Handwriting Learning Vector Quantization / Mobile Application / Extraction Feature

First Author[1,2], Next Author[1] & Last Author[2] (Use full name for all Author(s))

[1]Author Address
[2]Author Address
Email: Email Address

Novelty: The novelty of previous research is the use of feature extraction in java android as well as passing the value of extraction features from java android to python via http url connection Ngrok.

Highlight:

- the learning vector quantization model created has an accuracy of 67.5%.
- the Skewness and relative smoothness features are the most difficult to apply to distinguish document ownership.
- the system can be used to help handwriting recognition to find out the owner of the handwriting quickly and easily.

Abstract. Today's technological developments increasingly lead us towards digital and mobile. One of the usual manual work that humans do that can be facilitated by mobile application technology is handwriting recognition to find out the owner of the handwriting. One of the machine learning methods that can do handwriting recognition is Learning Vector Quantization, Learning Vector Quantization is one of the methods of Artificial Neural Networks (ANN). In this study the researchers aimed to build a handwriting recognition system using the Learning Vector Quantization method in mobile applications. using feature extraction as a basic step in interpreting and classifying images. The results obtained from testing prediction learning vector quantization of 16 new data with 80 times the total test. The result is that 54 data are correct and 26 data are false, so that the accuracy is 67.5%, with learning rate = 0.005, alpha value = 0.05, and iteration = 100.

Keywords: *Android, Feature Extraction, Handwriting, Learning Vector Quantization, Mobile Application.*

1 Introduction

The term Industry 4.0 refers to the fourth industrial revolution marked by growing trends in the fields of automation, Internet of Things (IoT), Big Data, and Cloud Computing technology. Like the earlier versions of steam, electricity

Received _____, Revised _____, Accepted for publication _____

and computers in the past, the integration of these technologies will transform future industries into smarter industries. Today's technological developments increasingly lead us towards digital and mobile. Equipment that used to be only in the form of human imagination, now one by one can be realized. Modern equipment specifically for mobile phones, personal tablet computers, etc., with a variety of existing mobile applications.

One of the usual manual work that humans do that can be facilitated by mobile application technology is handwriting recognition to find out the owner of the handwriting. the handwriting recognition technique that is done manually is comparing one by one the handwritten file that has been saved with the handwritten file that you want to recognize, this technique usually uses the senses of human sight. of course to do it requires a lot of time and effort when done manually

Handwriting recognition can be done by machine learning/data mining, so the user only needs to enter the handwriting data as training data, and check the new handwriting from the results of the training model that has been made.

One of the machine learning methods that can do handwriting recognition is learning vector quantization, Learning Vector Quantization is one of the methods of Artificial Neural Networks (ANN). Principle is to reduce its neighbor nodes so that ultimately there is only one node is selected, and then calculating the minimum distance difference. Learning Vector Quantization will automatically learn to classify input vectors in the competitive layer. The resulting class is based on the distance of the vector. If there are two vectors that have a distance that is close enough or close to the same then the two vectors are grouped into the same class.

Several Related Works have been conducted before, Satia Budhi conducted research on Handwritten Javanese Character Recognition Using Several Artificial Neural Network Methods, the results of the study found the combination of the Chi2 method and the backpropagation neural network method performed better than the evolutionary neural network method with 1 layer or 2 layers for Javanese character recognition. The level of recognition accuracy reaches 98% for data that has been previously trained and 73% for data that has not been previously trained [1].

Jasril has conducted research on Learning Vector Quantization 3 (LVQ3) and Spatial Fuzzy CMeans (SFCM) for Beef and Pork Image Classification, the research results obtained are the application of Spatial Fuzzy C-Means in image segmentation and several other processes such as cropping area objects, feature extraction HSV color and GLCM texture feature extraction of meat object images and LVQ3 classification can recognize beef and pork images with the highest percentage of accuracy values 91.67% [2].

Shamim has conducted research on Handwritten Digit Recognition using Machine Learning Algorithms, obtained the results of research that is in any recognition process, the important problem is to address the feature extraction and correct classification approaches The proposed algorithm tries to address both the factors and well in terms of accuracy and time complexity. The overall highest accuracy 90.37% is achieved in the recognition process by Multilayer Perceptron [3].

