Classification of soil types based on suitable plants using Multiclass Classification Artificial Neural Network

Ivan Budianto, Nova El Maidah, Saiful Bukhori

Abstract— **Soil conditions are one of the factors that determine plant growth. For plants, soil is a place for plant growth, a place for air supply, a place for nutrient supply, and a place for plant growth. Soil conditions are divided into two, namely chemically and physically. Chemical soil conditions include the content of Sodium, Phosphorus, Hydrogen, Potassium and Calcium. Meanwhile, physically it includes daily temperature, humidity, pH, and rainfall. This research develops a neural network model to recognize soil condition data patterns with predetermined parameters. The parameters used in this research were the chemical conditions of the soil, namely levels of Nitrogen, Phosphorus and Potassium, as well as the physical condition of the soil which included temperature, humidity, pH and rainfall. After identifying the soil condition data pattern, it is used to classify soil types based on the appropriate plants. This research develops a model with 9 scenarios that vary in the ratio of data splitting and the number of layers used. Based on all trials conducted, the best scenario is the splitting of 90% training data, 5% validation data, and 5% test data with 4 layers. This model has a training accuracy of 99.30%, a validation accuracy of 99.24%, and a test accuracy of 98.93%. Model testing in this scenario is also the best with 99.24% precision, 99.49% recall, and 99.32% F1 score.**

Keywords—**soil condition, chemical soil condition, physical soil condition, classification, neural network model.**

I. Introduction

A crop is plant product that can be grown and harvested for profit or subsistence [1]. According to their use, crops divided into 6 types, namely: food crops, industrial crops, feed crops, ornamental crops, fiber crops, and oil crops [2]. The success of plant cultivation is very important, because the success of plant cultivation is one of the factors to improve the welfare of farmers. The success of plant cultivation is influenced by various factors, including plant material factors, essential factors, climatic factors and plant disturbance factors [3].

The planting material factor is very important because the planting material is the initial key to the success of plant cultivation. The planting material used by farmers is seeds [4]. The variety of plant seeds to be planted must be in accordance with soil conditions, altitude and climate. Essential factors are intake factors that affect plant growth. This factor is important because it will be processed by plants through the process of photosynthesis which will form plant biomass. Essential factors consist of nutrients, sunlight, water, and oxygen [5]. Cultivated plants are influenced by the circumstances around them. Many climatic factors affect plant growth either directly or indirectly. Direct climate influence on plants, for example, less rainfall, strong winds, and less sunlight [6]. Indirect climate influences, for example, high humidity will trigger disease attacks. Factors of good planting material, adequate supply of essential factors and a supportive climate have not guaranteed the success of crop production if the disturbance factors cannot be controlled. Disturbance factors include weeds, pests and diseases. Weeds are planting whose existence is unwanted. In order for crop cultivation to be successful, weeds, pests and diseases must be controlled [7].

Soil conditions are one of the factors that determine plant growth [8]. For plants, the soil functions as a place for plant growth, a place for air supply for breathing of plant roots and the life of microorganisms, a place for supplying nutrients for plant growth, both in the form of organic and inorganic substances, and a place for water supply to dissolve nutrients so that they can be absorbed by plants.

Soil can fertilize plants if the soil contains organic matter, inorganic substances, water and air. These substances can fertilize plants if the amount is sufficient according to plant growth. Organic substances are substances that are formed from weathering or decay of plant and animal remains [9]. Usually, organic matter is found in the top layer of soil. While inorganic substances are substances that come from the destruction of rocks and minerals, usually scattered in the subsoil. Soil becomes fertile if it contains these materials with a composition of 45% organic matter, 5% inorganic matter, 25% water and 25% air [10].

Soil conditions are divided into two types, namely soil according to chemical conditions and soil according to physical conditions [11]. Chemically, the elements present in the soil include the elements Sodium, Phosphate, Hydrogen, Potassium, and Calcium [12]. Soil physical condition factors include the daily temperature of the land, humidity, pH, and rainfall of the area [13]. Certain types of plants will grow well in suitable soil conditions. This research will develop a soil type classification system based on suitable plants.

Several researches have examined the most suitable plant classification based on soil conditions using machine learning[14][15][16][17]. These researches developed a recommendation system using Support Vector Machine (SVM), Naïve Bayes, Multi-layer perceptron, and Random Forest algorithms. The data used in this research is a dataset with soil attributes in the form of Depth, Texture, Ph, Soil

Color, Permeability, Drainage, Water holding and Erosion [18].

