Ensemble learning model on Artificial Neural Network - Backpropagation (ANN-BP) architecture for coal pillar stability classification

Gabriella Aileen Mendrofa, *Bevina Desjwiandra* Handari, and *Gatot Fatwanto* Hertono*

Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Abstract. Pillars are important structural units used to ensure mining safety in underground hard rock mines. Unstable pillars can significantly increase worker safety hazards and sudden roof collapse. Therefore, precise predictions regarding the stability of underground pillars are required. One common index that is often used to assess pillar stability is the Safety Factor (SF). Unfortunately, such crisp boundaries in pillar stability assessment using SF are unreliable. This paper presents a novel application of Artificial Neural Network-Backpropagation (ANN-BP) and Deep Ensemble Learning for pillar stability classification. There are three types of ANN-BP used for the classification of pillar stability distinguished by their activation functions: ANN-BP ReLU, ANN-BP ELU, and ANN-BP GELU. These three activation functions were chosen because they can solve the vanishing gradient problem in ANN-BP. In addition, a Deep Ensemble Learning process was carried out on these three types of ANN-BP to reduce the prediction variance and improve the classification results. This study also presents two labeling alternatives for pillar stability by considering its suitability with the SF. Thus, pillar stability is expanded into four categories: failed with a suitable safety factor, intact with a suitable safety factor, failed without a suitable safety factor, and intact without a suitable safety factor. There are five features used for each model: pillar width, mining height, bord width, depth to floor, and ratio. In constructing the model, the initial dataset is divided into training data, validation data, and testing data. In this case, four type of proportions are used. For training-testing division the proportions are: 80 % : 20 %, 70 % : 30 %, for training-validation-testing division the proportions are: 80 % : 10 % : 10 %, 70 % : 15 % : 15 %. Average accuracy, $F1$-score, and $F2$-score from 10 trials were used as performance indicators for each model. The results showed that the ANN-BP model with Ensemble Learning could improve ANN-BP performance with an average accuracy 86.48 % and an $F2$-score 96.35 % for the category of failed with a suitable safety factor.

Keywords. Artificial neural network, ensemble learning, pillar stability

* Corresponding author: gatot-f1@ui.ac.id

1 Introducton

Pillars are important structural components in underground hard rock and coal mining. This is because pillars can provide temporary or permanent support for mining and tunneling operations [1]. Pillars can protect the machine and ensure the safety of workers [2]. Unstable pillars can increase employee safety risks and potential roof collapse [3]. In addition, as the mining depth increases, the increased ground pressure can also lead to more frequent and serious pillar failures [4]. Therefore, a proper assessment of the stability of the underground pillars is necessary. Assessment of the stability of the existing pillars can provide a reference for the designer to avoid unwanted accidents [5]. In general, pillar stability can be divided into three categories, namely: stable, unstable and failed [6]. Safety Factor (SF) is a common index used in several pillar design methodologies to assess pillar stability in relation to pillar strength and average pillar stress [2]. SF is calculated by dividing the pillar strength by the pillar stress [1]. Theoretically, a rock or coal pillar is considered "unstable" if the SF value is less than 1, and "stable" if it is greater than 1. However, such rigid boundaries are often unreliable, because the occurrence of unstable pillars is also frequently appears when the SF value is above 1 [2, 6].

Machine learning techniques have been currently used effectively to evaluate the stability of pillars with better accuracy compared to other methods. This is due to an increase in the availability of pillar stability data [4]. Unfortunately, from the literature research on pillar stability prediction using ANN that has been carried out [1, 7, 8], the use of the activation function in the ANN algorithm is still limited to the sigmoid (logistic) function. The use of this function can cause vanishing gradient problems when there are many layers in an ANN [9, 10]. Vanishing gradient is a situation where the value of the partial derivative of the loss/error function gets closer to zero, and the partial derivative disappears. As a result, ANN will not improve the model weight [11] and ANN performance will not increase. Therefore, the prediction of pillar stability using the ANN algorithm needs to be improved.