Haviluddin has conducted research on Handwriting Character Recognition using Vector Quantization Technique. This research explores the processing stages of Learning Vector Quantization (LVQ) to recognize Lontara Bugis Makassar script and to explain its accuracy. The LVQ test results obtained an accuracy rate of 66.66%. The most optimal network architecture variant in the

recognition process is a variation of the learning rate of 0.02, a maximum epoch of 5000 and a hidden layer of 90 neurons which is the result of recognition based on feature 8 [4].

Rasika R. Janrao has conducted research on Handwritten English Character Recognition using LVQ and KNN. This study makes a handwritten character recognition system using soft computing methods. using two datasets, the first is the database itself which consists of 26 letters, 10 numbers and 5 special characters written by various people and the second is the CEDAR standard database. the test results of this study LVQ has an accuracy of 77.80% and KNN has an accuracy of 100% [5].

The difference that can be observed from this research with previous research is the use of new methods and the use of mobile applications. So in this study the researchers aimed to build a handwriting recognition system using the Learning Vector Quantization method in mobile applications.

2 Methodology

2.1 Learning Vector Quantization

Learning Vector Quantization (LVQ) is a training method for learning in supervised learning layers with single layer network architecture. The classes obtained as a result of this competitive layer only depend on the distance between the input vectors. If two input vectors approach the same, the competitive layer will put both input vectors into the same class. LVQ is a method of classifying patterns of each output unit representing a particular category or class (several output units should be used for each class). The The advantage of the LVQ method is its ability to provide training to competitive layers so that it can automatically classify the given input vectors.

The steps in the LVQ training algorithm consist of :

1. Initial weight initialization (W) and LVQ parameters, that is maxEpoch, α, decα and minα.
2. Enter input data (X) and target class (T).
3. Set the initial conditions: epoch = 0
4. Do it if: (epoch <maxEpoch) and ($\alpha \geq \min\alpha$).
   a. epoch = epoch+1.
   b. Determine J such that $\| Xi-Wj \|$ is minimal using the calculation of the euclidian distance formula. $D(j) = \sum (Wij - x\ i)^2$
   c. Correct Wj with the following provisions:

      If T = Cj then

      $Wj(t + 1) = wj\ (t)\ \alpha\ (t)[x(t) – wj]$

      If $T \neq Cj$ then

      $Wj(t + 1) = wj\ (t) + \alpha\ (t)[x(t)– wj(t)]$

   d. Reduce the value of α by:

      $\alpha = \alpha - \alpha * Dec\alpha$
5. Stop condition test with optimal weight output[6].

2.2 Extraction Feature

Feature Extraction is the process of transforming input data into a feature set to retrieve relevant information from input data with the aim of retrieving a minimal representation of the input data[7]. Feature extraction (or sometime called by indexing) is a basic step in conducting an image interpretation and classification. There are many ways to extract the feature, depends on the data. In image data, features that can be extracted are color, shape, and texture. And in this study, researcher uses texture. Texture is an intrinsic character of image which related with roughness level, granulation, and regularity of structural arrangement of pixels. Texture is defined as spatial distribution of greylevel in a set of

neighbouring pixels . Image feature extraction based on texture in orde one can use statistical method, ie by looking at the greylevel distribution statistic on the image histogram. From the histogram values, it can calculate feature parameters :

1. Variance (σ^2)

   Indicates element variations on the histogram of an image. Variance value is used by researchers to assess the extent of word variaton which present in each handwritten document. The greater the value of variance, the more varied the existing pattern in writing. The formula is :

$$\sigma^2 = \sum_n (f_n - \mu)^2 P(f_n)$$

   Where is an average value of pixels on image, is a value of grey intensity, ($p(f_n)$ is a value of histogram (intensity occurrence probability in document image) and is σ^2a value of variance.

2. Skewness (α_3)

   Skewness will indicate the relative historic level of the histogram curve of an image. In this research, value of skewness will be used to assess the inclination of the stroke direction of a post on each document. If the value is close to 0, the direction of the skew is symmetric. The formula is:

$$\alpha^3 = \frac{1}{\alpha^3} \sum_n (f_n - \mu)^3 P(f_n)$$

3. Entropy (H)

   Entropy will indicate irregularity level of a pattern, and this is also will be used in research on handwritten document. The

higher value of entropy, the variations and information contained in the writing pattern more and more irregularity.

$$H = -\sum_{n} p(f_n).Log_2 P(f_n)$$

4. Relative Smoothness (R)

Relative smoothness is a value that will indicate the relative degree of smoothness of shape and pattern of an image, and the researcher would also use it to determine how fine or rough the handwriting strokes of each document are. The higher the value of relative smoothness, the smoother the strokes pattern and the faster the writing movement.

$$R = 1 - \frac{1}{(1 + \sigma^2)}$$

In this research, another feature added is the mean feature, which will indicate the exact feature value to represent the entire representation of handwriting patterns[8].

