Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 22, No. 1, November 2022, pp. 97∼104
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v22i1.1880

The Improvement of Artificial Neural Network Accuracy Using Principle Component Analysis Approach

Arief Hermawan[1], Adityo Permana Wibowo[2], Akmal Setiawan Wijaya[3]
[1,2]Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
[3]Universitas Islam Indonesia, Yogyakarta, Indonesia

Article Info

Article history:

Received May 26, 2022
Revised June 02, 2022
Accepted November 04, 2022

Keywords:

Principle Component Analysis
Neural Network
Mushroom Classification

ABSTRACT

An important problem in a classification system is how to get good accuracy results. A way to increase the accuracy of a classifier system is to improve the number of input data attributes. Improving the number of input data attributes can be done using the Principal Component Analysis (PCA) method. The aim of this research is to reduce the number of input data attributes to increase the accuracy in a mushroom classification system. The research method used in this study started from collecting datasets from Kaggle.com related to mushroom-classification, then the data visualization process was carried out using pie charts then a dimension reduction process was carried out to reduce the number of variables using the PCA method. The next step is the training and testing of the artificial neural network. The architecture of artificial neural network used is backward error propagation with the number of hidden layers as much as 2 layers with the number of cells as many as 3 and 2. The training data used is 80%, while the testing data is 20%. Based on the test results, obtained an accuracy of 100% with 150,000 iterations and using 11 input variables from 22 existing input variables. By adding Principal Component Analysis part of the development that can improve the accuracy and performance of Artificial Neural Networks.

Corresponding Author:

Adityo Permana Wibowo,
Information System of Diploma Program'
Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia,
Email: adityopw@uty.ac.id

How to Cite:A. Hermawan, A. Wibowo, and A. Setiawan Wijaya, The Improvement of Artificial Neural Network Accuracy Using Principle Component Analysis Approach, MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 1, Nov. 2022.

1. INTRODUCTION

Mushroom is one of the types of plants that can easily grow in Indonesia because Indonesia is a tropical country with enough rainfall and sunlight [1]. Some kinds of edible mushrooms even contain important substances for medication [2, 3]. However, there are also many of inedible mushrooms that are poisonous [4–6]. Some cases of death due to mushrooms poisoning also happened in Indonesia in the last 10 years. There were 76 cases with total of 550 victims, 9 of whom died [7]. With these many dangers caused by poisonous mushrooms, we need a system to classify whether a mushroom is poisonous or not. Artificial Neural Network is one of the classification methods [8]. One evidence that artificial neural network can classify the families of mushrooms was found in a study in 2018 [9], where the classification accuracy of the system was 92%.

In 2019, another study to classify poisonous and non-poisonous mushrooms using Back-propagation Artificial Neural Network started to be done and resulting in the best accuracy of 99.25% [10]. The study used 22 input variables. The architecture of artificial neural network used in the study consisted of 5 layers; 1 input layer, 3 hidden layers and 1 output layer. The network could recognize the pattern well at 161.501 of iteration [10]. Other study related to mushroom classification that has been done resulted in the accuracy value at 98%. Those results were obtained using ANN architecture which consisted of 21 cells in input layer, 1 cell in hidden layer, and 3 cells in output layer [11]. Other than that, there is a study of mushroom classification using image processing based on its genus using the Convolutional Neural Network (CNN) method. The study resulted in accuracy of 62% using Adam optimizer and 100 iterations, the comparison of training and testing data at 80:20, kernel size at 3x3 and a learning rule at 0,001 [12]. Another study which uses image processing to detect mushrooms also has also been done using Backpropagation Neural Network. The study processed the mushrooms based on the color segmentation, edge detection, and contours, resulting in 80% of accuracy [13]. CNN were also used to classify spices with VGG 16 architecture modification with 10 layers consist of 7 convolution layers and 3 classifier layers. The accuracy result was 81% [14]. So far, there are no studies about mushroom classification that are able to get the perfect accuracy result of 100%.

Some techniques to increase the classification accuracy of Artificial Neural Network are the application of dimension reduction [15], one of which is Principal Component Analysis (PCA) [16–18]. Reducing the number of dimensions can make the classification process faster and more accurate. Dimension reduction can speed up the training time of artificial neural network and increase the accuracy of the artificial neural network, proven to be able to produce an accuracy value of 81,5% with 5000 epoch in classifying college students stress levels [19]. Principal component analysis is one of the methods used to perform dimension reduction. While the use of principal component analysis can produce a greater accuracy value of 96.4% in processing Pap Smear images using a fold validation value of 5 performance [20]. Based on several previous studies, no one has yet produced an accuracy value of 100% in the use of Principal Component Analysis. This study tries to prove that using Principal Component Analysis can produce an accuracy value of 100%.