One way to overcome the vanishing gradient problem in ANN is to replace the activation function. Several activation functions that have been proposed in recent years to overcome the vanishing gradient problem include ReLU, ELU, and GELU. On the other hand, ensemble learning techniques are also used in ANN to improve ANN performance. Ensemble learning combines the decisions of several predictors in two ways, namely majority vote and averaging. With this combination, the variance of the prediction results will be lower, and consequently the accuracy of the prediction results will increase [12].

In this study, the authors use the South African coal mining data used in [13] by expanding the initial pillar stability categories (intact and failed) into 4 categories based on the stability and its suitability with SF value. In addition, the authors also add a ratio variable referring to research [14]. Next, the authors use Artificial Neural Network-Backpropagation (ANN-BP) with ReLU, ELU, and GELU activation functions for pillar stability classification. Furthermore, classification of pillar stability using ensemble learning with the ANN-BP basic model was also carried out. The results of the classification using ensemble learning are then compared with the performance of a single ANN-BP.

2 Materials and method

2.1 Dataset

The dataset used in this study was taken from a journal written by J.N. van der Merwe and M. Mathey [13] entitled “Update of Coal Pillar Database for South African Coal Mining”. The data consists of 423 case histories of coal mines in South Africa with 4 types of variables

(depth, mining height, bord with, and pillar width) and 2 types of pillar stability labels (intact and failed). The intact label means that the pillar is stable, while the failed label means that the pillar is unstable which can cause it to collapse. Based on the results of data exploration, it was found that the data contained no missing values, and 337 of the 423 cases in the data were labeled intact. This means that there is a class imbalance in the data. In addition to the 4 variables contained in the data, in this study one additional variable was added, namely Ratio. The value of the Ratio variable is calculated by dividing the pillar width by mining height.

Furthermore, the SF value is also calculated for each case in the data using the Equation (1) to (3) [13, 15, 16].

$$SF = \frac{S_p}{\sigma_p} \tag{1}$$

$$S_p = 5.47 \frac{w^{0.8}}{h} \text{ MPa} \tag{2}$$

$$\sigma_p = \frac{25\text{HC}^2}{w^2} \text{ kPa} \tag{3}$$

where w denotes pillar width (meters), h denotes pillar height (meters), 5.47 is the strength of coal material (MPa), H denotes the depth to the bottom of the mine (meters), C denotes the amount of pillar width and the distance between the pillars (bord width) in meters, and 25 is the multiples of the overburden density and gravitation (kPa/m). Because there are cases of pillar stability that are not in accordance with the theory, in this study an extension of the pillar stability label was carried out in the dataset used by considering its suitability with the safety factor value. The initial labels in the dataset will be expanded from 2 categories (intact and failed) to 4 categories (F0, F1, I0, and I1) with the description of each label shown in Table 1. Because the expanded label contains additional information on the conformity of the pillar stability to the SF, it is necessary to specify the boundaries that determine whether the stability of the pillars is suitable with the calculation of the SF.

Table 1. Expanded label descriptions.

Label	Description	Safety Factor (SF) Value
F0	Failed with a suitable safety factor	SF value is lower than the specified boundary
F1	Failed without a suitable safety factor	SF value is higher than the specified boundary
I0	Intact with a suitable safety factor	SF value is lower than the specified boundary
I1	Intact without a suitable safety factor	SF value is higher than the specified boundary

There are two alternative boundary calculations used in this study. Alternative 1, the boundary is calculated by taking the midpoint between the two SF averages on each label (intact and failed) as in Equation (4).

$$\text{Boundary} = \frac{Avg.SF\ Failed + Avg.SF\ Intact}{2} = 2.23 \quad (4)$$

Alternative 2, cases with class labels F1 and I1 are considered as outliers in the SF data for each label (intact and failed). Referring to Yang et al. [17], assuming the data is normally distributed, the general equation for calculating the outlier threshold in a data is presented in Equation (5) and (6).