In this research, a neural network model is proposed to recognize soil condition data patterns with predetermined parameters. The parameters used in this research were chemical soil conditions, namely levels of Nitrogen, Phosphorous, Potassium, and also physical soil conditions which included temperature, humidity, pH and rainfall. The data is taken from Kaggle, namely the "Crop Recommendation Dataset" [19]. Research that develops this artificial neural network model is expected to solve problems in the form of selecting the best plant seeds for specific soil conditions based on the parameter sodium, phosphate, potassium, temperature, humidity, pH and rainfall of an area.

The rest of this paper is organized as follows: The proposed model for classification of soil types based on suitable plants using multiclass classification ANN is discussed in section 2. Section 3 provides a system design for classification of soil types based on suitable plants using multiclass classification ANN. Section 4 discusses the results and analysis of ANN model and integrated system applications. Finally, conclusions are given in section 5.

II. PROPOSED MODEL

The model developed in this research is a multinomial classification model with 4 layers with the ReLU activation function, 22 Softmax activation function units and using the RMSProp optimizer. (Fig. 1). The ReLU activation function allows the model to solve nonlinear problems and is suitable for multinomial classification. The Softmax activation function provides probabilities for each possible output class and is also suitable for multinomial classifications.

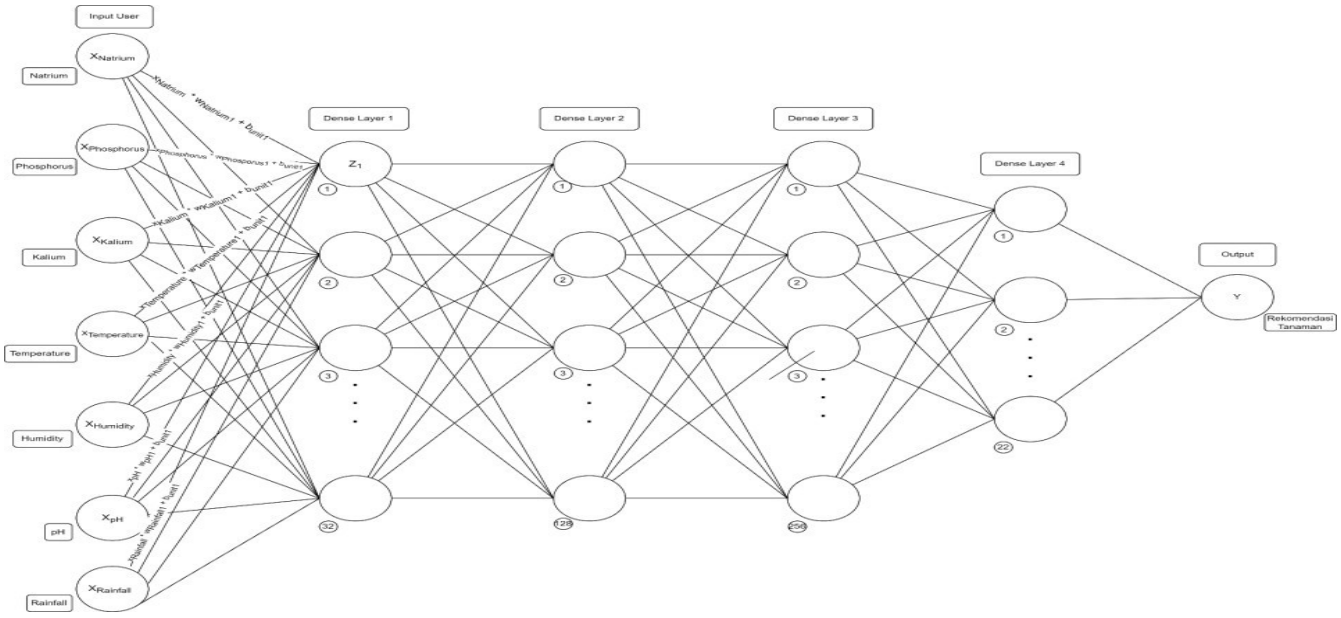


Fig. 1. The Proposed Model

The model consists of 7 inputs represented by X. X is the input matrix consisting of Sodium = x_1, Phosphate = x_2, Potassium = x_3, Temperature = x_4, Humidity = x_5, pH = x_6, and Rainfall = x_7. All these variables are normalized through MinMax Scaling.

