Journal of Theoretical and Applied Information Technology
15th October 2024. Vol.102. No. 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MODELING OPTIMAL MLP NEURAL NETWORKS WITH DATA MINING FEATURE SELECTION IN CLASSIFYING LUNG CANCER PATIENTS

AGUS WAHYU WIDODO[1], TITIS HANDAYANI[2], FATMA AGUS SETYANINGSIH[3], IMAM NURHADI PURWANTO[4], RATNO BAGUS EDY WIBOWO[5], SAMINGUN HANDOYO[6,7*]

[1]Informatics Engineering Department, Brawijaya University, Malang 65145, Indonesia

[2]Information System Study Program, Semarang University, Semarang 50197, Indonesia

[3]Mathematics Department, Yogyakarta State University, Yogyakarta 94043, Indonesia

[4]Mathematics Department, Brawijaya University, Malang 65145, Indonesia

[5]Mathematics Department, Brawijaya University, Malang 65145, Indonesia

[6]Data Science Study Program, Brawijaya University, Malang 65145, Indonesia

[7]Electrical Engineering and Computer Science-IGP Department, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

E-mail: [1]a_wahyu_w@ub.ac.id, [2]titis@usm.ac.id, [3]fatmastyaningsih@uny.ac.id, [4]wanto_inp@ub.ac.id, [5]rbagus@ub.ac.id, [6,7]samistat@ub.ac.id

ID 55224 Submission	Editorial Screening	Conditional Acceptance	Final Revision Acceptance
31-07-2024	01-08-2024	12-09-2024	30-09-2024

ABSTRACT

In building a machine learning model, the quality of the model input has a significant role in producing the model with satisfactory performance. This study deploys the data mining feature selection to acquire the independent predictor as the input of multilayer perceptron neural networks (MLP NN) with tuning hyperparameters: node number in the hidden layer and the L2 penalty regularization. The complete predictor dataset is used to build the optimal MLP NN benchmark. Both MLP NN models have the same L2 penalty regularization of 0.05, whereas the node number in the hidden layer of 12 and 7 respectively for the dataset with the complete predictor and independent predictor. The performance evaluation of both MLP NN models in the testing data employing three metrics: accuracy, Mathew's Correlation Coefficient (MCC), and Area under curve (AUC) shows that the optimal MLP NN with independent predictors is not only producing the simpler model but also performing a slightly better than the optimal MLP NN with complete predictors.

Keywords: *Chi-square test, Cross-entropy loss, Data mining, Neural Networks, Regularization method*

1. INTRODUCTION

Machine learning (ML) classifier models applied in medical datasets range from simple logistic regression[1] to complex deep-learning models [2]. In the case of binary classification, linear classifier models including Fisher linear discriminant, Linear Discriminant Analysis, Logistic Regression, perceptron learning, and so on have the same performance as nonlinear models including decision trees, ensemble models, multi-layer perceptron neural network (MLP NN), deep neural network (DNN) and so on when two classes in a data set are separated by a linear boundary[3]. Input features or predictor variables have a major contribution to producing a high-performance model for any applied classifier model. Therefore, ensuring high-quality input features is an important stage that should be given serious attention in building machine-learning models[4]. High-quality input features can be obtained by either feature selection or feature extraction approaches. However, many works only use popular works such as the forward selection or backward elimination which involves the model candidate in the selection process. Input features

must not only have relevance to the target feature but also be independent of each other.

Feature selection involving model candidates in the selection process is known as the filter method such as forward selection or backward elimination[5]. The filter feature selection method is computationally expensive because the model candidate was executed as many times as the total of predictor feature combinations. The method is a good choice for building a model involving the linear model candidate. om the other hand, a data mining (heuristic) feature selection method works directly by evaluating relevant and independent features based on input-output data pairs without involving a model candidate[6]. The selection employs several statistical tests including chi-square, Pearson correlation, and one-way ANOVA. Relevant predictor features mean that the target features depend on them, and independent predictors mean that the predictor features are independent of each other[7]. However, assessing independent predictors is not an easy task if the dataset has a high dimension and various scales of predictor features. The predictor features involve not only the categorical and numerical scales but also various class labels in categorical features. The assessment of independent predictors with the mixture scales and various class labels is not easy work.

MLP NN implementations have satisfactory performance in either classification tasks as in the work done by[8] or regression tasks as done by[9]. However, the performance of MLP NN is directly determined not only by the network architecture such as the number of nodes in the hidden layer[10], but also by the hyperparameters of the learning algorithm such as the L2 penalty regularization[11]. Hyperparameters of machine learning models should be tuned systematically to acquire the optimal pair leading to producing the best model.

This research aims to implement and evaluate the optimal MLP NN classifier model in two scenarios: complete and independent datasets with the model hyperparameters tuned using the five folds of cross-validation data. Both hyperparameters of the number of nodes in the hidden layer and the L2 penalty regularization will be searched using the gris search method based on the average accuracy performance metric of the five folds of validation data. The optimal models will be explored relating to the loss curves and the weights preceding and succeeding of the hidden layer. The performance of both models in the testing data is evaluated in some metrics. The remainder of this article is structured as follows: related literature work in the field is reviewed in section 2, a detailed description of the data employed as a case study, a summary of the research stages, and an explanation of the proposed methods are presented in section 3. Results of hyper-parameter tuning, loss curves and model weights, and model evaluation on the testing data followed by discussion are presented in section 4. Conclusions and recommendations for the future work are given in the last section

2. LITERATURE REVIEW

Feature engineering is an important part that will determine the model input quality. A better input quality will lead to producing a better model output with a high-performance[12]. The high quality of model inputs can be acquired through either feature selection[13] or feature extraction[14]. Feature selection is a method that picks up relevant and independent features based on some statistical tests or some goodness of fit criteria. A heuristic feature selection is the feature selection method based on not only the evaluation of dependency between predictor features and the target feature but also the evaluation of the independence among the predictor features[15]. The statistical test employed to evaluate either the dependency or independence between 2 features is distinct by measurement scales of both features. The Chi-square test is employed to evaluate the dependency between 2 categorical features[16], Pearson's correlation is employed to evaluate the dependency between 2 numerical features[17], and the One-way ANOVA is employed to evaluate the dependency between the numerical and categorical features[18]. The independent features mean that predictor features do not depend on each other. On the other side, the filter feature selection involves directly a model candidate to yield the optimal input of the model by using a goodness of fit criterion[19]. Some goodness of fit criteria employed in the filter method include the Bayes information criterion (BIC), the Akaike information criterion (AIC), and the Adjusted R squared. The filter feature selection uses the fitted model candidate to calculate the goodness of fit criteria. On the other side, feature extraction projects predictor features into a commensurate measure of the orthogonal or independent components.