3 Results and Discussion

3.1 Collecting Data

Collecting data in this research is done through direct searches of a number of resources (randomly close person with researcher) or a certain time interval. They are recorded in a book in white background, with intervals of three days in a row one time, and intervals of one week one time, and intervals of one month one time. This is done to see if there's any change in the shape of a person's handwriting in a certain period[9]. So that the number of documents collected for each resource amount to

two documents. The total of data that has been collected is 34 from 17 authors.

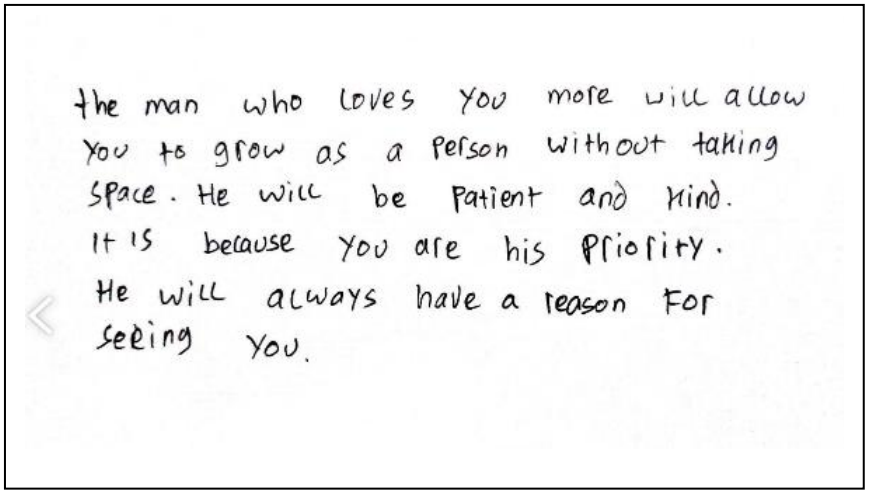

Figure 1 Handwritten Sample.

3.2 Extraction of datasets by java on android

The system is built using the java android language using the Android Studio IDE. The data used in this study were handwritten images (500x200 pixel) in a greyscale format in jpg type. The data will be taken from several feature values including Variance, Skewness, Entropy, and Relative Smoothness. The value of the feature that has been taken will be stored in the SQLite database which is used as a reference to determine the document owner against the data examiner. The difference in the feature interval will be automatically converted by the system to a value of 0 to 1. If the feature produces a very small value so that it can produce a NaN value, it will be converted to a value of 0. in this study perform feature extraction from the 34 handwritten image data, the results are shown in Table 1.

Table 1 Results of feature extraction

id	name	greyscale	variance	Entropy	Skewness	Relative Smoothness	Energy	Contrast
1	Steven	241.101	1.05153e+06	1734	0	1	110471	241.101
2	Steven	241.388	1.02413e+ 06	1734	0	1	110682	241.388
3	Dimas	243.589	896754	1792	0	1	117097	243.589
4	Dimas	245.788	692442	1802	0	1	118146	245.787
5	ichsan	246.396	638417	1810	0	1	119229	246.395
6	ichsan	244.944	743871	1787	0	1	116855	244.943
7	uray	245.464	728959	1810	0	1	119408	245.463
8	uray	244.546	790898	1808	0	1	119734	244.546
9	tri	234.951	1.49205e+06	1660	0	1	102959	234.951
10	tri	236.654	1.37725e+06	1684	0	1	105871	236.654
11	sabrina	244.068	757641	1714	0	1	105786	244.068
12	sabrina	243.07	832903	1632	0	1	91779	243.07
13	putry	242.303	975404	1782	0	1	117245	242.303
14	putry	243.317	898195	1804	0	1	119655	243.317
15	puspa	242.091	972550	1756	0	1	113501	242.091
16	puspa	243.606	868877	1789	0	1	117399	243.605
17	panji	242.798	925371	1770	0	1	115457	242.798
18	panji	241.149	1.04017e+06	1742	0	1	112221	241.149
19	novenia	239.485	1.18948e+06	1735	0	1	111402	239.485
20	novenia	240.179	1.11447e+06	1734	0	1	111593	240.179
21	niko	238.972	1.19305e+06	1707	0	1	108583	238.971
22	niko	239.86	1.11348e+06	1713	0	1	109538	239.859
23	martin	242.088	968328	1755	0	1	113798	242.088
24	martin	243.949	836173	1786	0	1	117117	243.949
25	Samudera	241.431	972916	1724	0	1	110547	241.431
26	Samudera	242.419	911348	1743	0	1	112612	242.419
27	kemal	243.08	900427	1780	0	1	116734	243.08
28	kemal	243.954	839282	1793	0	1	117972	243.954
29	carmelita	240.041	1.04483e+06	1681	0	1	103782	240.041
30	carmelita	242.211	884113	1717	0	1	107661	242.221
31	alberties	243.657	715081	1667	0	1	100893	243.657
32	alberties	244.769	640764	1682	0	1	101382	244.769
33	jeany	243.794	850408	1790	0	1	117819	243.794
34	jeany	242.638	918146	1758	0	1	114450	242.638