Forward pass is a process that flows input data through each neuron in the dense layer to the output layer. Calculations for each layer (Z) using normalized inputs x_1, x_2, x_3, x_4, x_5, x_6, and x_7 and using weights for each input, namely w_1, w_1, w_3, w_4, w_5, w_6, and w_7 and also using b is biased. The activation function uses equation 1.

$$Z_1 = \sum_{n=1}^{n=7} (X_n W_n) + b. \quad (1)$$
$$= ((x_1 * w_1) + (x_2 * w_2) \dots (x_7 * w_7)) + b_{unit1}$$

Input in the form of 7 soil parameters in the artificial neural network to be calculated into prediction results in the form of a list of confidence models for each existing label (22 labels). The greatest confidence value of the model is used as the output after reverse label encoding is done. The expected final result is y which is the most suitable plant name for the soil with the 7 parameters that have been entered.

The developed model is evaluated using accuracy metrics. The accuracy tested uses two scenarios, namely the accuracy of the training data and the accuracy of the validation process carried out through the test dataset. If the accuracy is not satisfactory, then the model that has been developed is modified by some of its hyperparameters to learn more data patterns, so that the results of the training can be maximized.

III. System Design

Soil condition data obtained from the Crop Recommendation Dataset. The data is the result of augmentation of soil parameters chemically and physically. The dataset consists of 8 columns and 2200 data records. The process of classification of soil types based on suitable plants using multiclass classification ANN consists of 4 stages, namely (1) Data preprocessing, (2) Data division into training data, validation data, and test data, (3) Development of the ANN Model, and (4) Model Evaluation.

At the preprocessing stage, the data is normalized so that it is ready to be processed with ANN. Preprocessing is carried out in two stages, namely the encoding labeling process to change the string data type label to an integer with a range between 0 to 21 and the separation and labeling process to become X and y, the one-hot encoding process to change the label to a binary list. In this process there are 1,980 training data, 110 validation data, and 110 test data. Each data category is normalized with MinMax Scaler.

The dataset is split into training data, validation data, and test data using 3 scenarios based on the split ratio and the number of hidden layers as shown in Table 1.

Table 1. Scenario of data splitting

Scenario	Train Ratio	Validation Ratio	Test Ratio
1	70%	15%	15%
2	80%	10%	10%
3	90%	5%	5%

In the third scenario, the 2,200 data records are split into 1,980 training datasets used in ANN model training, 110 validation datasets to test the validity of the model in each epoch, and 110 test datasets to test the accuracy of the developed model. In the distribution of data, randomization was also using a predetermined seed, so that there was no change in the training data and test data even though the model was repeatedly trained.

Data is normalized using MinMaxScaler. MinMaxScaler is to rescale a variable into value between 0 and 1. Data normalization is a very helpful process for a better computation. Normalization is also able to increase the accuracy. Data normalization is the last step before the modelling process. Table 2 shows the first 10 rows of normalization results.

Table 2. The first 10 rows of normalization results

No	N	P	K	Temp.	Humidity	pH	Rainfall
1	90	42	43	20,890	82,002	6,5	202,9
2	85	58	41	21,770	80,320	7,038	226,656
3	60	55	44	23,004	82,321	7,840	263,964
4	74	35	40	26,491	80,158	6,980	242,864
5	78	42	42	20,130	81,605	7,628	262,717
6	69	37	42	23,058	83,370	7,073	251,055
7	69	55	38	22,709	82,639	5,701	271,325
8	94	53	40	20,278	82,894	5,719	241,974
9	89	54	38	24.5159	83.535	6.685	230.446
10	68	58	38	23.224	83.033	6.336	221.209

In this research, 3 models were developed based on different amounts of data and hidden layers. The first model uses 3 layers dense, the second model uses 4 layers dense, and the last model uses 5 layers dense. The architecture with the best results (accuracy, validation accuracy, test accuracy, combined with Precision, Recall, and F1 Score). Combined with 3 data splitting scenarios, there are 9 scenarios as shown in Table 3.