$$T_{min} = mean - a \times SD \quad (5)$$

$$T_{max} = mean - a \times SD \quad (6)$$

where T_{min} and T_{max} denote the minimum and maximum threshold respectively. The mean denotes the average of the data, and SD denotes the standard deviation of the data. Thus, the boundary is determined by calculating SF threshold for each label (intact and failed). The threshold used in this study was calculated from the average and standard deviation of the safety factor for each label using Equation (5) and (6) with the value a = 1. The threshold used for each label is as follows.

$$T_{failed} = Avg.SF\ Failed + Std.Failed = 2.48 \quad (7)$$

$$T_{intact} = Avg.SF\ Failed + Std.Failed = 1.42 \quad (8)$$

where T_{failed} denotes the maximum SF value limit is said to be suitable for failed labels and T_{intact} denotes the minimum SF value limit is said to be suitable for intact labels. The labeling rules using Alternative 1 and Alternative 2 are presented in Table 2.

Table 2. Labeling condition for Alternative 1 and Alternative 2 data.

Alternative 1 Condition	Alternative 2 Condition	Label
$Failed\ \&\ SF < 2.23$	$Failed\ \&\ SF \leq 2.48$	F0
$Failed\ \&\ SF \geq 2.23$	$Failed\ \&\ SF > 2.48$	F1
$Intact\ \&\ SF \geq 2.23$	$Intact\ \&\ SF \geq 1.42$	I0
$Intact\ \&\ SF < 2.23$	$Intact\ \&\ SF < 1.42$	I1

2.2 Preprocessing

Before conducting model training, preprocessing is carried out on the dataset used. This step is needed so that the data can be used as input to the model and the model can learn the characteristics of the dataset better. The preprocessing performed on the dataset includes three stages, namely oversampling, label encoding, and split dataset. The oversampling method used in this study is SMOTE. After oversampling, the number of cases for each label in Alternative 1 data was 209, while the number of cases for each label in Alternative 2 data was 312. Next, label encoding was also performed for each class. This needs to be done because ANN-BP cannot read categorical data types. Class F0 is encoded as 0, class F1 is encoded as 1, class I0 is encoded as 2, and class I1 is encoded as 3. Furthermore, the data is finally divided into training data sets, validation data, and testing data. There are four

combinations of data proportions used to test each model, each of which can be seen in Table 3.

Table 3. Types of proportions and percentages of training, validation, and testing data.

Data proportion	% Training, validation, and testing
1	80 % Training, 20 % Testing
2	70 % Training, 30 % Testing
3	80 % Training, 10 % Validation, 10 % Testing
4	70 % Training, 15 % Validation, 15 % Testing

2.3 Model training

The ANN-BP architecture used in this study is multilayer perceptron (MLP). The number of neurons used in the input layer and the output layer in this architecture are 5 and 4 respectively (because there are 5 inputs, namely depth, pillar width, mining height, bord width, and ratio, and 4 output labels, namely F0, F1, I0 , and I1). In this study, it was determined that the number of hidden layers was 4 and the number of neurons contained in the hidden layer were 512, 256, 256, and 128 respectively. There were 3 ANN-BP models used in this study which were differentiated based on the type of activation function used, namely ReLU [18], ELU [19], and GELU [9].

The type of ANN-BP ensemble learning used in this study is bagging. The ensemble learning process begins with bootstrapping, or sampling with replacement from the training data to serve as new training data for each basic model of ensemble learning (ANN-BP ReLU, ANN-BP ELU, and ANN-BP GELU). There are 3 types of bootstrap percentages used in this study, they are 70 %, 80 %, and 90 %. After bootstrapping, the training process for each ANNBP is carried out independently.

The process of determining the class in ensemble learning is done by doing a majority vote on the prediction results of each basic model. If the prediction results of the three ANN-BP are different, then the class that becomes the final prediction for ensemble learning is the class with the smallest encoding label. In this study, class with the smallest encoding label is F0. Class F0 is the best choice when there are differences in predictions among the three basic models, because class F0 is a class with the smallest risk of prediction error compared to the other classes (F1, I0, and I1). Class F0 is said to have the smallest risk of prediction error because the stability of the pillars with class F0 is in accordance with the predicted stability based on the calculation of the safety factor. In addition, because the predicted stability in class F0 is failed, the possibility of the pillar being built is also lower, and consequently the operational costs incurred are also smaller. The ensemble learning scheme carried out in this study is shown in Fig. 1.