The organization of this article, after the introduction then start the second subsection which is the research method that discusses the steps of this study that start from taking datasets to producing the accuracy values. The third subsection discusses about the results and analysis that explain the results of calculations starting from ANN modeling, the construction of ANN architecture, to testing the ANN model. The last part of this article discusses about the conclusion of the overall study.

2. RESEARCH METHOD

The method used in this study can be seen in Figure 1. The study was began by downloading dataset related to Mushroom-Classification in kaggle.com and doing the initial processing then was followed by data visualization process to determine the composition of the comparison between poisonous and non-poisonous mushrooms. Next was dimension reduction process by reducing the number of input variables using PCA method. After the data was ready, it was then generated using an artificial neural network model to produce the best architecture that produces the highest accuracy value. The process of training and testing of the Artificial Neural Network was by reducing the number of variables which was around 8 to 21 variables. The result of it was shown in the form of accuracy value. The research method can be seen in Figure 1.

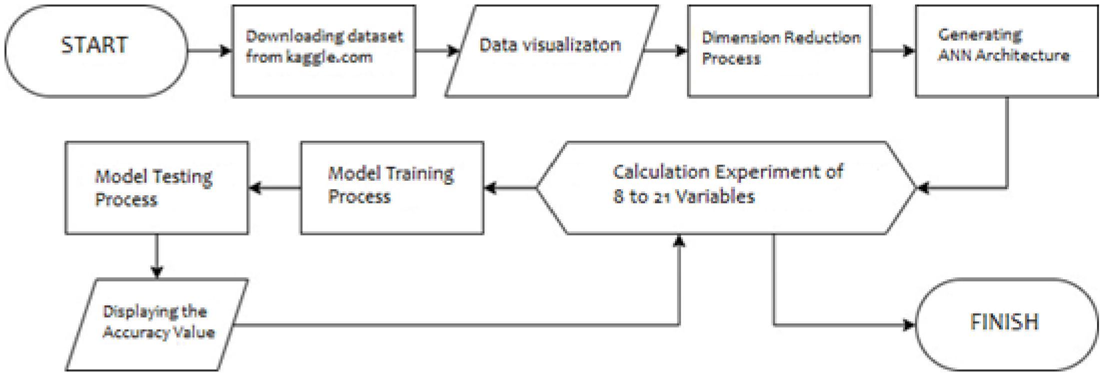

Figure 1. Research Method

More about the research method will be explained below.

2.1. Dataset

The data was taken from https://www.kaggle.com/uciml/mushroom-classification. There were total of 8124 records with 22 input attributes consist of cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, and habitat. Those were input attributes that determined whether a mushroom was poisonous or not. Meanwhile, the only one output attribute used in this study was class, namely class of poisonous or non-poisonous mushroom. The summary of the dataset used in this study is shown in Table1.

Table 1. Pembagian data untuk Training dan Testing

No.	**Class**	**Cap Shape**	**Cap Surface**	**Cap Color**	**...**	**Ring type**	**Spore Print Color**	**Population**	**Habitat**
1	p	x	s	n	...	p	k	s	u
2	e	x	s	y	...	p	n	n	g
3	e	b	s	w	...	p	n	n	m
4	p	x	y	w	...	p	k	s	u
5	e	x	s	g	...	e	n	a	g
6	e	x	y	y	...	p	k	n	g
7	e	b	s	w	...	p	k	n	m
8	e	b	y	w	...	p	n	s	m
9	p	x	y	w	...	p	k	v	g
10	e	b	s	y	...	p	k	s	m
...	...	...	...	...	...	...	...	...	...
8115	p	f	y	c	...	n	w	c	d
8116	e	x	s	n	...	p	o	v	l
8117	p	k	y	n	...	e	w	v	l
8118	p	k	s	e	...	e	w	v	d
8119	p	k	y	n	...	e	w	v	d
8120	e	k	s	n	...	p	b	c	l
8121	e	x	s	n	...	p	b	v	l
8122	e	f	s	n	...	p	b	c	l
8123	p	k	y	n	...	e	w	v	l
8124	e	x	s	n	...	p	o	c	l

The next initial processing consists of:

1. Removing inclomplete data
   There were total of 8124 records in dataset from Kaggle. However, not all of those data were complete in each column, some were empty (null). In order to produce maximum accuracy, it was necessary to select the incomplete data first. The dataset

selection was done manually by removing data in the incomplete columns. The results of the selection left a dataset of 5644 records.