Model development on medical datasets is a challenge in data science because many datasets are collected within neither an experimental design nor a planned sampling frame. Data collection only has the aim of recording data for documentation purposes[20]. Several machine learning models on medical datasets including Nugroho et al.[21], Marji

and Handoyo[22], and Widodo et al.[23]. Medical data sets are very likely to contain important information such as underlying patterns in the data that can be investigated using a clustering algorithm[24], or causal relationships between features that refer to and lead to predictive modeling[25]. However, medical datasets tend to not only have a large number of feature dimensions but also have a large number of examples. Generally, medical datasets consist of categorical features in the form of qualitative or discrete data types and also have a varying number of labels[26]. Meanwhile, numerical features have various units of measurement[27]. In addition, some features may be completely unrelated to the target feature to be predicted, and some features may be redundant in the sense that two or more of them contain the same prediction information[28].

Artificial neural networks (ANN) model is a machine learning model that has evolved from simple single perceptron models to complex deep learning models[29]. A single perceptron model is equivalent to a regression model in statistical learning in that this model only performs well for modeling a linear system. On the other hand, the multi-layer perceptron (MLP) model has proven reliable for modeling complex nonlinear systems[30]. Deep learning models, especially convolution neural networks (CNN), have compelling performance in predictive modeling for datasets in the form of images[31]. The architecture or topology of the ANN model directly affects the complexity of training the model being built[32]. The more complex the ANN structure, the more model parameters, and hyperparameters are involved there. In the MLP NN model topology, the number of nodes in the hidden layer is a hyperparameter whose value must be determined before training to obtain optimal weights[33]. Apart from that, there is a trade-off in ANN models in general, namely the problem of overfitting where the model performance on testing data gets worse. Adding the L2 norm penalty to the loss function can overcome the overfitting problem[34]. The amount of this penalty must also be determined before training the model. The magnitude of the value that must be set before the model training process and does not change after the training process is complete is called a hyperparameter. Both the hidden layer node number and the L2 penalty are hyperparameters of the MLP NN model[35]. The grid search method is a simple technique to find the optimal hyperparameters based on the model candidate's performance in cross-validation data[36].

A better input quality will produce a better model output. The high-quality model input can be acquired through either the data mining approach (heuristic method) or the filter selection method. The heuristic feature selection works directly based on the dependency test using the statistical methods whereas the filter feature selection works by involving the model candidate and employing the model goodness of fit as the criterion selection. Medical data are available in a huge amount provided by many hospitals that were collected within neither an experimental design nor a planned sampling frame. The data were recorded as a part of a comprehensive service to patients. Developing an optimal MLP NN classifier model needs hyperparameter tuning to acquire the optimal pair of hyperparameters and to ensure finding the best model expected with satisfactory performance. The acquiring of relevant and independent predictors of medical datasets with high dimensions as the input of MLP NN with hyperparameter tuning is a challenging study.

3. MATERIAL AND METHODS

This raw dataset used in the work comes from the world's largest data science community named Kaggle. It can be downloaded at: https://www.kaggle.com/datasets/fdcellat/cancer-prediction-dataset. The dataset properties are given in Table 1 as follows:

Table 1: Response and predictor features with their values.

No.	Feature Name	Feature Description	Measurement Scale	Feature value
1	ID	A unique 5-digit identifier assigned to each respondent and randomly generated	String	As index
2	Gender	The respondent gender	Categorical binary	'Male' or 'Female'
3	Age	The respondent age (years)	Numerical interval	[18, 90]
4	Marital Status	The respondent Marital Status	Categorical	'Married', 'Single',

				'Widowed', or 'Separated'
5	Children	The respondent number of children	Numerical discrete	0 to 5
6	Smoker	Indicates whether the respondent smokes	Categorical binary	'Yes' or 'No'
7	Employed	The respondent employed status	Categorical binary	'Yes' or 'No'
8	Years Worked	The total number of years the respondent has been employed	Numerical interval	[0, 40]
9	Income Level	The self-assessed income level of the respondent	Categorical	'High', 'Medium', or 'Low'
10	Social Media	Indicates whether the respondent uses social media platforms	Categorical binary	'Yes' or 'No'
11	Online Gaming	Denotes whether the respondent engages in online gaming	Categorical binary	'Yes' or 'No'
12	Cancer	Indicates whether the respondent has been diagnosed with lung cancer	Categorical binary	'Yes' or 'No'

The dataset consists of 10 predictor features: three numerical and seven categorical features, and one target feature with binary class. The description of each feature is given in Table 1. The encoding process to the categorical features and commensurate measures to the numerical features were conducted on the dataset to produce the preprocessed one. This study employs two dataset scenarios: the preprocessed dataset and the produced dataset with the data mining feature selection. The optimal MLP NN classifier models are developed and evaluated in the training and testing data on each dataset scenario.

The process conducted to acquire relevant and independent predictor features using the heuristic method including:

a. evaluating the dependence of the target feature on each predictor feature and dropping the predictor features with no significant influence.

b. evaluating the independence among predictor features and choosing one feature as the representation of dependent features.

By executing both processes, it will be produced the dataset with relevant and independent predictor features which was the second dataset.

The stages conducted on each dataset to develop and evaluate the MLP NN model are summarized as follows.

1. Format the dataset in the input-output pairs: the list of predictors followed by the target feature
2. Divide the input-output pairs into the 90^ training and 10% testing data
3. Randomly Divide the training data into 5 folds and create 5 pairs of the training and validation sets.
4. Determine pairs of hyperparameter values: the number of hidden nodes and L2 regularization, supposed to cover the optimal one.
5. Execute the MLP NN model candidate on each training set and evaluate the accuracy metric on the corresponding validation set.
6. Calculate the average accuracy in all validation sets of each pair of hyperparameters.
7. Create the heat map of the average accuracy values on the hyperparameters axis.
8. Do the grid search method to find the optimal pair of hyperparameters.
9. Train the MLP NN model using the optimal hyperparameters on the training data.
10. Evaluate the model performance on the testing data.