In this study, the model simulation was implemented using the Python programming language and executed on Google Colab with GPU running time. The model is built using the TensorFlow library and trained for 10 times. In TensorFlow, the batch_size parameter specifies the number of samples in a batch that are inputted into the neural network before updating the model parameters, while the epochs parameter specifies the number of times the entire dataset is inputted into the neural network. In this simulation the batch_size used for each model is 16 with the maximum number of epochs for one training is 400. Models without data validation are trained using accuracy Early Stopping with a value of

patience = 10, which means the model will stop the training process if the training accuracy value is not improved after 10 epochs. Meanwhile, models with validation data are trained using val_loss Early Stopping with a value of patience = 10, which means the model will stop the training process if the validation loss value does not improve after 10 epochs.

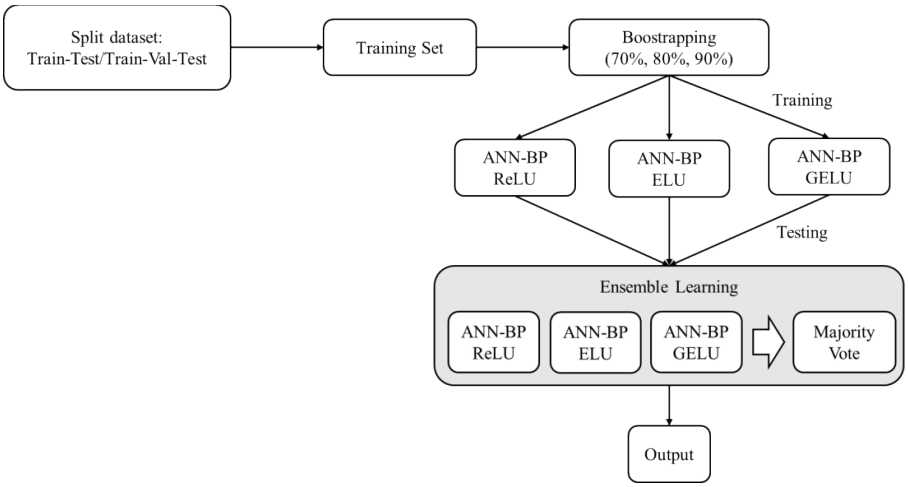

Fig. 1. Skema Ensemble Learning pada ANN-BP ReLU, ANN-BP ELU, dan ANN-BP GELU.

3 Results and discussion

The average accuracy and standard deviation of accuracy from 10 trials of the four models (ANN-BP ReLU, ANNBP ELU, ANN-BP GELU, and Ensemble Learning) for each data (Alternative 1 and Alternative 2) are presented in Table 4 to 7.

Based on the results of the average model accuracy using Alternative 1 data, ANN-BP with GELU activation function produces a better average accuracy compared to ANN-BP with ReLU and ELU activation functions for each data proportion, which is at least 0.08 % higher in the 4th type of data proportion. The best average accuracy produced by ANN-BP GELU is obtained in the 3rd data proportion, which is 92.02 %. On the other hand, with the same type of data proportion, the use of ensemble learning can increase the average accuracy produced by a single ANN-BP by at least 0.12 % in the 3rd type of data proportion.

Table 4. Average accuracy of the four models for Alternative 1 data.

Data Proportion	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
1	85.30 %	82.44 %	87.50 %	89.88 %	89.29 %	89.29 %
2	83.20 %	84.30 %	85.26 %	90.20 %	89.44 %	88.73 %
3	88.69 %	88.33 %	**92.02 %**	92.98 %	**92.14 %**	92.86 %
4	**84.68 %**	83.17 %	**84.76 %**	87.94 %	89.37 %	89.29 %

Table 5. Average accuracy of the four models for Alternative 2 data.