2. Turning nominal data into numerical
The default dataset was in alphabetical form, while the process using Rapidminer needed numerical data. Therefore, to make the data processing easier, the data needed to be converted into numerical using Excel 365 software. Turning the data type from nominal to numerical was done by giving 0 minimum score and 1 maximum score. If there were more than two categories, the other categories were scored between 0 and 1. The results of converting dataset into numerical can be seen in Table 2.

Table 2. The Converted Dataset of Mushroom Classification in Numerical Form

No.	Class	Cap Shape	Cap Surface	Cap Color	...	Ring type	Spore Print Color	Population	Habitat
1	0	0	0	0	...	0	0	0	0
2	1	0	0	0.14	...	0	0.2	0.2	0.2
3	1	0.2	0	0.28	...	0	0.2	0.2	0.4
4	0	0	0.3	0.28	...	0	0	0	0
5	1	0	0	0.42	...	0.3	0.2	0.4	0.2
6	1	0	0.3	0.14	...	0	0	0.2	0.2
7	1	0.2	0	0.28	...	0	0	0.2	0.4
8	1	0.2	0.3	0.28	...	0	0.2	0	0.4
9	0	0	0.3	0.28	...	0	0	0.6	0.2
10	1	0.2	0	0.14	...	0	0	0	0.4
...	...	...	...	...	...	...	...	...	...
5635	1	0	0.3	0.98	...	0	1	0.6	0.8
5636	1	0.6	0	0.98	...	0	1	0.6	0.8
5637	1	0	0.3	0	...	0	1	0.8	0.8
5638	0	0.8	0.3	0.98	...	0.9	1	1	0.6
5639	1	0.6	0	0	...	0	1	0.8	0.6
5640	1	0.2	0.3	0	...	0	1	0.8	0.8
5641	1	0	0.3	0	...	0	1	0.8	0.8
5642	1	0	0.3	0.42	...	0	1	0.8	0.8
5643	0	0	0.3	0.98	...	0.9	1	1	0.6
5644	0	0.6	0.3	0.98	...	0.9	1	1	0.6

2.2. Data Visualization

After the initial processing, the next step was to do the data visualization. Data visualization aimed to see the comparison of the total mushrooms categorized as poisonous and non-poisonous. The software used to visualize the data was Excel 365. This process was done by using pie chart.

2.3. Dimension Reduction

A Dimension Reduction Technique used in this study was Principal Component Analysis (PCA). The concept of PCA was to compare all variables with similar characteristics, then one of them was chosen between those similar variables. This process of comparing 22 variables produced 8 to 13 variables. An application used to generate those variables was Rapidminer. The flow of the principal analysis component used in this study is as shown in Figure 2.

