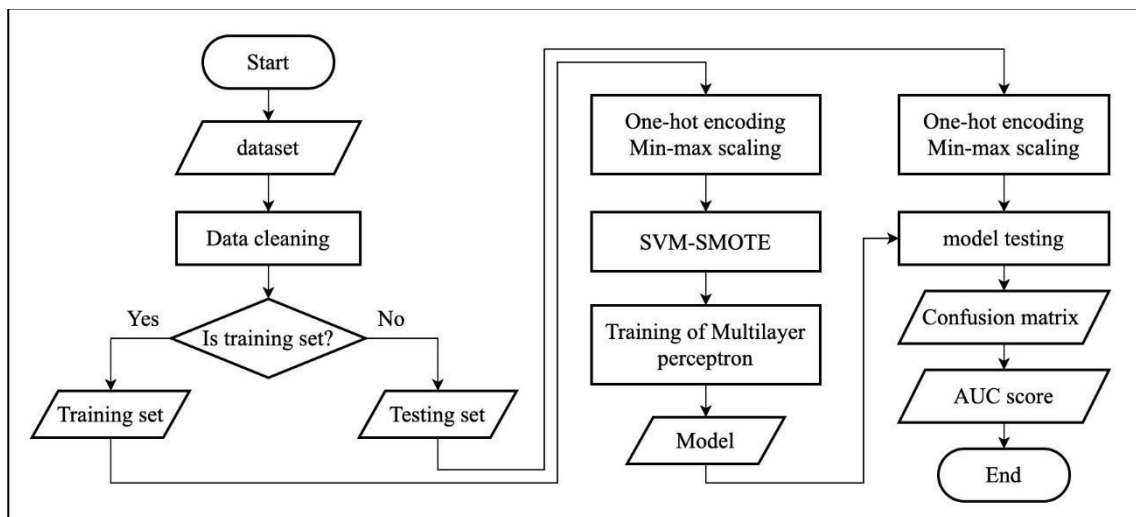
a fairly high level of accuracy of 81.7% and 81.049%. Despite obtaining a high level of accuracy, this study did not use a calculation metric that took into account the imbalanced data.

Binary classification cases that use real-world datasets are prone to imbalanced data which causes the accuracy model to be biased and cannot represent its real performance to distinguish the two classes. It happens because the trained model by datasets that have an unbalanced class distribution causes the algorithm to only focus on the majority class and ignore the minority class (Alam et al., 2020). Another study (Vishwakarma et al., 2021) that used MLP for classification on a similar dataset as Pasha et al. and Koklu, produced an AUC score of 0.5. AUC score is a method of calculating the level of a model's performance on imbalanced datasets (Vo et al., 2021). That result indicates that the model cannot differentiate the two classes properly.

The results of high accuracy but have a low AUC score indicate that the model is experiencing bias-to-majority when the used dataset is imbalanced and handled by the classification algorithm without any proper optimization phases (Liu et al., 2019). To cope with these problems, random sampling using synthetic minority oversampling technique (SMOTE) can be used to deal with imbalanced data (Suksut et al., 2019). The SMOTE technique has several development variants, one of them is SVM-SMOTE (Zhang et al., 2019). In addition to the oversampling, the encoding on the dataset also needs to be considered. This is because the encoding in the original dataset uses ordinal encoding on all categories, both on categorical ordinal and nominal. The use of ordinal encoding on features with categorical nominal values can lead to sorting bias. To avoid these problems, one-hot encoding (OHE) can be used to improve the ability of the classification algorithm (Potdar, 2017). Based on the problem mentioned above, this research focuses on optimizing the multilayer perceptron which is carried out by the optimization process using SVM-SMOTE and OHE to overcome the imbalanced data.

## 2 Method

This stage describes the analysis of the methods carried out in this study, specifically by applying OHE for data transformation in the categorical nominal class, min-max scaling in the numeric and categorical ordinal class, and SVM-SMOTE to perform oversampling on minority class instances. These two techniques are used to optimize the multilayer perceptron to classify credit card users who fail to pay. The flowchart of the proposed method in this study is shown in Figure 1.