From the seven texture features that have been used to investigate the authenticity of handwriting, the Skewness and relative smoothness features are the most difficult to apply to distinguish document ownership because they have the same value for all tested data. This can be because a perfectly symmetrical dataset will have a slope of 0, because skewness is usually described as a measure of a dataset's symmetry - or lack of symmetry. Whilefor Relative Smoothness, which is related to flexibility or speed, where handwriting is basically made automatically by individual abilities or without fabrication, so that all documents have a maximum relative smoothness value of 1 [8].

3.3 LVQ Python Extraction Results Processing

After extracting the features as in table 1 which is stored in the SQLite database of the mobile application, it can be continued by making predictions using a new handwritten image that will be checked. This new data will be carried out by a feature extraction process, where the results will be compared with the data already stored in the database to see which data has the closest similarity seen from the smallest feature value distance and the name of the owner of the image is shown as the prediction result from the classification results.

To calculate the closest distance and classification of image owners, the closest distance calculation in this study uses the Learning Vector Quantization (LVQ) method.

The Learning Vector Quantization classification process is carried out on a python flask, the flask is used as a framework for processing data from mobile application feature extraction to the python algorithm. Learning Vector Quantization on the flask. the data transfer process between the mobile

Application and the flask using JSON HttpURLConnection in packages through the link ngrok.

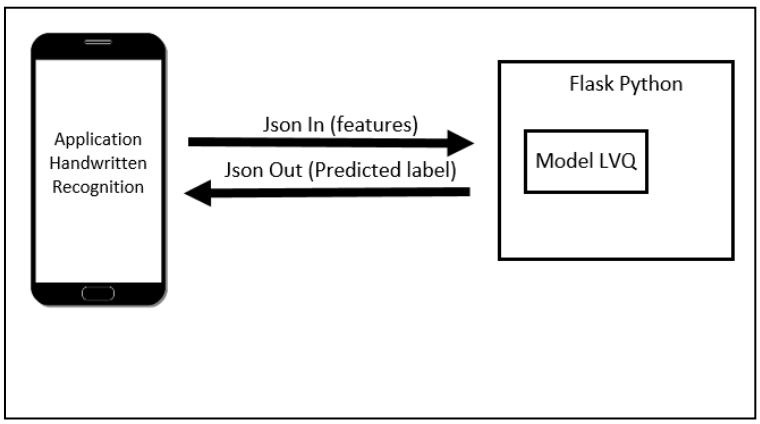

Figure 2 Prediction process schema model.

After the extraction results are obtained from the input data, the extracted values will be entered into new methods, namely array_train and array_target, this process is used when doing LVQ training and LVQ testing. With a summary of the parameters in the LVQ model which is formed as in Figure 3.