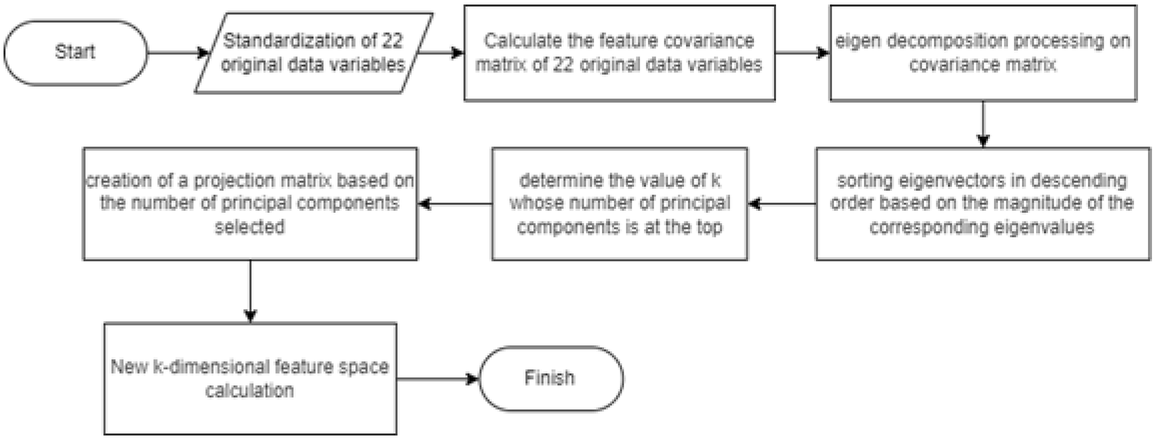

Figure 2. The Flow of The Principal Analysis Component Used

2.4. ANN Architecture Development

The architecture of artificial neural network in this study consisted of 4 layers namely 1 input layer, 2 hidden layers, and 1 output layer. Input layer consisted of several cells according to the result of the principal component analysis. The first hidden layer consisted of 3 cells and the second one of 2 cells. Output layer consisted of 2 cells. The architecture of ANN can be seen in Figure 3.

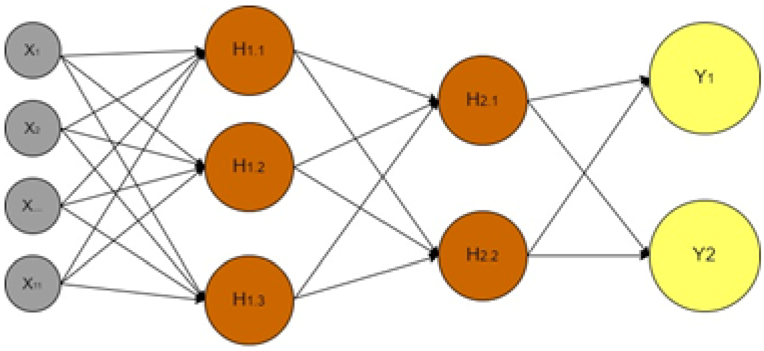

Figure 3. ANN Architecture in This Study

2.5. Training

The training aimed to create an Artificial Neural Network model in the classification process of poisonous mushroom. The modeling was based on some provisions that had been made and produced accuracy as shown in Figure 4. 80% of 5705 dataset records were used in the training. Rapidminer was used in this Artificial Neural Network training/modeling.

2.6. Testing and Evaluation

In this last step, the goal of testing and evaluation was to find out the performance of the Artificial Neural Network model. The testing used 20% of the training data. Meanwhile, the evaluation was done using Rapidminer with Split Validation operator.

3. RESULT AND ANALYSIS

3.1. Dataset

Dataset taken were 8124 records. Those consisted of 22 characteristics of poisonous and non-poisonous mushrooms which were cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, and habitat. The obtained data were then processed initially by removing the incomplete data. By this step, the remaining data were 5705 records. Changing the data form from nominal into numerical was done by giving 0 as minimum score and 1 as maximum score. For any data with more than two categories, the other category was scored between 0 and 1.

3.2. Data Visualization

The results of the data selection showed the percentage ratio of mushroom P (Poisonous) and mushroom E (Edible). The poisonous mushroom was 38% while non-poisonous mushroom was 62%. The visualization data can be seen in Figure 4.

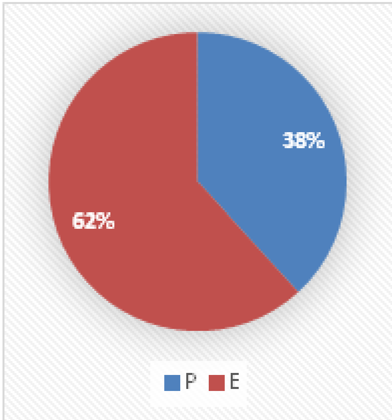

Figure 4. Data Visualization Class

Based on Figure 4, it can be concluded that the comparison between the categories of poisonous and non-poisonous mushrooms were still not ideal. The comparison should be around 50% for to be considered as ideal.

3.3. Dimension Reduction

Dimension reduction was done using Principal Component Analysis (PCA). This aimed to extract a number of variables, resulting in some low-dimension variables. This study conducted an experiment that regulated the use of some variables that later would be processed. The experiment was started by using 8 to 21 variables. Figure 6 shows the data of 11 variables. The results processing principal component analysis with 11 variables can be seen in Table 3.