In this study, the proposed method for yielding the optimal MLP NN model includes four main processes: the data mining feature selection, MLP NN classification model, and classification model performance metrics.

3.1 Feature selection with data mining approach

Data mining approach (heuristic method) feature selection works directly based on the results of either dependency or independent statistical tests involving both the predictor and target features[37]. The unit measurement scales of the evaluated features determine the type of statistical test that is employed for the evaluation relationship dependency between two features. The dependency between two features

with the interval or ratio scales is evaluated using Pearson's correlation test[38], where the statistical test is given in Eq. (1) as follows:

$$t_{(n-2)} = \frac{r\times(n-2)}{\sqrt{1-r^2}}, \text{ with } r_{xy} = \frac{cov(x,y)}{S_x \times S_y} \quad (1)$$

The dependency between two categorical features is evaluated using the Chi-square test with the statistical test given in Eq. (2) as follows[39]:

$$\chi^2 = \sum \frac{(Expected - Obs\ )^2}{Expected} \quad (2)$$

The value of the Chi-square statistic is calculated based on a contingency table which consists of the row and column numbers associated with the class number of both categorical features. Whereas, the evaluation of dependency between two features with different measurement scales, one is categorical and the other is the numerical features or it versa, can be assessed using the ANOVA test[40]. The F statistic given in Eq. (3) has the main role as the statistic test.

$$F = \frac{SS_B(Variance\ _{between\ groups})}{SS_W(Variance\ _{within\ groups})} \quad (3)$$

Two features are called dependent on each other when the P-value of the statistical test in Eq. (1), Eq. (2), or Eq. (3) is less than 0.05 which is the level of significance.

3.2 Multi-Layer Perceptron Neural Network Classification Model

Modeling a complex system using a Multilayer perceptron neural network (MLP NN) produces a reliable nonlinear model with satisfactory performance[41]. MLP neural network is characterized by the presence at least of one hidden layer laid between the input layer and the output layer. Two main elements have a dominant effect on the MLP NN performance: the model topology and the learning process[42]. The model topology like in Fig. 1, presents three elements: the number of inputs, the number of nodes in the hidden layer, and the output number connected to form an MLP NN. Both the numbers of input and output are related directly to the input-output data pair, whereas the node number in the hidden layer is a hyperparameter, which should be tuned systematically.

The MLP NN is considered a nonlinear function mapping the input space onto the output space through the model training using an iterative algorithm known as backpropagation[43].

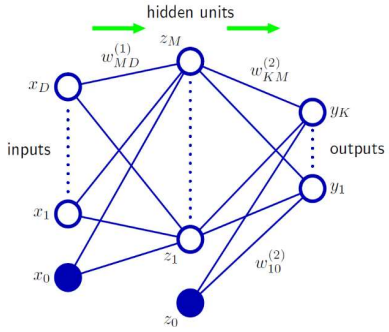

Figure 1. The MLP topology with one hidden layer

Mathematically, the model of an MLP NN with the model structure in Fig. 1, is given in Eq. (4) as follows.

$$y_k(X,W) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)}\, h\left(\sum_{i=1}^{D} w_{ji}^{(1)}\, x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \quad (4)$$

The hidden activation function $h(.)$ Is a differentiable nonlinear function that nonlinearly transforms the activation values, whereas the output activation function $\sigma(.)$ for binary class classification is a logistic sigmoid function of $\sigma(a) = 1/(1 + exp(-a))$.

The model can serve as either a classification or regression model depending on the defined loss function. The MSE loss function is associated with a numerical target feature that produces the regression model. On the other hand, the cross-entropy loss function is associated with a categorical target feature that produces the classification model[44]. The main purpose of learning the MLP NN model is to acquire the optimal weights of the model through the model trained using the training data and learning algorithm. The loss function of the MLP NN classification model with an L2 penalty is given in Eq. (5) as follows.

$$L(W;X,y) = -\frac{1}{D}[\sum_{i=1}^{D}[t_i log(y_i) + (1 - t_i) log(1 - y_i)]] + \lambda \cdot \|W\|^2 \quad (5)$$

The regularization form of the L2 penalty is added to the loss function to ensure the model does not suffer an overfitting problem[45]. The loss function of MLP NN must be minimized using the gradient descent method as the learning algorithm to acquire the optimal weights and the λ value is a hyperparameter that should be tuned systematically[46].

3.3 Performance Metrics of Classification Model

Several performance metrics are usually used to fairly evaluate classifier models to produce the best model. A simple and very popular metric in assessing the performance of a classifier model is the accuracy metric. It is calculated as the ratio between the total correctly classified instances and the total

instances in the test data[47]. The accuracy metric formula is given in Eq. (6) as follows.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

The study uses the accuracy metric as a criterion in the hyperparameter tuning using k-fold cross-validation data. Other metrics employed to evaluate the optimal MLP NN performance are the Matthews Correlation Coefficient (MCC) and Area Under Curve (AUC). The MCC metric is widely used in the evaluation of a classifier model performance in biomedical research and it is calculated based on the confusion matrix elements[48], whereas the AUC is obtained using a numerical integration approach[49]. Both MCC and AUC metrics have a range value between 0 and 1 describing a binary classifier's ability to classify instances from the positive class. Both metrics are calculated by using Eq. (7) and (8) as follows.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

$$AUC = \int_0^1 ROC(x)dx \quad (8)$$

TN, FN, TP, and TP stand for True Negative, False Negative, False Positive, and True Positive respectively. Both MMC and AUC metrics can measure the classifier model sensitivity well.

4. RESULTS AND DISCUSSION

Both datasets: complete predictor features and independent features are divided into the training and testing data. The training data are split into 5 folds and formatted in 5 pairs of cross-validation data. The hyperparameter tuning of the node number and penalty value is carried out on the cross-validation data with the average accuracy of 5 validation data as the criterion in the gird search method. The MLP NN models with the optimal hyperparameters are respectively trained and evaluated using the associated training and testing data.

4.1 Relevant and Independent Features

The dataset with the complete predictor was yielded through preprocessing the raw dataset. The encoding process of the predictor features with two, three, four, and five categorical values to be discrete was conducted. The numerical predictor features were transformed using the min-max transformation to acquire the commensurate measure in the range of [0, 1]. The selected subset predictors called the relevant and independent predictors, were acquired by conducting either dependency or independency statistical tests to evaluate not only the relationship between the predictor and target feature but also the relationship among predictor features of the complete predictor dataset. A predictor feature is called the relevant one if the target feature depends on it. In the case of a classification model, the categorical predictors were evaluated using the Chi-square test, and the numerical predictors were evaluated using the ANOVA F-test. Table 2 presents the statistical test results in selecting relevant predictors.