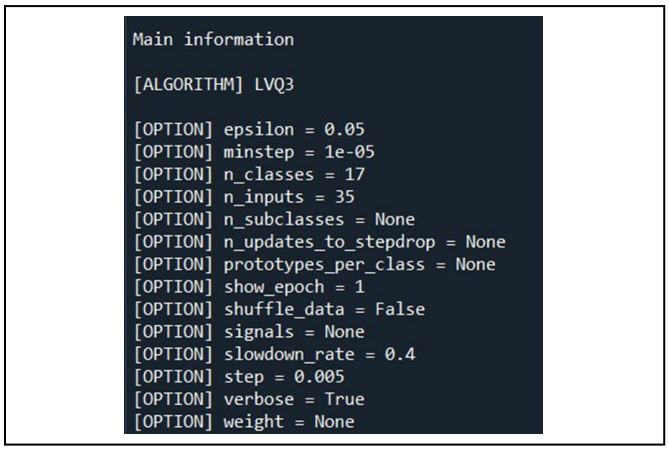

Figure 3 Summary Parameter LVQ

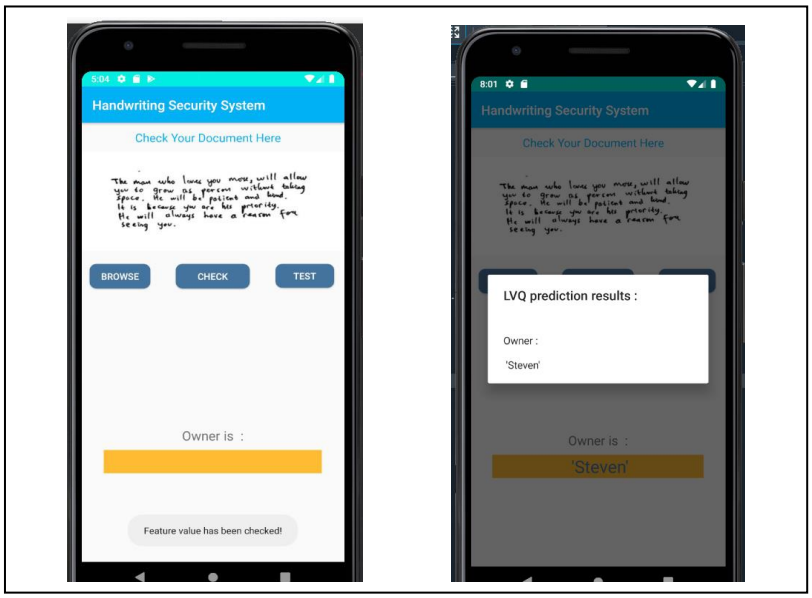

Figure 4 Display of prediction process results.

3.4 Test Results

Conducted tests to see the performance of the learning vector quantization method by inputting 16 new image data with 5 tests for each image, so that the total testing was 80 times. conducted tests to see the performance of the learning vector quantization method by inputting 16 new image data with 5 tests for each image, so that the total test was 80 times. The test results can be seen in table 2 learning vector quantization test.

Table 2 Learning Vector Quantization Test

ID	Owner's	Testing	LVQ Prediction Results	(True/False)
35	Steven	1st	Steven	T
		2nd	Steven	T
		3rd	Puspa	F
		4th	Putry	F
		5th	Steven	T
36	Dimas	1st	Dimas	T
		2nd	Dimas	T
		3rd	Carmelita	F

		4th	Kemal	F
		5th	Dimas	T
37	Ichsan	1st	Alberties	F
		2nd	Ichsan	T
		3rd	Ichsan	T
		4th	Alberties	F
		5th	Ichsan	T
38	Tri	1st	Tri	T
		2nd	Tri	T
		3rd	Tri	T
		4th	Tri	T
		5th	Tri	T
39	Tri	1st	Tri	T
		2nd	Niko	F
		3rd	Tri	T
		4th	Tri	T
		5th	Tri	T
40	Uray	1st	Uray	T
		2nd	Martin	F
		3rd	Uray	T
		4th	Kemal	F
		5th	Uray	T
41	Puspa	1st	Putry	F
		2nd	Kemal	F
		3rd	Puspa	T
		4th	Puspa	T
		5th	Puspa	T
42	alberties	1st	alberties	T
		2nd	ichsan	F
		3rd	alberties	T
		4th	ichsan	F
		5th	alberties	T
43	alberties	1st	alberties	T
		2nd	sabrila	F
		3rd	alberties	T
		4th	alberties	T
		5th	alberties	T
44	alberties	1st	alberties	T
		2nd	dimas	F
		3rd	alberties	T
		4th	alberties	T
		5th	dimas	F
45	Carmelita	1st	carmelita	T
		2nd	Niko	F
		3rd	carmelita	T
		4th	carmelita	T
		5th	steven	F
46	Carmelita	1st	carmelita	T
		2nd	samudera	F
		3rd	jeany	F
		4th	carmelita	T
		5th	carmelita	T