Table 3. Principal Component Analysis Result with 11 Variables

Class	Pc_1	Pc_2	Pc_3	Pc_4	Pc_5	Pc_6	Pc_7	Pc_8	Pc_9	Pc_10	Pc_11
1	0.292	0.088	0.103	0.250	0.844	0.916	0.541	0.509	0.446	0.648	0.555
1	0.413	0.430	0.694	0.652	0.464	0.783	0.974	0.275	0.439	0.701	0.430
1	0.260	0.089	0.070	0.331	0.348	0.802	0.557	0.318	0.387	0.553	0.572
0	0.120	0.546	0.977	0.328	0.617	0.739	0.597	0.627	0.397	0.714	0.432
0	0.143	0.550	0.954	0.348	0.608	0.657	0.720	0.495	0.459	0.699	0.654
0	0.135	0.530	0.997	0.292	0.568	0.692	0.674	0.632	0.304	0.732	0.540
1	0.173	0.694	0.713	0.055	0.502	0.694	0.364	0.404	0.344	0.553	0.421

3.4. The Architecture of Artificial Neural Network (ANN)

The artificial neural network used in this study consisted of 4 layers; 1 input layer, 2 hidden layers, and 1 output layer. The input layer consisted of some nodes according to the principal component analysis result. The first hidden layer consisted of 3 nodes and the second hidden layer consisted of 2 nodes. There were 2 nodes in the output layer.

3.5. Training and Testing

After the architecture of ANN was formed, the next step was to do training and testing of the model to know the performance and the optimization. The training and testing was done using Python 3.9 Programming Language.

Artificial Neural Network was trained using the maximum iteration at 150.000. Then, it was tested using split validation model. The testing was done by calculating the accuracy value based on the total variables with a scenario of 6 predetermined variables. The testing result can be seen in Table 4.

Table 4. The Testing Result of Artificial Neural Network

Variables	8	9	10	11	12	13
Accuracy	97,9	99,2	99,4	100	97,5	82,7

Table 1 shows that artificial neural network could classify with perfect accuracy of 100% for the lowest number of 11 variables. Based on the result, it can be concluded that this classification result is better than study [10]. This improvement was done by reducing the input variables using the principal component analysis method. Several input variables used in study [10] became confounding variables in the process of mushroom classification. This finding is in line with previous studies [19, 20]. Comparison of previous studies can be seen in Table 5.

Table 5. Comparison of previous Studies

Author	Method	Dataset	Accuracy
[19]	Artificial Neural Network Backpropagation with dimensional reduction	Stress resistance among college students	81,5%
[20]	Artificial Neural Network Backpropagation with principal component analysis	Pap Smear Image for classification Cervix Dysplasia	96,4%
Proposed Method	Artificial Neural Network Backpropagation with principal component analysis	Mushroom	100%

4. CONCLUSION

The Artificial Neural Network processing produced perfect accuracy value at 100% with 11 input variables from 22 existing input variables. The architecture of ANN used in this study consisted of 4 layers; 1 input layer, 2 hidden layers, and 1 output layer. The total iteration used was 150.000. This finding showed that the developed artificial neural network could classify better than the previous studies. Further study can be done to reduce the computational load. This can be done by using dimensional reduction technique with variable selection method. In order to achieve reducing computational load, the total of variables should be lower than 11 and the accuracy value should be 100%.

5. ACKNOWLEDGEMENTS

The Acknowledgments section is optional. Research sources can be included in this section.

6. DECLARATIONS

AUTHOR CONTIBUTION

FUNDING STATEMENT

COMPETING INTEREST

REFERENCES

[1] M. T. Sibero, I. P. Putra, and R. Murwani, "Deskripsi dan Potensi Jamur Makro Asal Hutan Adat Penembahen, Desa Juhar, Kabupaten Tanah Karo, Sumatra Utara," *Jurnal Mikologi Indonesia*, vol. 5, no. 1, pp. 48–54, 2021.

[2] I. P. Putra, “Note on Macro Fungi on Belitong Island: Description and Potential,” *BIOEDUSCIENCE: Jurnal Pendidikan Biologi dan Sains*, vol. 4, no. 1, pp. 11–20, 2020.

[3] ——, “Studi Taksonomi dan Potensi Beberapa Jamur Liar di Pulau Belitong,” *Justek : Jurnal Sains dan Teknologi*, vol. 3, no. 1, pp. 24–31, 2020.