Data Proportion	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
1	83.48 %	82.48 %	86.32 %	88.32 %	88.68 %	88.76 %
2	83.60 %	85.44 %	87.73 %	90.64 %	90.37 %	90.21 %
3	**82.16 %**	82.56 %	**83.52 %**	87.12 %	87.92 %	86.48 %
4	85.48 %	83.24 %	**87.98 %**	89.15 %	89.20 %	**89.31 %**

Table 6. Standard deviation of model accuracy using Alternative 1 data.

Data Proportion	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
1	3.74 %	4.14 %	2.24 %	1.12 %	**0.56 %**	**1.64 %**
2	**4.42 %**	3.99 %	3.69 %	0.96 %	0.63 %	1.33 %
3	3.60 %	3.21 %	**2.03 %**	1.43 %	1.00 %	1.59 %
4	2.43 %	2.11 %	3.17 %	1.62 %	1.25 %	1.07 %

Table 7. Standard deviation of model accuracy using Alternative 2 data.

Data Proportion	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
1	1.83 %	**6.84 %**	3.65 %	0.75 %	0.98 %	1.18 %
2	2.32 %	3.42 %	1.71 %	0.98 %	0.61 %	0.78 %
3	2.30 %	2.92 %	3.83 %	1.66 %	**1.86 %**	1.53 %
4	2.31 %	4.62 %	**1.57 %**	1.55 %	0.97 %	**0.47 %**

The best average accuracy produced using ensemble learning is obtained using 70 % percentage of bootstrap in the 3rd type data proportion, which is 92.98 %. This means that the model can provide the right pillar stability predictions for around 92.98 % of the data from all the testing data used, or in other words, the chance that the model gives the right predictions for the testing data used is 92.98 %.

Based on the results of the average model accuracy using Alternative 2 data, ANN-BP with GELU activation function produces a better average accuracy compared to ANN-BP with ReLU and ELU activation functions for each data proportion, which is at least 1.36 % higher in the 3rd type of data proportion. The best average accuracy produced by ANN-BP GELU is obtained in the 4th data proportion, which is 87.98 %. On the other hand, the use ensemble learning can increase the average accuracy produced by a single ANN-BP, at least by 1.17 % in the 4th type data proportion. The best average accuracy produced using ensemble learning is obtained using 70 % percentage of bootstrap in the 2nd type data

proportion, which is 90.64 %. This means that the model can provide the right pillar stability predictions for around 90.64 % of the data from all the testing data used, or in other words, the chance that the model gives the right predictions for the testing data used is 90.64 %. Overall, of the two data alternatives, ensemble learning increases the average accuracy produced using single ANN-BP for each data proportion. However, the difference of average accuracy in ensemble learning for each bootstrap percentages is still very small (not reaching 1 %).

In Alternative 1 data, the standard deviation of ensemble learning accuracy only ranges from 0.56-1.64 %, with the smallest standard deviation of accuracy is obtained using Ensemble Learning 80 % in the 1st data proportion and the largest standard deviation of accuracy is obtained using Ensemble Learning 90 % in the 1st data proportion. Meanwhile, the standard deviation of single ANN-BP accuracy ranges from 2.03-4.42 %, with the smallest standard deviation of accuracy obtained using the GELU activation function in the 3rd data proportion and the largest standard deviation of accuracy obtained using the ReLU activation function in the 2nd data proportion.

In Alternative 2 data, the standard deviation of ensemble learning accuracy only ranges from 0.47-1.86 %, with the smallest standard deviation of accuracy obtained using Ensemble Learning 90 % in the 4th data proportion and the largest standard deviation of accuracy obtained using Ensemble Learning 80 % in 3rd data proportion. Meanwhile, the standard deviation of single ANN-BP accuracy ranges from 1.57-6.84 %, with the smallest standard deviation of accuracy obtained using the GELU activation function in the 4th data proportion and the largest standard deviation of accuracy obtained using the ELU activation function in the 1st data proportion.