**Figure 1.** Flowchart of the proposed method

### 2.1 Dataset

This study uses a public dataset as an object for the research experiment. The dataset was obtained from the UCI machine learning repository with the title of “the default of credit card clients” dataset from a bank in Taiwan. This data was produced by I-Cheng Yeh from the Department of Information Management, Chung Hua University Taiwan. This dataset has 30000 individual data, 23 features, and one class feature.

## 2.2 Preprocessing

The preprocessing stage includes two processes, these are data cleaning and data splitting processes. Data cleaning is the process of cleaning data from errors which can be in the form of missing entries and incorrect entries. Some of the causes of errors are the errors that occur due input, large amounts of data entered manually, and some records that are too expensive to obtain (Aggarwal, 2015). The next process is data splitting. For large amounts of data, several sample sets can be made for evaluation, namely training and testing datasets (Kuhn and Johnson, 2013). The training dataset is a sample that is used to create a model, while the testing or validation dataset is a term for a sample used for the process of evaluating the performance of a model. In this study the dataset is splitted into two parts with proportions as shown in Table 1.

**Table 1.** Distribution datasets of research

Training dataset	Testing dataset
80% of the dataset	20% of the dataset

## 2.3 Encoding and Normalization

The data encoding and normalization stages are carried out by changing the data process on the dataset into a numeric form that makes it easier for the MLP algorithm in the learning phase to create a model. The dataset used in this study has 23 features and one class feature whose values are numeric and categorical as shown in Table 2.

**Table 2.** Features in the dataset

No	Features	Attribute types
1	X1: LIMIT_BAL	Numeric
2	X2: SEX	Categorical
3	X3: EDUCATION	Categorical
4	X4: MARRIAGE	Categorical
5	X5: AGE	Numeric
6	X6: PAY_0	Categorical
7	X7: PAY_2	Categorical
8	X8: PAY_3	Categorical
9	X9: PAY_4	Categorical
10	X10: PAY_5	Categorical
11	X11: PAY_6	Categorical
12	X12: BILL_AMT1	Numeric
13	X13: BILL_AMT2	Numeric
14	X14: BILL_AMT3	Numeric
15	X15: BILL_AMT4	Numeric
16	X16: BILL_AMT5	Numeric
17	X17: BILL_AMT6	Numeric
18	X18: PAY_AMT1	Numeric
19	X19: PAY_AMT2	Numeric

No	Features	Attribute types
20	X20: PAY_AMT3	Numeric
21	X21: PAY_AMT4	Numeric
22	X22: PAY_AMT5	Numeric
23	X23: PAY_AMT6	Numeric
24	Class	Categorical

All features in the original dataset that have categorical values have been converted by the dataset maker to numeric, using the ordinal encoding. Even though not all categorical features always have ordered values (Potdar, 2017), to cope with this problem, some of categorical features that have unordered values will be encoded using OHE. Meanwhile, for other features, the encoding uses ordinal encoding according to the original conditions in the dataset and then the data normalization process will be carried out using min-max scaling. OHE works by transforming a single feature of  $n$  observers and  $d$  different values into  $d$  binary variables with  $n$  observers. OHE will change the values to 0 and 1, then on feature  $x$  with  $S = \{a, b, c\}$  with  $x_1 = a, x_2 = b, x_3 = c$  then this technique will change feature  $x$  to  $(1, 0, 0), (0, 1, 0), (0, 0, 1)$ . This process is shown in Table 3 and Table 4. In this study, the features changed by OHE are features X2, X3, and X4.

**Table 3.** Feature  $x$  before OHE

Data	$X$
1	a
2	b
3	c

**Table 4.** Feature  $x$  after OHE

Data	$x_a$	$x_b$	$x_c$
1	1	0	0
2	0	1	0
3	0	0	1

Min-max scaling is a data normalization technique that transforms raw data by rescaling the original data into a new data with a certain fixed range. This technique generally uses a range of 0 to 1 [21]. In this study, features with numerical values and categorical ordinal will be normalized using this technique.