Table 1 Scenario of Data Splitting and Modelling

Scenario	Number of Layers	Train Ratio	Validation Ratio	Test Ratio
1	3	70%	15%	15%
2	4	80%	10%	10%
3	5	90%	5%	5%
4	3	70%	15%	15%
5	4	80%	10%	10%
6	5	90%	5%	5%
7	3	70%	15%	15%
8	4	80%	10%	10%
9	5	90%	5%	5%

IV. Results and Analysis

The first layer uses the ReLU activation function while the last layer uses the softmax activation function by 22 units, because the resulting output is 22 recommended plant species. In this study, RMSProp was also used for optimization. The accuracy value of each scenario in the modeling process is shown in Table 4.

Table 2 Accuracy Result of Each Scenario

Scenario	Number of Layers	Split Ratio	Train Acc	Validation Acc	Test Acc
1	3	70%:15%:15%	98.96%	98.48%	96.66%
2	4	70%:15%:15%	99.18%	98.18%	95.75%
3	5	70%:15%:15%	99.29%	98.79%	96.25%
4	3	80%:10%:10%	98.96%	98.18%	97.02%
5	4	80%:10%:10%	98.88%	97.97%	97.72%
6	5	80%:10%:10%	99.25%	96.36%	93.63%
7	3	90%:5%:5%	98.96%	98.79%	97.57%
8	4	90%:5%:5%	99.30%	99.24%	98.93%
9	5	90%:5%:5%	99.49%	97.27%	98.99%

Based on testing using the 9 scenarios, the accuracy in the 8th scenario is the best among all scenarios. This scenario uses 4 layers, split ratio of 90% training data, 5% validation data, and 5% test data. The training accuracy reaches 99.30%, while the validation accuracy drops slightly by 0.06% which indicates that the model is fit. The test accuracy is 98.93%. The training accuracy in the 9th scenario achieves the highest training accuracy, but slightly shows overfitting, because the validation accuracy drops by 2%. These results are also supported by the Precision, Recall, and F1 Scores from each scenario shown in Table 5.

Table 5. The Precision, Recall, and F1 Scores from Each Scenario

Scenario	Number of Layers	Split Ratio	Precision	Recall	F1 Score
1	3	70%:15%:15%	97.17%	97.31%	97.15%
2	4	70%:15%:15%	96.78%	96.55%	96.41%
3	5	70%:15%:15%	97.13%	96.56%	96.63%
4	3	80%:10%:10%	97.96%	98.37%	98.07%
5	4	80%:10%:10%	98.03%	98.49%	98.00%
6	5	80%:10%:10%	98.11%	98.26%	98.00%
7	3	90%:5%:5%	97.42%	97.68%	97.05%
8	4	90%:5%:5%	99.24%	99.49%	99.32%
9	5	90%:5%:5%	97.58%	97.58%	97.58%

Based on the test results using precision, recall, and F1 scores from the 9 scenarios tested, the 8th scenario also has the best score. The resulting score is more than 99%. Therefore, based on the accuracy test results in Table 4 and Table 5, the 8th architecture was chosen as a model for classifying soil conditions. The architecture detail of the 8th scenario is shown in the Figure 2.

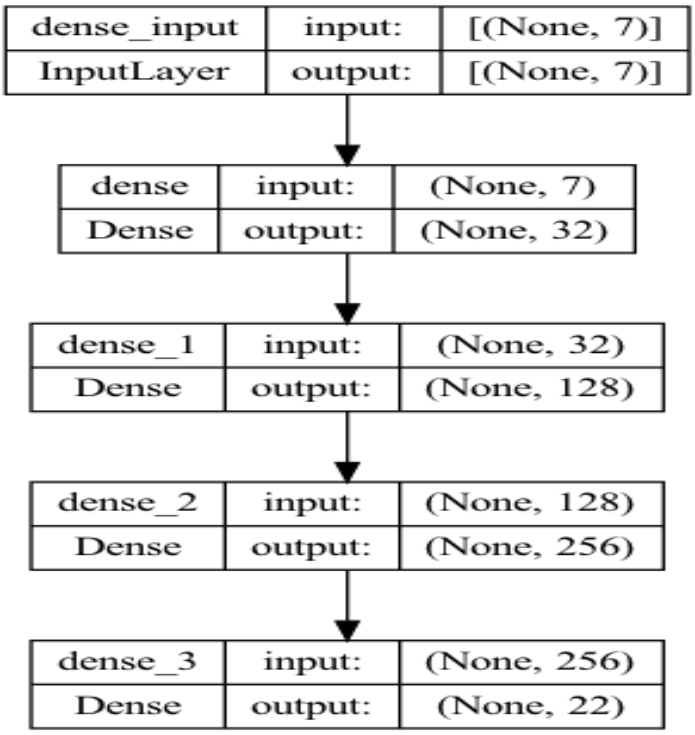


Fig. 2. ANN Model Architecture

The best model that has been developed in the modeling process is evaluated. In the evaluation section, accuracy model plots (Fig. 3) and lost model plots (Fig. 4) are carried out for up to more than 200 epochs.