Table 2: The dependence test of the target to predictors

Categorical feature		
Name	**Chi-square**	**P value**
Children	44.1620	0
Gender	0.3052	***0.5807***
Marital Status	15.3076	0.0016
Smoker	2.6719	***0.2629***
Employed	38.1220	0
Income Level	22.3321	0
Social Media	27.2624	0
Online Gaming	36.4426	0
Numerical feature		
Name	F statistic	**P value**
Age	36.1300	0
Years Worked	29.7540	0

The categorical features of 'Gender' and 'smoker' did not influence the target feature shown by their p-values of 0.5807 and 0.2629, greater than 0.05. Both features are dropped from the predictor features. On the other side, both numerical features on the complete dataset have p-values of 0 in the ANOVA F-test which confirm that both features influence the target feature. The acquired relevant features consist of six categorical and two numerical predictors.

Furthermore, the relevant predictors should be independent of each other. The evaluation of independence among the relevant predictors was conducted in three steps: independence among categorical predictors, among numerical predictors, and between the categorical and numerical predictors. Table 3 presents the two combination features in the independent test among predictor features.

Table 3: The independence test among predictor features

Among categorical predictors		
Combination	**Chi-square**	**P value**
['Employed', 'Income Level']	3.6871	0.0548
['Employed', 'Social Media']	7.9725	0.0048

['Employed', 'Online Gaming']	0.5574	0.4553
['Income Level', 'Social Media']	1.4977	0.221
['Income Level', 'Online Gaming']	0.1549	0.6939
['Social Media', 'Online Gaming']	0.2816	0.5956
['Marital Status', 'Children']	9.2619	0.8634
['Children', 'Income Level']	3.4068	0.6375
['Children', 'Social Media']	4.1559	0.5272
['Children', 'Online Gaming']	8.4211	0.1345
['Marital Status', 'Income Level']	1.795	0.616
['Marital Status', 'Social Media']	1.8672	0.6004
['Marital Status', 'Online Gaming']	4.5253	0.21
Among numerical predictors		
Combination	**Corr. Value**	**P value**
['Age', 'Years Worked']	0.0416	0.1891
Between numerical and categorical predictors		
Combination	**F statistic**	**P value**
['Age', 'Children']	0.8625	0.3537
['Age', 'Marital Status']	1.7499	0.1865
['Age', 'Income Level']	1.6075	0.2051
['Age', 'Social Media']	0.3394	0.5603
['Age', 'Online Gaming']	2.8129	0.0938
['Years Worked', 'Children']	0.0005	0.9827
['Years Worked', 'Marital Status']	0.6106	0.4349
['Years Worked', 'Income Level']	1404.8910	0
['Years Worked', 'Social Media']	0.3024	0.5825
['Years Worked', 'Online Gaming']	0.1175	0.7318

There are four features with two labels, one with four labels, and one with six labels. Independent tests are presented in Table 3. The first six rows evaluate the independence of two features with two labels, and the second row shows a p-value of 0.0048, less than 0.05, which means the 'Employed' and 'Social Media' features depend on each other. The 'Employed' feature is dropped from the predictor set based on considering the fourth and sixth row results. The seventh row is the result that confirms independence between the 'Marital Status' and 'Children' features. The 8 to 13 rows show the results evaluating the independence between two labels and four or six labels, where all results show independence among them. Pearson's correlation test on the 14-row confirms that two numerical features are independent of each other. Furthermore, the ANOVA F-test on the 15-row to 24-row shows that the 'Years Worked' and 'Income Level' features are dependent on each other. The 'Income Level' feature is dropped from the predictor set and the numerical feature of 'Years Worked' is preserved. Finally, the set of relevant and independent predictors consists of four categorical predictors ('Children', 'Marital Status', 'Social Media', 'Online Gaming') and two numerical predictors ('Age', 'Years Worked') that influence the 'Cancer' target feature.

4.2 Hyperparameter tuning of the node number in the hidden layer and L2 penalty regularization

The training data were divided into 5 folds and formatted into 5 pairs of training sets and 5 validation folds. Hyperparameters tuning of the MLP NN model candidate was conducted by employing the training set and validation fold and setting the values of hyperparameters covering the optimal one. Both dataset scenarios employ the same l2 penalty values as [0.001,0.005,0.01,0.05,0.1,0.5] and the node number of [3, 5, 7, 9, 11] and [6,9,12,15,18] for the complete and independent datasets respectively. The other hyperparameters needed in executing the MLP NN model candidate were set up as follows: epoch number = 300, batch size = 30, optimization method = 'Adam', and validation metric of Accuracy.

Both hyperparameters create 30 combinations, and the optimal gird search task produces the highest average accuracy in the five validation folds. Each MLP NN model with a hyperparameter pair is trained using the training set, and the accuracy performance is calculated using the corresponding validation fold. Because of five training sets and five validation folds, each hyperparameter pair is employed five times to produce the average accuracy.

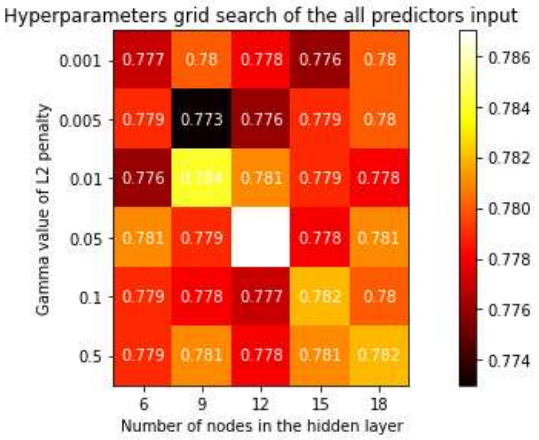

Figure 2. Heat map of the average accuracy in the 5 validation folds for all predictors as the model input