Because the value is smaller, the standard deviation of ensemble learning accuracy is said to be better when compared to the standard deviation of single ANN-BP accuracy on both data alternatives. This means that the percentage accuracy of the prediction results using ensemble learning does not fluctuate too much from the average accuracy.

In addition to the average accuracy and standard deviation of accuracy, the F_1 score for each class label in the two data alternatives is also calculated and presented in Table 8 to 15. In particular, cases with class label F1 must be prioritized compared to cases with other class labels (F0, I0, and I1) because F1 is the most dangerous class. It is said to be the most dangerous because based on the calculation of the safety factor, pillars in the F1 class are categorized as intact, but in reality these pillars failed. Cases in this class will cause greater operational losses compared to other classes. Therefore, the authors assume that it is much worse to miss a failed stability prediction than to give a false alarm for an intact pillar stability. This means, in the F1 category, recall in more important than precision.

Table 8. F_1 score for 1st data proportion (Alternative 1).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	81.87 %	79.44 %	83.12 %	86.27 %	86.73 %	86.13 %
F1	94.54 %	91.81 %	96.39 %	97.42 %	96.90 %	97.01 %
I0	85.54 %	81.31 %	89.92 %	90.66 %	89.01 %	89.88 %
I1	78.83 %	75.80 %	80.50 %	84.69 %	83.76 %	83.53 %

Table 9. F_1 score for 2nd data proportion (Alternative 1).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	79.90 %	79.24 %	80.47 %	85.69 %	84.79 %	83.98 %
F1	89.40 %	91.94 %	93.40 %	96.36 %	95.98 %	96.05 %
I0	84.50 %	86.65 %	88.86 %	93.56 %	93.16 %	92.22 %
I1	79.40 %	79.86 %	79.53 %	86.34 %	85.01 %	83.78 %

Table 10. F_1 score for 3rd data proportion (Alternative 1).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	89.62 %	89.57 %	90.47 %	90.56 %	89.00 %	90.40 %
F1	91.83 %	90.85 %	96.94 %	98.11 %	97.34 %	97.83 %
I0	85.13 %	85.13 %	92.81 %	94.80 %	94.98 %	94.94 %
I1	87.48 %	87.34 %	88.34 %	89.08 %	88.07 %	88.98 %

Table 11. F_1 score for 4th data proportion (Alternative 1).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	79.27 %	80.46 %	81.42 %	87.43 %	87.77 %	88.21 %
F1	94.75 %	91.45 %	95.00 %	94.50 %	96.14 %	96.13 %
I0	88.63 %	84.90 %	87.25 %	87.42 %	91.01 %	90.31 %
I1	74.40 %	73.75 %	72.69 %	79.88 %	80.40 %	80.03 %

Table 12. F_1 score for 1st data proportion (Alternative 2).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	72.83 %	74.46 %	78.60 %	80.52 %	81.37 %	81.98 %
F1	94.91 %	95.49 %	96.61 %	97.06 %	97.13 %	97.42 %
I0	80.91 %	81.98 %	84.51 %	86.59 %	87.48 %	87.94 %
I1	83.58 %	76.91 %	84.73 %	88.28 %	88.16 %	87.37 %

Table 13. F_1 score for 2nd data proportion (Alternative 2).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	72.57 %	74.54 %	79.27 %	82.96 %	82.33 %	82.16 %
F1	95.58 %	96.16 %	96.84 %	98.19 %	97.99 %	97.94 %
I0	82.72 %	85.12 %	86.53 %	91.01 %	91.70 %	91.45 %
I1	82.82 %	85.19 %	87.66 %	90.24 %	89.54 %	89.42 %

Table 14. F_1 score for 3rd data proportion (Alternative 2).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	70.15 %	70.56 %	74.14 %	77.55 %	77.62 %	74.68 %
F1	94.85 %	94.23 %	93.87 %	96.68 %	97.37 %	96.68 %
I0	79.28 %	79.77 %	**76.52 %**	88.05 %	88.93 %	86.44 %
I1	79.78 %	81.33 %	84.53 %	83.84 %	84.91 %	84.42 %