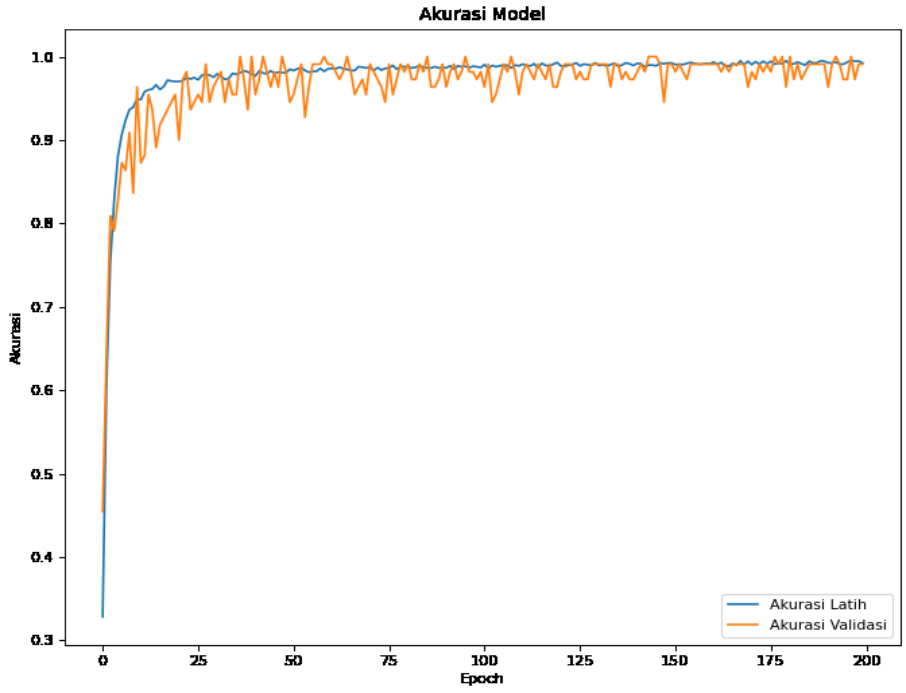


Fig. 3. Accuracy Model Plot

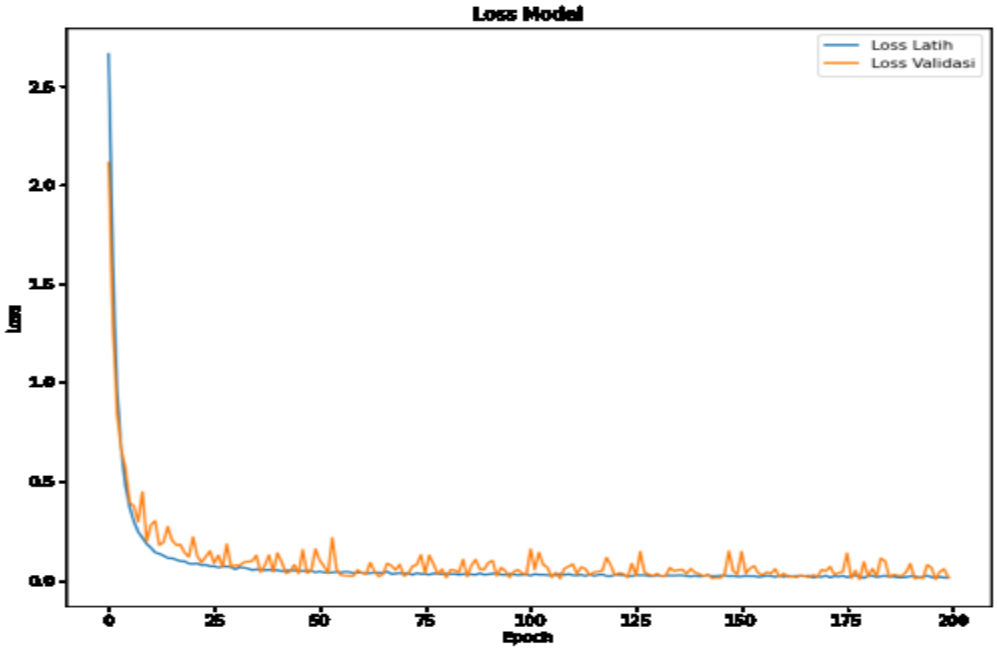


Fig. 4. Loss Model Plot

The model converges rapidly on the first 15 epochs. In these epochs, the accuracy of the model increases sharply. After the first 15 epochs, the growth of the accuracy decreases as the accuracy almost reaches maximum point of 100%. The loss plot also depicts the same result. The loss of the model decreases significantly on the first 15 epochs.

Model growth plot shows the indication of wellfit. This statement is able to be confirmed through the growth of the accuracy. The growth of the training accuracy is raising along with the validation accuracy. The training loss also decreases along with the validation loss.

The model is then tested to validate its accuracy. Fig. 5 shows the confusion matrix generated from the testing process.

```
([[5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 9, 0, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 9, 0, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0],
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4]])
```

Fig. 5. Confusion Matrix of Model

Fig. 5 shows that there is one false prediction in the model. Based on Fig. 5 it can be observed that the 13th label, mungbean, was falsely predicted from the 10th label, maize.

V. Conclusion

Process of classification of soil types based on suitable plants using multiclass classification ANN consists of 4 stages, namely data preprocessing, data splitting into training data, validation data, and test data, development of the ANN Model, and model evaluation. Parameters used in this research were the chemical conditions of the soil, namely levels of Nitrogen, Phosphorus