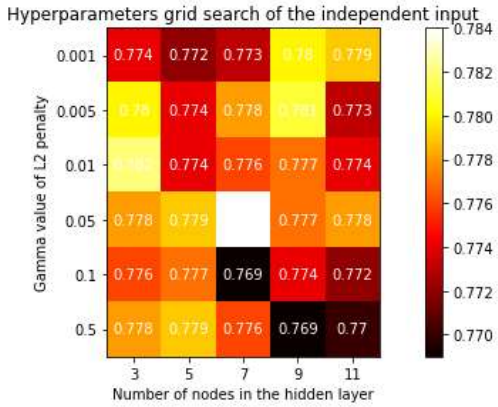

Figure 3. Heat map of the average accuracy in the 5 validation folds for relevant and independent predictors as the model input

Each cell in the heatmaps presented in Figure 2 and Figure 3 contains the average accuracy. The complete training data yields the heatmap of Figure 2, whereas the relevant and independent training data produces the heatmap of Figure 3. The optimal hyperparameter pair is found easily: the cell with a white color. data respectively. The optimal hyperparameter pair is found easily: the cell with a white color. Figure 2 shows the highest average accuracy of 0.787 with the hyperparameter pair of (7, 0.05), whereas Figure 3 shows the highest average accuracy of 0.784 with the hyperparameter pair of (12, 0.05). Both datasets yield the L2 penalty regularization value the same as 0.05, whereas the node numbers in the hidden layer are 12 and 7 for the complete and independent training data, respectively.

4.3 The Optimal MLP NN Training and The Performance Metrics

Both optimal MLP NN models are obtained using the optimal hyperparameter pairs (12, 0.05) and (7, 0.05) corresponding to the complete predictor and independent predictor datasets, respectively. Additionally, the models are trained by setting the epoch number = 300, batch size = 30, and optimization method = 'Adam'. The loss functions yielded in the training process of both models are presented in Figures 4 and 5.

Both Figures have similar losses in either training or testing curves, where the curves are flat following the epoch of 50. The training process on both datasets occurs quickly to reach the convergent condition. The training and testing loss curves almost overlap in Figure 4 and total overlap in Figure 5 without any gradual pattern occurrence. Based on both Figures, both models ensure no overfitting problem.

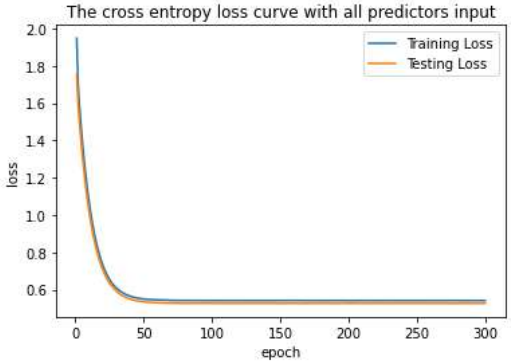

Figure 4. All predictor dataset loss curves of the training and testing data

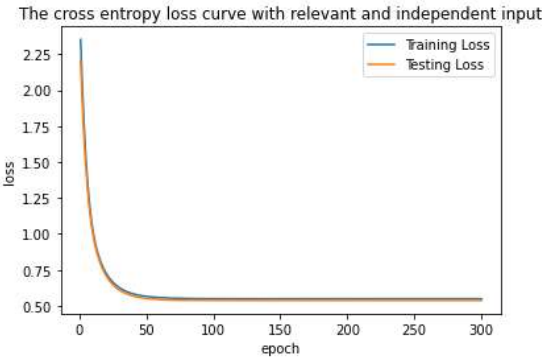

Figure 5. Independent predictor dataset loss curves of the training and testing data

Subsequently, the number of inputs and hidden layer nodes affect the weights and biases, which must be calculated through training. Both MLP NN models have two types of weights: weights preceding and succeeding the hidden layer, and two types of biases: input and output biases. Exploration of the weights preceding and succeeding the hidden layer of both models is presented in Figure 6 to Figure 9. Meanwhile, the number of input and output biases equals the number of nodes in the hidden and output layers.

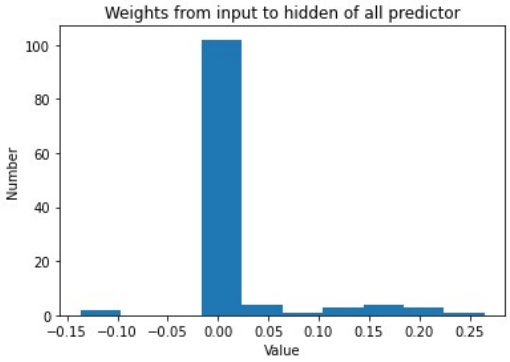

Figure 6. Weights preceding the hidden layer of the dataset of all predictors

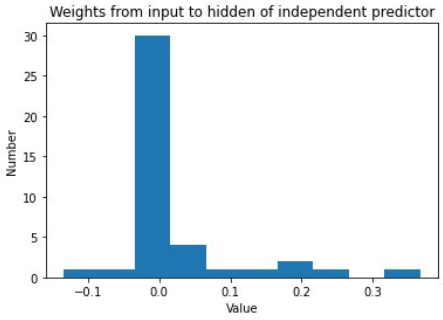

Figure 7. Weights preceding the hidden layer of the dataset of independent predictors

The MLP NN with 10 inputs (all predictors) and 12 nodes on the hidden layer has 120 weights preceding the hidden layer presented in Figure 6. In contrast, the MLP NN with six inputs (independent predictors) and 7 nodes on the hidden layer has 42 weights preceding the hidden layer presented in Figure 7. Both bar charts of Figure 6 and Figure 7 have a similar distribution with the modus in the 0 value where the frequencies are 100 and 30, respectively, for Figure 6 and Figure 7.

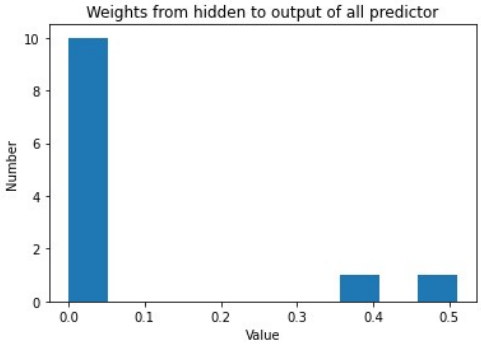

Figure 8. Weights succeeding in the hidden layer of the dataset of all predictors

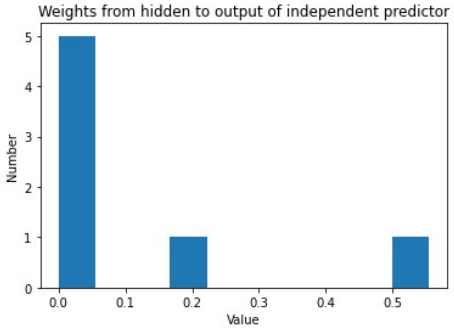

Figure 9. Weights succeeding in the hidden layer of the dataset of independent predictors

A similar distribution also occurs in Figure 8 and Figure 9. Both models have weights that succeed the hidden layer of (12, 7), which is the same as the number of nodes of the model with 12 inputs and 7 inputs, respectively. Both bar charts of Figure 8 and Figure 9 have the modus close to 0 value with frequencies of 10 and 5 for Figure 8 and Figure 9, respectively.