[4] ——, “Kasus Keracunan Inocybe sp. di Indonesia,” in *Prosiding Seminar Nasional Biologi di Era Pandemi COVID-19*, 2020, pp. 148–153.

[5] E. I. Haksoro and A. Setiawan, “Pengenalan Jamur yang Dapat Dikonsumsi Menggunakan Metode Transfer Learning pada Convolutional Neural Network,” *Jurnal ELTIKOM : Jurnal Teknik Elektro, Teknologi Informasi dan Komputer*, vol. 5, no. 2, pp. 81–91, may 2022.

[6] R. Hayami, S. Soni, and I. Gunawan, “Klasifikasi Jamur Menggunakan Algoritma Naïve Bayes,” *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 1, pp. 28–33, 2022.

[7] I. P. Putra, “Kasus-Kasus Keracunan Jamur Liar di Indonesia,” *Jurnal Ekologi Kesehatan*, vol. 20, no. 3, pp. 215–230, mar 2022.

[8] S. K. Verma and M. Dutta, “Mushroom Classification using ANN and ANFIS Algorithm,” *IOSR Journal of Engineering (IOSRJEN)*, vol. 08, no. 01, pp. 94–100, 2018.

[9] J. U. Lidasan and M. P. Tagacay, “Mushroom Recognition using Neural Network,” *IJCSI International Journal of Computer Science Issues*, vol. 15, no. 5, pp. 52–57, 2018.

[10] E. S. Alkronz, K. A. Moghayer, M. Meimeh, M. Gazzaz, B. S. Abu-Nasser, and S. S. Abu-Naser, “Prediction of Whether Mushroom is Edible or Poisonous Using Back-Propagation Neural Network,” *International Journal of Academic and Applied Research*, vol. 3, no. 2, pp. 1–8, 2019.

[11] R. Hanseliani, “Klasifikasi Berbagai Jenis Jamur Layak Konsumsi dengan Metode Backpropagation,” Ph.D. dissertation, 2019.

[12] O. N. Putri, “Implementasi Metode CNN dalam Klasifikasi Gambar Jamur pada Analisis Image Processing (Studi Kasus: Gambar Jamur dengan Genus Agaricus dan Amanita),” Ph.D. dissertation, 2020.

[13] S. Enggari, A. Ramadhanu, and H. Marfalino, “Peningkatan Digital Image Processing dalam Mendeskripsikan Tumbuhan Jamur dengan Segmentasi Warna, Deteksi Tepi dan Kontur,” *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 4, no. 1, pp. 70–75, 2022.

[14] E. Tanuwijaya and A. Roseanne, “Modifikasi Arsitektur VGG16 untuk Klasifikasi Citra Digital Rempah-Rempah Indonesia,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 1, pp. 189–196, nov 2021.

[15] A. S. Ritonga and I. Muhandhis, “Teknik Data Mining untuk Mengklasifikasikan Data Ulasan Destinasi Wisata Menggunakan Reduksi Data Principal Component Analysis (PCA),” *Jurnal Ilmiah Edutic*, vol. 7, no. 2, pp. 124–133, 2021.

[16] A. L. Unihehu and I. Suharjo, “The Klasifikasi Jenis Ikan Berbasis Jaringan Saraf Tiruan Menggunakan Algoritma Principal Component Analysis (PCA),” *Jurnal Ilmiah Ilmu Komputer*, vol. 7, no. 2, pp. 27–32, 2021.

[17] H. Harizahayu, “Pengenalan Ekspresi Raut Wajah Berbasis Jaringan Saraf Tiruan Backpropagation dengan Metode Principal Component Analysis,” *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 15, no. 1, pp. 037–046, mar 2021.

[18] D. Frenza and R. Mukhaiyar, “Aplikasi Pengenalan Wajah dengan Metode Adaptive Resonance Theory (ART),” *Ranah Research : Journal of Multidisciplinary Research and Development*, vol. 3, no. 3, pp. 35–42, 2021.

[19] A. Hermawan and D. Avianto, “The Implementation of Neural Network on Determining The Determinant Factors Towards Students’ Stress Resistance,” *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 3, pp. 129–133, sep 2017.

[20] K. A. Widagdo, K. Adi, and R. Gernowo, “Kombinasi Feature Selection Fisher Score dan Principal Component Analysis (PCA) untuk Klasifikasi Cervix Dysplasia,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 3, pp. 565–572, 2020.