		1st	Novenia	F
		2nd	niko	T
47	Niko	3rd	Niko	T
		4th	Novenia	F
		5th	Niko	T
		1st	Niko	T
		2nd	Niko	T
48	Niko	3rd	Niko	T
		4th	Niko	T
		5th	Niko	T
		1st	Niko	F
		2nd	Novenia	T
49	Novenia	3rd	novenia	T
		4th	niko	F
		5th	novenia	T
		1st	Puspa	F
		2nd	M samudera	T
50	M Samudera	3rd	M samudera	T
		4th	Puspa	F
		5th	M samudera	T
			True	54
			False	26
			Accuracy	67,5 %

The results obtained from testing prediction learning vector quantization of 16 new data with 80 times the total test. The result is that 54 data are correct and 26 data are false, so that the accuracy is 67.5%. This accuracy is obtained because the Learning Vector Quantization model is made using a learning rate = 0.005, alpha value = 0.05, and iteration = 100.

Accuracy can be improved again if more data is used as training data in the SQLite database, because the value of the features to be matched will be more varied to allow details of various post features to be more visible and the system can more easily recognize patterns. So that the system built by researchers can be used to help handwriting recognition to find out the owner of the handwriting quickly and easily.

4 Conclusion

The implementation of texture-based feature extraction with Learning Vector Quantization in developing mobile applications has a fairly good success rate of 67.5%. Better than the previous research conducted by Haviluddin where the LVQ model made had an accuracy of 66.66%. helps handwriting recognition to find out the owner of the handwriting quickly and easily. proven by the test results of the learning vector quantization prediction obtained of 16 new data with 80 times the total test. The result is that 54 data are correct and 26 data are false. From the seven texture features that have been used to investigate the authenticity of handwriting, the Skewness and relative smoothness features are the most difficult to apply to distinguish document ownership because they have the same value for all tested data.

Acknowledgement

This research was fully funded and highly supported by Trisakti University.

References

[1] G. S. Budhi and R. Adipranata, "Handwritten Javanese Character Recognition Using Several Artificial Neural Network Methods," *J. Eng. Technol. Sci.*, vol. 8, no. 3, pp. 195–212, 2015.

[2] Jasril and S. Sanjaya, "Learning Vector Quantization 3 (LVQ3) and Spatial Fuzzy C- Means (SFCM) for Beef and Pork Image Classification," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 2, pp. 60–65, 2018.

[3] S. M. Shamim, A. S. Mohammad Badrul Alam Miah, M. Rana, and A. Al Jobair, "Handwritten Digit Recognition using Machine Learning Algorithms," *Glob. J. Comput. Sci. Technol. D Neural Artif. Intell.*, vol. 18, no. 1, pp. 17–23, 2018.

[4] Haviluddin, R. Alfred, and H. Santoso, "Handwriting Character Recognition using Vector Quantization Technique," *Knowl. Eng. Data Sci.*, vol. 2, no. 2, pp. 82–89, 2019.

[5] R. R. Janrao and D. D. Dighe, "International journal of engineering sciences & research technology handwritten english character recognition using lvq and knn *," *Int. J. Eng. Sci. Res. Technol.*, vol. 5,

no. 8, pp. 904–912, 2016.
[6] R. Hamidi, M. T. Furqon, and B. Rahayudi, "Implementasi Learning Vector Quantization (LVQ) untuk Klasifikasi Kualitas Air Sungai," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1758–1763, 2017.
[7] H. Masrani, Ilhamsyah, and I. Ruslianto, "APLIKASI PENGENALAN POLA PADA HURUF TULISAN TANGAN MENGGUNAKAN JARINGAN SARAF TIRUAN DENGAN METODE EKSTRAKSI FITUR GEOMETRI," *J. Coding*, vol. 6, no. 2, pp. 69–78, 2018.
[8] D. Pratiwi, Syaifudin, T. Rahardiansyah, W. Anggriani, A. Hilman, and N. Chairunnisa, "Texture Traits with Uniform-Quantization in Handwriting Documents for Digital Forensics Investigation," *Int. Conf. Aerosp. Aviat.*, vol. 645, pp. 1–7, 2019.
[9] Syaifudin, D. Pratiwi, and Trubus, "Handwritten Security Modeling based on Cosine and Inner Product Method," *Technol. Reports Kansai Univ.*, vol. 62, no. 2, pp. 107–114, 2020.