4.4. Discussion

The heatmap of the MLP NN with all predictors input in Figure 2 has an average accuracy value in the range of 77.6% to 78.7%, and the heatmap of the MLP NN with independent predictors input in Figure 3 has an average accuracy value in the range of 77.2% to 78.4%. The average accuracy gap of the first heatmap is 1.1%, and the second one is 1.2%. The results confirm that the hyperparameter tuning on both types of input just improves the model performance accuracy by around 1% or 2%. The condition is affected by many features that influence the target feature that is absent in the dataset. The pattern of the loss curves in Figure 4 and Figure 5 shows sharp decreases and quick convergence before the 50 epochs, supporting the fact that the feature predictors could not completely determine the class label of a patient. The weight distribution of both MLP NN models in Figure 6 to Figure 9 was dominated by zero value: 100 of 120 and 30 of 42 for Figure 6 and Figure 7, respectively, and by close to zero value: 10 of 12 and 5 of 7 for Figure 8 and Figure 9 respectively.

Implementation of the MLP NN, either for regression or classification purposes, still does not involve the hyperparameter tuning described in [40], [41], [44] and the selection of relevant and independent model inputs, including those in [34], [45], and [46]. The acquired MLP NN models were potentially not optimal because nonoptimal hyperparameters and inefficient model inputs could have been employed. Subsequently, this study successfully acquired highly qualified independent

and relevant predictor features as an input of the MLP NN model.

To produce an optimal MLP NN model, the set of complete predictors or the subset of independent and relevant predictors are treated equally at all stages of the modeling process. Table 4 presents the confusion matrix and performance metrics in the testing data of both models.

Table 4: The performance comparison of both MLP NN models in the corresponding dataset.

Performance Metrics	All Predictor Features	Independent Predictors
Confusion matrix	[[10, 11], [10, 69]]	[[12, 9], [10, 69]]
Accuracy	0.79	0.81
MCC	0.356	0.4374
AUC	0.6748	0.7224

The confusion matrices of both models in Table 4 show that the MLP NN with the independent correctly classified two instances more significant than the MLP NN with complete predictors, which was also shown by the accuracy metric: 0.81 and 0.79. Both MCC metrics are less than 50%: 35.6% and 43.74% for the model with complete and independent predictors, respectively, indicating that some important features influencing the target feature are still not covered by the dataset. Both models have AUC values greater than 0.5: 0.6748 and 0.7224 for the first and second models, respectively, implying that both models still have a moderate performance to increase to higher ones. The MLP NN with independent predictors performs slightly better than the MLP NN with complete predictors. The MLP NN with independent predictors also has a simple model structure: six predictor inputs and seven nodes in the hidden layer. This leads to faster model training due to optimizing fewer weights.

The results of this study recommend that in medical data classification modeling with MLP NN, it is essential to select a subset of relevant and independent predictor features as input to the MLP NN model and to adjust the hyperparameters of the number of nodes in the hidden layer and the L2 penalty value systematically using k-folds cross-validation data.

Feature selection that is relevant and independent will face serious challenges when the dataset has high-dimensional predictor features composed of various measurement scales, namely categorical features involving multiple class labels. Testing the independence between predictor features will be a tiring job. In addition, implementing k-fold cross-validation data will require a long computing time in high-dimensional datasets with many samples, especially in the MLP NN model, which has many model inputs and nodes in the hidden layer. Calculation of the average accuracy on each cell heatmap takes a long time.

5. CONCLUSION

The dataset with complete predictors consists of 10 predictor features: two numeric and eight categories and a target feature with a binary class. The dependency test between the predictor and target feature yields two categorical features named 'Gender' and 'Smoker', which were dropped from the dataset, and two numerical features are preserved in the dataset due to the target feature depending on both features. The evaluation independence among predictor features yields two categorical features named 'Employed' and 'Social Media', which are deleted from the predictor feature set. The dataset with independent predictors consists of four categorical and two numerical features. Both datasets employed in this study were divided into 90% training and 10% testing data. The training data are split into five folds for hyperparameter tuning. The optimal pair of hyperparameters acquired using the gird search method are (12, 0.05) and (7, 0.05) for the dataset with complete and independent predictors. The execution of MLP NN models employing the corresponding optimal hyperparameter pairs and setting up the other hyperparameters: epoch number is 300, batch size is 30, and optimization method is 'Adam' yielded the optimal MLP NN with independent predictors performing a slightly better than the optimal MLP NN with complete predictors in three metrics: accuracy, MCC, and AUC. The independent predictor dataset produces a simpler model and better performance than the complete predictors dataset. In future works, deploying the data mining feature selection in a high dimensionality, either the medical or not medical datasets, to produce the independent predictors as an input in either the linear or nonlinear machine learning models is the challenging issue.

REFERENCES:

[1] P. Kumar, G. Pradeepini, and P. Kamakshi, "Feature selection effects on gradient descent logistic regression for medical data classification," *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 5, 2019, doi 10.22266/ijies2019.1031.28.

[2] R. Ashraf *et al.*, "Deep Convolution Neural Network for Big Data Medical Image Classification," *IEEE Access*, Vol. 8, 2020, pp. 105659-105670, doi: 10.1109/ACCESS.2020.2998808.

[3] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, Vol. 7, 2021, pp. e623, doi: 10.7717/PEERJ-CS.623.

[4] S. Handoyo, N. Pradianti, W. H. Nugroho, and Y. J. Akri, "A Heuristic Feature Selection in Logistic Regression Modeling with Newton Raphson and Gradient Descent Algorithm," *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 3, 2022, pp. 119-126, doi: 10.14569/IJACSA.2022.0130317.