and Potassium, as well as the physical condition of the soil which included temperature, humidity, pH and rainfall. After identifying the soil condition data pattern, it is used to classify soil types based on the appropriate plants. This research develops a model with 9 scenarios that vary in the ratio of data splitting and the number of layers used. Based on all trials conducted, the best scenario is the splitting of 90% training data, 5% validation data, and 5% test data with 4 layers. This model has a training accuracy of 99.30%, a validation accuracy of 99.24%, and a test accuracy of 98.93%. Model testing in this scenario is also the best with 99.24% precision, 99.49% recall, and 99.32% F1 score.

References

[1] B. Keulemans, W., Bylemans, D., De Coninck, *Farming without plant protection products: Can we grow without using herbicides, fungicides and insecticides?*, no. March. 2019.

[2] B. S. Adeleke and O. O. Babalola, "Oilseed crop sunflower (Helianthus annuus) as a source of food: Nutritional and health benefits," *Food Sci. Nutr.*, vol. 8, no. 9, pp. 4666–4684, 2020, doi: 10.1002/fsn3.1783.

[3] D. Serebrennikov, F. Thorne, Z. Kallas, and S. N. McCarthy, "Factors influencing adoption of sustainable farming practices in europe: A systemic review of empirical literature," *Sustain.*, vol. 12, no. 22, pp. 1–23, 2020, doi: 10.3390/su12229719.

[4] H. Mehbub *et al.*, "Tissue Culture in Ornamentals: Cultivation Factors, Propagation Techniques, and Its Application," *Plants*, vol. 11, no. 23, 2022, doi: 10.3390/plants11233208.

[5] H. Upadhyay *et al.*, "Exploration of Crucial Factors Involved in Plants Development Using the Fuzzy AHP Method," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/4279694.