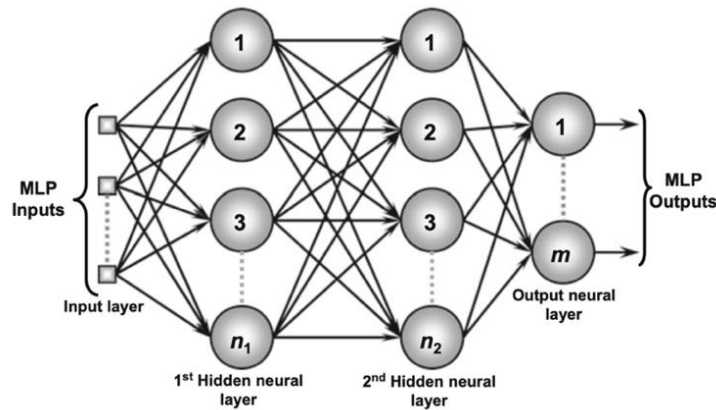
## 2.4 SVM-Synthetic minority over-sampling technique

Synthetic minority over-sampling technique (SMOTE) is an oversampling technique that is used to deal with imbalanced dataset problems by modifying the training dataset to produce a balanced training dataset for each class. One of the developments of SMOTE was carried out by Nguyen, Cooper, and Kamei by developing the SVM-SMOTE technique (Luo et al., 2019). The SVM-SMOTE technique utilizes interpolation, extrapolation, and SVM techniques to create new synthetic instances. The steps of this process are as follows (Nguyen et al., 2011).

- 1) Extract the value of  $X^+$  from  $X$ .
- 2) Calculate the value of  $T$  with the equation  $T = \left(\frac{N}{100}\right) \times |X^+|$
- 3) Calculate  $SV^+$  by training an SVM on  $X$ .
- 4) Calculate  $amount$  by evenly distributing  $T$  among  $SV^+$ .
- 5) Calculate  $nn$  of  $X^+$ .
- 6) For each  $sv_i^+ \in SV^+$  find  $m$  nearest neighbors of  $X$ . If the value of majority nearest neighbors is less than  $\frac{m}{2}$ , along the extension line connecting  $sv_i^+$  to  $k$  minority nearest neighbors, then extrapolate  $amount[i]$  artificial minority instances by using the formula in the equation  $x_{new}^+ = sv_i^+ + \rho(sv_i^+ - nn[i][j])$ . If not, then interpolate using the equation  $x_{new}^+ = sv_i^+ + \rho(nn[i][j] - sv_i^+)$ .
- 7) Returns  $X_{new}$  which is the group of  $X$  and the set  $x_{new}^+$ .

## 2.5 Multilayer Perceptron

Multilayer perceptron (MLP) is a neural network consisting of an input layer, an output layer, and one or more hidden layers that uses supervised learning paradigms. MLP is famous for its implementation of solving problems in a wide area and is considered as one of the most versatile architectures (Silva et al., 2017). An illustration of MLP is shown in Figure 2. MLP was used in this study to create a predictive model for classification. MLP has several hyperparameters that need to be configured to get the best possible model. Table 5 is the configuration arrangement of the hyperparameters that will be used.



**Figure 2.** Illustration of an MLP [24]

**Table 5.** Configuration of MLP hyperparameter

Hyperparameter	Value
Testing size	20% of the dataset
Hidden layer	2
Nodes of hidden layers	10 and 30
Activation function	Tanh and ReLU
Solver	LBFGS, SGD, Adam
Alpha	0.1, 0.3, and 0.5

## 2.6 Model Evaluation

The evaluation stage is a phase for the model that has been produced to be evaluated using a testing dataset that has never been seen by the model. The evaluation will be projected on a confusion matrix which will then be used to calculate the AUC score. The evaluation of the model will use the AUC score metric rather than using the accuracy score. The usage of accuracy score as a measurement method is not reliable for imbalanced datasets (Korkmaz, 2020) because it is prone to experiencing the accuracy paradox (Artetxe, 2020). AUC score can be calculated using Equation 1 (Devi et al., 2017).

$$AUC\ Score = \frac{1}{2} \times \left( \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right) \quad (1)$$

Where TP is the positive proportion in the dataset classified as positive, TN is the negative proportion in the dataset classified as negative, FP is the negative proportion in the dataset classified as positive, and FN is the positive proportion in the dataset classified as negative.