[5] S. K. Nayak, P. K. Rout, A. K. Jagadev, and T. Swarnkar, "Elitism based Multi-Objective Differential Evolution for feature selection: A filter approach with an efficient redundancy measure," *Journal of King Saud University - Computer and Information Sciences*, Vol. 32, No. 2, 2020, pp. 174-187, doi: 10.1016/j.jksuci.2017.08.001.

[6] Y. T. Mursityo, I. Rupiwardani, W. H. N. Putra, D. S. Susanti, T. Handayani, and S. Handoyo, "Relevant Features Independence of Heuristic Selection and Important Features of Decision Tree in the Medical Data Classification," *Journal of Advances in Information Technology*, Vol. 15, No. 5, 2024, pp. 591–601, doi: 10.12720/jait.15.5.591-601.

[7] R. Bernardes, "Machine learning - Basic principles," *Acta Ophthalmol*, Vol. 102, No. S279, 2024, doi: 10.1111/aos.16281.

[8] Y. Wang, L. Wang, F. Yang, W. Di, and Q. Chang, "Advantages of direct input-to-output connections in neural networks: The Elman network for stock index forecasting," *Inf Sci (N Y)*, Vol. 547, 2021, pp.1066-1079, doi: 10.1016/j.ins.2020.09.031.

[9] E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," *Expert Syst Appl*, Vol. 38, No. 8, 2011, pp.10389-10397, doi: 10.1016/j.eswa.2011.02.068.

[10] S. Gupta and R. R. Sedamkar, "Genetic Algorithm for Feature Selection and Parameter Optimization to Enhance Learning on Framingham Heart Disease Dataset," in *Lecture Notes in Networks and Systems*, 2021, pp. 11-25, doi: 10.1007/978-981-15-7421-4_2.

[11] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Comput Electron Agric*, Vol. 163, 2019, p. 104859, doi: 10.1016/j.compag.2019.104859.

[12] V. Liu and L. B. Chilton, "Design Guidelines for Prompt Engineering Text-to-Image Generative Models," in *Conference on Human Factors in Computing Systems - Proceedings*, 2022, pp. 1-23, doi: 10.1145/3491102.3501825.

[13] S. Campagnini, C. Arienti, M. Patrini, P. Liuzzi, A. Mannini, and M. C. Carrozza, "Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review," 2022, p. 54, doi: 10.1186/s12984-022-01032-4.

[14] A. Aggarwal *et al.*, "Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning," *Sensors*, Vol. 22, No. 6, 2022, p. 2378, doi: 10.3390/s22062378.

[15] W. H. Nugroho, S. Handoyo, H. C. Hsieh, Y. J. Akri, Zuraidah, and D. DwinitaAdelia, "Modeling Multioutput Response Uses Ridge Regression and MLP Neural Network with Tuning Hyperparameter through Cross Validation," *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 9, 2022, pp. 777-787, doi: 10.14569/IJACSA.2022.0130992.

[16] S.K. Gajawada, "Chi-Square Test for Feature Selection in Machine learning," *Towards Data Science,* No. Oct 4, 2019.

[17] L. Berus, S. Klancnik, M. Brezocnik, and M. Ficko, "Classifying parkinson's disease based on acoustic measures using artificial neural networks," *Sensors (Switzerland)*, Vol. 19, No. 1, 2018, p.16, doi: 10.3390/s19010016.

[18] M. J. Siraj, T. Ahmad, and R. M. Ijtihadie, "Analyzing ANOVA F-test and Sequential Feature Selection for Intrusion Detection Systems," *International Journal of Advances in Soft Computing and its Applications*, Vol. 14, No. 2, 2022, pp. 185-194, doi: 10.15849/IJASCA.220720.13.

[19] D. Gallacher, P. Kimani, and N. Stallard, "Extrapolating Parametric Survival Models in Health Technology Assessment: A

Simulation Study," *Medical Decision Making*, Vol. 41, No. 1, 2021, pp. 37-50, doi: 10.1177/0272989X20973201.

[20] S. Shilo, H. Rossman, and E. Segal, "Axes of a revolution: challenges and promises of big data in healthcare," *Nature medicine*, Vol. *26, No.* 1, 2020, pp. 29-38, doi: 10.1038/s41591-019-0727-5.

[21] W. H. Nugroho, S. Handoyo, and Y. J. Akri, "An influence of measurement scale of predictor variable on logistic regression modeling and learning vector quntization modeling for object classification," *International Journal of Electrical and Computer Engineering*, Vol. 8, No. 1, 2018, pp. 333-343, doi: 10.11591/ijece.v8i1.

[22] Marji and S. Handoyo, "PERFORMANCE OF RIDGE LOGISTIC REGRESSION AND DECISION TREE IN THE BINARY CLASSIFICATION," *J Theor Appl Inf Technol*, Vol. 100, No. 15, 2022, pp. 4698-4709.

[23] A. W. Widodo, S. Handoyo, I. Rupiwardani, Y. T. Mursityo, I. N. Purwanto, and H. Kusdarwati, "The Performance Comparison between C4.5 Tree and One-Dimensional Convolutional Neural Networks (CNN1D) with Tuning Hyperparameters for the Classification of Imbalanced Medical Data," *International Journal of Intelligent Engineering and Systems*, Vol. 16, No. 5, 2023, pp. 748–759, doi: 10.22266/ijies2023.1031.63.

[24] Marji, S. Handoyo, I. N. Purwanto, and M. Y. Anizar, "The effect of attribute diversity in the covariance matrix on the magnitude of the radius parameter in fuzzy subtractive clustering," *J Theor Appl Inf Technol*, vol. 96, No. 12, 2018, pp. 3717-3728.

[25] H. Kusdarwati and S. Handoyo, "Modeling Treshold Liner in Transfer Function to Overcome Non Normality of the Errors," in *IOP Conference Series: Materials Science and Engineering*, Vol. 546, No. 5, 2019, p. 052039, doi: 10.1088/1757-899X/546/5/052039.

[26] C. Ju, M. Combs, S.D. Lendle, J.M. Franklin, R. Wyss, S. Schneeweiss, & M.J. van der Laan, "Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods," *J Appl Stat*, Vol. 46, No. 12, 2019, pp. 2216–2236, doi: 10.1080/02664763.2019.1582614.