Table 15. F_1 score for 4th data proportion (Alternative 2).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
F0	70.15 %	70.56 %	74.14 %	77.55 %	77.62 %	74.68 %
F1	94.85 %	94.23 %	93.87 %	96.68 %	97.37 %	96.68 %
I0	79.28 %	79.77 %	**76.52 %**	88.05 %	88.93 %	86.44 %
I1	79.78 %	81.33 %	84.53 %	83.84 %	84.91 %	84.42 %

In this study, F1 class recall is considered to be twice as important as F1 class precision. Therefore, the F_2 score for each model is calculated for the F1 category by choosing the value $\beta = 2$ in the F_β evaluation metric function. The results of calculating the F_2 score from class F1 for each data alternative are presented in Table 16 to 17.

Based on the F_1 score obtained in Alternative 1 data, among the other two types of activation function, GELU activation function provides better model performance in classifying the majority of class labels in each type of data proportion (except data with label I1 in the 2nd and 4th data proportion). The best F_1 score obtained using GELU activation function in Alternative 1 data can be found in the 3rd data proportion, with the F_1 score from class F0 is 90.47 %, class F1 is 96.94 %, class I0 is 92.81 %, and class I1 is 88, 34 %. On the other hand, ensemble learning provides better performance for detecting the stability of each label for almost every data proportion. The best F_1 score obtained using ensemble

learning in Alternative 1 data can be found in the 3rd data proportion. The bootstrap percentage that was used is 70 %, with the F_1 score from class F0 is 90.56 %, class F1 is 98.11 %, class I0 is 94.80 %, and class I1 is 89, 08 %.

Table 16. F_2 score of class F1 (Alternative 1).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
1	94.59 %	92.08 %	97.00 %	**98.95 %**	98.61 %	98.78 %
2	91.75 %	93.72 %	94.11 %	98.51 %	98.35 %	98.38 %
3	93.54 %	90.56 %	97.05 %	98.38 %	96.35 %	97.12 %
4	97.83 %	94.34 %	97.77 %	97.72 %	98.42 %	98.41 %

Table 17. F_2 score of class F1 (Alternative 2).

Label	ANN-BP ReLU	ANN-BP ELU	ANN-BP GELU	Ensemble Learning (70 %)	Ensemble Learning (80 %)	Ensemble Learning (90 %)
1	96.57 %	96.61 %	98.17 %	98.80 %	98.83 %	98.95 %
2	97.24 %	96.99 %	98.09 %	**99.27 %**	99.19 %	99.17 %
3	96.13 %	93.94 %	94.13 %	97.85 %	98.93 %	97.85 %
4	97.72 %	96.70 %	98.08 %	97.90 %	98.13 %	98.13 %

Based on the F_1 score obtained in Alternative 2 data, among the other three types of activation function, GELU activation function provides better model performance in classifying the majority of class labels in each type of data proportion (except data with label I1 in the 1st, 2nd and 3rd data proportion, and data with label I0 in the 3rd data proportion). The best F_1 score obtained using GELU activation function in Alternative 2 data can be found in the 2nd data proportion, with the F_1 score from class F0 is 79.27 %, class F1 is 96.84 %, class I0 is 86.53 %, and class I1 is 87.66 %. In this data alternative, ensemble learning provides equal or better performance for detecting the stability of each label for almost every proportion of data. The best F_1 score obtained using ensemble learning in Alternative 2 data can be found in the 2nd data proportion. The bootstrap percentage that was used is 70 %, with the F_1 score from class F0 is 82.96 %, class F1 is 98.19 %, class I0 is 91.70 %, and class I1 is 90.24 %.

Based on Table 16-17, it can be seen that ensemble learning provides better performance in detecting class F1 for all types of data proportions compared to single ANN-BP. In Alternative 1, the F_2 score of class F1 obtained using ensemble learning reaches 98.95 % (Ensemble Learning 70 %). In Alternative 2, the F_2 score of class F1 obtained using ensemble learning reaches 99.27 % (Ensemble Learning 70 %). This means that the ensemble learning model has a very good ability in focus on detecting the presence of class F1.