### 3 Results and Discussion

The results and analysis of this study is based on the experiments that have been carried out. In this study, four scenarios were used that used several combinations of methods. These scenarios aim to find out the importance of the combined methods in the MLP optimization process. Each scenario was carried out 18 times with different configurations of MLP hyperparameters to obtain the best results and the average AUC score. The results are shown in Table 6.

**Table 6.** The results of the confusion matrix and AUC score of each scenario

Code	Scenario	TP	TN	FP	FN	Best Results	Average
I	MLP	10	4580	31	1300	0.5005	0.5000
II	MLP Norm	533	4355	256	777	0.6757	0.6572
III	MLP Norm, OHE	545	4360	251	765	0.6808	0.6628
IV	MLP Norm, OHE, SVM-SMOTE	842	3661	950	468	0.7184	0.7078

The confusion matrix's values (TP, TN, FP, and FN) shown in Table 6 is the result of the confusion matrix from the experiment that produces the best AUC score. The best result is the result of one of the 18 experiments that have been carried out that produced the best AUC score. While the average result is the average value of the AUC score results in all 18 trials of each scenario.

Based on Table 6, it can be seen that the AUC score obtained from each scenario has increased. In scenario I, the model gives a very low AUC score of 0.5005. But in the confusion matrix shows that the model is successful in classifying almost all negatives with 4580 instances, but in the positive class, which is a minority class, the model only manages to classify as many as 10 instances. This shows that the model experiences a bias-to-majority which makes the model only sensitive to the majority class (non-default) and ignores the minority class (default).

In scenario II, the model managed to give a score of 0.6757 which is quite high compared to the results of the MLP baseline in scenario I, this shows that the ability of the algorithm is influenced by the range of the data used. Data with a normalized range in the range of 0 to 1 gives a score more than raw data which has a wide and different range of variations for each feature.

The results of the AUC score in scenario III were higher than scenarios I and II. This shows that the use of OHE in creating categorical nominal features has succeeded in making MLP perform better classification. Also seen in the confusion matrix, this scenario managed to correctly guess the positive class (TP) as many as 545, more than the two previous scenarios.

Scenario IV produces the highest AUC score compared to other scenarios. This shows that optimization with OHE and SVM-SMOTE has succeeded in increasing the performance of MLP. SVM-SMOTE provides solutions from imbalanced data that causes bias-to-majority by creating instances of a new synthetic training set, this makes MLP conduct a training with a balanced amount of data in both classes. This process makes the MLP sensitive to both classes and makes the AUC score higher. This scenario resulted in a score of 0.7184, an increase of 0.2179 when compared to scenario I. In addition, this scenario also managed to classify a positive class of 842 instances, the highest compared to other scenarios.

**Table 7.** Comparison of AUC Score with related studies

Research	Methods	AUC Score
(Vishwakarma et al., 2021)	MLP	0.5000
(Souza and Torres, 2021)	MLP	0.6506
Proposed method	MLP Norm, OHE, SVM-SMOTE	0.7184

Comparisons were made with previous studies using MLP for classification on the default of credit card client datasets. As shown in Table 7, the related studies used for comparison were studies (Vishwakarma et al., 2021) and (Souza and Torres, 2021). In the study (Vishwakarma et al., 2021) used several classifiers, which resulted in the MLP classifier score of 0.5000. Then in research (Souza and Torres, 2021), MLP was also used as a classification algorithm on the similar dataset which resulted in an AUC score of 0.6506. While the proposed method succeeded in producing the highest score of 0.7184.

#### 4 Conclusion

Based on the results and discussion that have been described previously, the use of OHE as an encoding algorithm for categorical nominal features and SVM-SMOTE as an alternative to the oversampling process in the training dataset can be used to cope with imbalanced data on classification problems using MLP classifier. This study shows that these two techniques can be used to optimize the MLP algorithm and produce a highest AUC score of 0.7184, a significant increase by 0.2179 compared to an MLP without the optimization process.