[27] M. E. Cuffaro, "The Measurement Problem Is a Feature, Not a Bug–Schematising the Observer and the Concept of an Open System on an Informational, or (Neo-)Bohrian, Approach," *Entropy*, Vol. 25, No. 10, 2023, p. 1410, doi: 10.3390/e25101410.

[28] Y. Wang, X. Li, and R. Ruiz, "Feature Selection With Maximal Relevance and Minimal Supervised Redundancy," *IEEE Trans Cybern*, Vol. 53, No. 2, 2023, pp.707-717, doi: 10.1109/TCYB.2021.3139898.

[29] K. L. Du, C. S. Leung, W. H. Mow, and M. N. S. Swamy, "Perceptron: Learning, Generalization, Model Selection, Fault Tolerance, and Role in the Deep Learning Era," *Mathematics*, Vol. *10*, No. 24, 2022, p. 4730, doi: 10.3390/math10244730.

[30] H. Kusdarwati and S. Handoyo, "System for prediction of non stationary time series based on the wavelet radial bases function neural network model," *International Journal of Electrical and Computer Engineering*, Vol. 8, No. 4, 2018, pp. 2327-2337, doi: 10.11591/ijece.v8i4.

[31] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification," *IEEE Trans Biomed Eng*, Vol. 66, No. 5, 2019, pp. 1285-1296, doi: 10.1109/TBME.2018.2872652.

[32] R. Muh Ibnu Choldun, J. Santoso, and K. Surendro, "Determining the neural network topology: A review," in *ACM International Conference Proceeding Series*, 2019, pp. 357-362, doi: 10.1145/3316615.3316697.

[33] M. G. M. Abdolrasol *et al.*, "Artificial neural networks based optimization techniques: A review," *Electronics*, Vol. *10*, No. 21, 2021, p.2689, doi: 10.3390/electronics10212689.

[34] X. Xie, M. Xie, A. J. Moshayedi, and M. H. Noori Skandari, "A Hybrid Improved Neural Networks Algorithm Based on L2 and Dropout Regularization," *Math Probl Eng*, Vol. 1, 2022, p. 8220453, doi: 10.1155/2022/8220453.

[35] Z. S. Kadhim, H. S. Abdullah, and K. I. Ghathwan, "Artificial Neural Network Hyperparameters Optimization: A Survey," *International journal of online and biomedical engineering*, Vol. 18, No. 15, 2022, p. 34399, doi: 10.3991/ijoe.v18i15.34399.

[36] T. T. Ngoc, L. Van Dai, and C. M. Thuyen, "Support vector regression based on grid

search method of hyperparameters for load forecasting," *Acta Polytechnica Hungarica*, Vol. 18, No. 2, 2021, pp. 143-158, doi: 10.12700/APH.18.2.2021.2.8.

[37] M. Hamdi, I. Hilali-jaghdam, M. M. Khayyat, B. M. E. Elnaim, S. Abdel-khalek, and R. F. Mansour, "Chicken Swarm-Based Feature Subset Selection with Optimal Machine Learning Enabled Data Mining Approach," *Applied Sciences (Switzerland)*, Vol. 12, No. 13, 2022, p. 6787, doi: 10.3390/app12136787.

[38] S. Yu, Y. Cai, B. Pan, and M. F. Leung, "Semi-Supervised Feature Selection of Educational Data Mining for Student Performance Analysis," *Electronics (Switzerland)*, Vol. 13, No. 3, 2024, p. 659, doi: 10.3390/electronics13030659.

[39] M. Wójcik, T. Goździewicz, Z. Hudáková, and I. Siatkowski, "Endometriosis and the Temporomandibular Joint—Preliminary Observations," *J Clin Med*, Vol. 12, No. 8, 2023, p. 2862, doi: 10.3390/jcm12082862.

[40] G. H. Chan, "Therapeutic comparison in psychological capital," *Front Psychiatry*, Vol. 14, 2023, p. 1114170, doi: 10.3389/fpsyt.2023.1114170.

[41] M. Z. Doghmane, M. Kidouch, S. Eladj, and A. Ouali, "Identification and Modeling of a Rotary Kiln in Cement Plant Based on ANN (MLP)," in *Lecture Notes in Networks and Systems*, 2022, pp. 825-836, doi: 10.1007/978-3-030-92038-8_84.

[42] W. Almasri, D. Bettebghor, F. Adjed, F. Ababsa, and F. Danglade, "GMCAD: An original Synthetic Dataset of 2D Designs along their Geometrical and Mechanical Conditions," in *Procedia Computer Science*, 2022, p. 337-347, doi: 10.1016/j.procs.2022.01.232.

[43] A. Eser, E. Aşkar Ayyildiz, M. Ayyildiz, and F. Kara, "Artificial Intelligence-Based Surface Roughness Estimation Modelling for Milling of AA6061 Alloy," *Advances in Materials Science and Engineering*, Vol. 1, 2021, p.5576600, doi: 10.1155/2021/5576600.

[44] G. Tzougas and K. Kutzkov, "Enhancing Logistic Regression Using Neural Networks for Classification in Actuarial Learning," *Algorithms*, Vol. 16, No. 2, 2023, p.99, doi: 10.3390/a16020099.

[45] Y. Yang, H. Zhou, Y. Gao, J. Wu, Y. G. Wang, and L. Fu, "Robust penalized extreme learning machine regression with applications in wind speed forecasting," *Neural Comput Appl*, Vol. 34, No. 1, 2022, pp.391-407, doi: 10.1007/s00521-021-06370-3.

[46] X. Li, Y. Grandvalet, and F. Davoine, "A baseline regularization scheme for transfer learning with convolutional neural networks," *Pattern Recognit*, Vol. 98, 2020, p.107049, doi: 10.1016/j.patc.2019.107049.

[47] S. Handoyo, Y. P. Chen, G. Irianto, and A. Widodo, "The varying threshold values of logistic regression and linear discriminant for classifying fraudulent firm," *Mathematics and Statistics*, Vol. 9, No. 2, 2021, pp.135-143, doi: 10.13189/ms.2021.090207.

[48] S. Diwan and M. Sahu, "Cardio Vascular Disease Prediction Using Ensemble Machine Learning Techniques," in *Lecture Notes in Electrical Engineering*, 2023, p.78, doi: 10.1007/978-981-19-8136-4_28.

[49] P. R. Killeen, "From data through discount rates to the area under the curve," *J Exp Anal Behav*, Vol. 121, No. 2, 2024, pp.259-265, doi: 10.1002/jeab.888.