4 Conclusion

Among the three types of ANN-BP activation functions, GELU activation function provides the best performance in the classification of pillar stability measured based on average accuracy, standard deviation, and $F1$ score. The best average accuracy using ANN-BP GELU on Alternative 1 and 2 data are 92.02 % (in the 3rd data proportion) and 87.98 % (in the 4th data proportion) respectively. The best standard deviation of accuracy obtained using ANN-BP GELU on Alternative 1 and 2 data is 2.03 % (3rd data proportion) and 1.57 % (4th data proportion) respectively.

Ensemble learning (EL) provides excellent performance in the pillar stability classification measured by accuracy, standard deviation, F_1 score. The best average accuracy using EL in Alternative 1 and 2 data are 92.98 % (EL70 %) and 90.64 % (EL 70 %) respectively. The best standard deviation of accuracy using EL in Alternative 1 and 2 data are 0.56 % (EL 80 %) and 0.47 % (EL 90 %) respectively.

The use of ensemble learning can improve the performance of the pillar stability classification of a single ANN- BP, with an increase in average accuracy of at least 0.12 % for Alternative 1 data and 1.17 % for Alternative 2 data. Ensemble learning also reduces the standard deviation of accuracy to a maximum of 1 .64 % for Alternative 1 data and 1.86 % for Alternative 2 data.

Acknowledgements

The research is supported by Hibah Riset Penugasan FMIPA UI with contract number No. 003/UN2.F3.D/PPM.00.02/2022.

References

1. N. Li, M. Zare, C. Yi, and R. Jimenez, Int. J. Environ. Res. Public Health **19***,* 2136 (2022).
2. C. Li, J. Zhou, D. J. Amarghani and X. Li, Undergr. space **6**, 379-395 (2020).
3. J. A. Wang, X. Shang and H. Ma, H, Int. J. Rock Mech. Min. Sci. **6,** 1480-1499 (2008).
4. M. Ahmad, et.al, App. Sci. **10**, 6486 (2020).
5. S. Hidayat, A. Alpiana and D. Rahmawati, Int. Conf. Min. Environ. Tech, (2020).
6. P. Lunder, *Hard Rock Pillar Strength Estimation an Applied Empirical Approach*, *Ph.D. Thesis*, (University of British Columbia, Canada*,* 1994)
7. A. Tawadrous and P. Katsabanis, Int. J. Numer. Anal. Methods. Geomech. **31**, 917-931 (2007).
8. J. Zhou, X. Li, and H. S. Mitri, Nat. Hazards. **79***,* 291-316 (2015).
9. D. Hendrycks and K. Gimpel, *arXiv:1606.08415v4 [cs.LG]* (2020).
10. A. D. Rasamoelina, F. Adjailia and Sinčák, *IEEE 18th World Symposium on Applied Machine Intelligence and Informatics* (IEEE, 2020)*,* pp. 281-286.
11. S. Hochreiter, Int. J. Uncertain. Fuzziness Knowlege-Based Syst. **6**, 107-116 (1998).
12. C. Zhang and Y. Ma, in Ensemble Machine Learning: Methods and Applications 2012th Edition. (Springer, New York, 2012).
13. J. N. Van Der Merwe and M. Mathey, J. South Afr. Inst. Min. Metall. **113**, 825-840 (2013).
14. W. Liang, S. Luo, G. Zhao and H. Wu, Mathematics*,* (2020).

15. G. Song and S. Yang, Int. J. Min. Sci. Tech. **28**, 715-719 (2018).
16. B. J. Madden, J. South Afr. Inst. Min. Metall. **91**, 27-37 (1991).
17. J. Yang, S. Rahardja and P. Fränti, Proceedings of the International Conference on Artificial Intelligence, 1-6 (2019).
18. S. Sharma and A. Athaiya, Int. J. Appl. Sci. Eng. **6**, 310-316 (2020).
19. D. A. Clevert, T. Unterthiner, and S. Hochreiter, arXiv:1511.07289 (2016).