#### References

- Aggarwal, C. C. (2015). *Data mining: The textbook (1st.Ed)*. New York: Springer.
- Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the Imbalanced datasets. *IEEE Access*, 8, 201173-201198.
- Artetxe, A., Graña, M., Beristain, A., & Ríos, S. (2020). Balanced training of a hybrid ensemble method for imbalanced datasets: a case of emergency department readmission prediction. *Neural Computing and Applications*, 32(10), 5735-5744.
- Chen, K., Zhu, K., Meng, Y., Yadav, A., & Khan, A. (2020). Mixed credit scoring model of logistic regression and evidence weight in the background of big data. *Intelligent Systems Design and Applications*, 435-443.
- Da Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., & dos Reis Alves, S. F. (2017). *Artificial Neural Networks*. Cham: Springer International Publishing, 39.
- Devi, D., Biswas, S. K., & Purkayastha, B. (2017). Redundancy-driven modified totem-link based undersampling: A solution to class imbalance. *Pattern Recognition Letters*, 93, 3-12.
- Eesa, A. S., & Arabo, W. K. (2017). A normalization methods for backpropagation: A comparative study. *Science Journal of University of Zakho*, 5(4), 319-323.
- Faris, H., Aljarah, I., & Mirjalili, S. (2016). Training feedforward neural networks using multi-verse optimizer for binary classification problems. *Applied Intelligence*, 45(2), 322-332.
- Koklu, M., & Sabanci, K. (2016). Estimation of credit card customers payment status by using kNN and MLP. *International Journal of Intelligent Systems and Applications in Engineering*, 4(1), 249-251.
- Korkmaz, S. (2020). Deep learning-based imbalanced data classification for drug discovery. *Journal of Chemical Information and Modeling*, 60(9), 4180-4190.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
- Liu, H., Zhou, M., & Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, 6(3), 703-715.
- Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2), 117-134.
- Luo, M., Wang, K., Cai, Z., Liu, A., Li, Y., & Cheang, C. F. (2019). Using imbalanced triangle synthetic data for machine learning anomaly detection. *Computers, Materials & Continua*, 58(1), 15-26.
- Neagoe, V. E., Ciotec, A. D., & Cucu, G. S. (2018). Deep Convolutional neural networks versus multilayer perceptron for financial prediction. *2018 International Conference on Communications (COMM)*, 201-206.
- Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*,

- 3(1), 4-21.
- Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3), 128-138.
- Pasha, M., Fatima, M., Dogar, A. M., & Shahzad, F. (2017). Performance comparison of data mining algorithms for the predictive accuracy of credit card defaulters. *International Journal of Computer Science and Network Security*, 17(3), 178-183.
- Pertiwi, R. E., Syaikat, Y., & Rachmina, D. (2020). Faktor yang memengaruhi kredit bermasalah bank konvensional dan syariah di Indonesia. *Jurnal Aplikasi Bisnis dan Manajemen (JABM)*, 6(1), 118-127.
- Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4), 7-9
- Raj, S. B. E., & Portia, A. A. (2011). Analysis on credit card fraud detection methods. *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*, 152–156.
- Souza, P. V. C., & Torres, L. C. B. (2021). Extreme wavelet fast learning machine for evaluation of the default profile on financial transactions. *Computational Economics*, 57(4), 1263-1285.
- Suksut, K., Kaoungku, N., Kerdprasop, K., & Kerdprasop, N. (2019). Improvement the Imbalanced Data Classification with Restarting Genetic Algorithm for Support Vector Machine Algorithm. *International Journal of Future Computer and Communication*, 8(2), 63-67.
- Vishwakarma, S. K., Rasool, A., & Hajela, G. (2021). Machine learning algorithms for prediction of credit card defaulters—A comparative study. *Lecture Notes in Networks and Systems*, 176, 141-149.
- Vo, M. T., Vo, A. H., Nguyen, T., Sharma, R., & Le, T. (2021). Dealing with the class imbalance problem in the detection of fake job descriptions. *Computers, Materials & Continua*, 68(1), 521-535.
- Yildirim, P. (2017). Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 2, 193-198.
- Yüksel, S. (2017). Determinants of the credit risk in developing countries after economic crisis: A case of Turkish banking sector. *Global financial crisis and its ramifications on capital markets*, 401-415.
- Zhang, C., Tan, K. C., Li, H., & Hong, G. S. (2019). A cost-sensitive deep belief network for Imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1), 109-122.