[6] M. F. Seleiman *et al.*, "Drought stress impacts on plants and

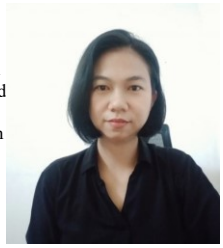
different approaches to alleviate its adverse effects," *Plants*, vol. 10, no. 2, pp. 1–25, 2021, doi: 10.3390/plants10020259.
[7] A. Tataridas, P. Kanatas, A. Chatzigeorgiou, S. Zannopoulos, and I. Travlos, "Sustainable Crop and Weed Management in the Era of the EU Green Deal: A Survival Guide," *Agronomy*, vol. 12, no. 3, pp. 1–23, 2022, doi: 10.3390/agronomy12030589.
[8] A. Javed, E. Ali, K. Binte Afzal, A. Osman, and D. S. Riaz, "Soil Fertility: Factors Affecting Soil Fertility, and Biodiversity Responsible for Soil Fertility," *Int. J. Plant, Anim. Environ. Sci.*, vol. 12, no. 01, pp. 21–33, 2022, doi: 10.26502/ijpaes.202129.
[9] N. M. Alzamel, E. M. M. Taha, A. A. A. Bakr, and N. Loutfy, "Effect of Organic and Inorganic Fertilizers on Soil Properties, Growth Yield, and Physiochemical Properties of Sunflower Seeds and Oils," *Sustain.*, vol. 14, no. 19, 2022, doi: 10.3390/su141912928.
[10] B. P. Akinde, A. O. Olakayode, D. J. Oyedele, and F. O. Tijani, "Selected physical and chemical properties of soil under different agricultural land-use types in Ile-Ife, Nigeria," *Heliyon*, vol. 6, no. 9, p. e05090, 2020, doi: 10.1016/j.heliyon.2020.e05090.
[11] R. E. Enescu, L. Dincă, M. Zup, Șerban Davidescu, and D. Vasile, "Assessment of Soil Physical and Chemical Properties among Urban and Peri-Urban Forests: A Case Study from Metropolitan Area of Brasov," *Forests*, vol. 13, no. 7, 2022, doi: 10.3390/f13071070.
[12] W. Food, *Soils for nutrition: state of the art*. 2022. doi: 10.4060/cc0900en.
[13] M. S. O'Donnell and D. J. Manier, "Spatial Estimates of Soil Moisture for Understanding Ecological Potential and Risk: A Case Study for Arid and Semi-Arid Ecosystems," *Land*, vol. 11, no. 10, 2022, doi: 10.3390/land11101856.
[14] T. Blesslin Sheeba *et al.*, "Machine Learning Algorithm for Soil Analysis and Classification of Micronutrients in IoT-Enabled Automated Farms," *J. Nanomater.*, vol. 2022, 2022, doi: 10.1155/2022/5343965.
[15] R. Thakur, "Recent Trends Of Machine Learning In Soil Classification : A Review," *Int. J. Comput. Eng. Res.*, vol. 08, no. 9, pp. 25–32, 2018.
[16] M. Uddin and M. R. Hassan, "A novel feature based algorithm for soil type classification," *Complex Intell. Syst.*, vol. 8, no. 4, pp. 3377–3393, 2022, doi: 10.1007/s40747-022-00682-0.
[17] J. Guo, K. Wang, and S. Jin, "Mapping of Soil pH Based on SVM-RFE Feature Selection Algorithm," *Agronomy*, vol. 12, no. 11, p. 2742, 2022, doi: 10.3390/agronomy12112742.
[18] N. Carvalho, L. C. Barbosa, H. Bellinaso, C. Danilo, and D. Mello, "Soil Erosion Satellite-Based Estimation in Cropland for Soil Conservation," pp. 1–24, 2023.
[19] A. Ingle, "Crop Recommendation Dataset." https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

Prof. Dr. Saiful Bukhori
e-mail: saiful.ilkom@unej.ac.id

Professor of Artificial Intelligence at the at the University of Jember. Research interest: artificial intelligence with research experience including: Development of graph mining for predicting payment system networks in RTGS based on Clearing Houses, Intelligent Agent for Serious Game of RTGS using Forest Fire Model, Parrondo's paradox based strategies in the serious game of RTGS using Forest Fire Model, Serious Game Supply Chain Management Agribusiness as a Production Planning using Cournot Model, Serious Game Relationship Between Socio-Economic and Territorial Conditions, etc.
https://orcid.org/0000-0002-2527-1080

Nova El Maidah, S.Si., M.Cs
e-mail: nova.pssi@unej.ac.id

Research interests: Artificial Intelligence with research experience including: Design of Fuzzy Controller Based on ATmega32 Microcontroller as Temperature and Humidity Controller, Comparison of Genetic Algorithm with Greedy Algorithm for Shortest Route Search, A Fuzzy Control System for Temperature and Humidity Warehouse Control



Ivan Budianto
e-mail: ivanbudianto0603@gmail.com

Research interests: Artificial Intelligence with research experience including: Classification of Soil Conditions Based on Recommendations for Agricultural and Plantation Crops