Classification of Planktonic Foraminifera Fossil Types with Feature Optimization of Support Vector Machine (SVM) Algorithm Using Particle Swarm Optimization (PSO)

Herlina Jayadianti[1], Dhimas Wahyu Asshidiq[1], Budi Santosa[1], Siti Umiyatun Choiriah[4] , Frans Richard Kodong[1] , Bambang Yuwono[1]

[1]Department of Informatics, Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta 55281, Indonesia
herlina.jayadianti@upnyk.ac.id
4 Department of Geological Engineering, Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta 55281, Indonesia

Abstract. Objective: To build a Support Vector Machine model with a good Particle Swarm Optimization parameter selection algorithm for object classification in images. Can determine the best features used for classification in the Support Vector Machine algorithm using Particle Swarm Optimization feature optimization. Design/method/approach: Image data will be extracted GLCM features namely homogeneity, contrast, correlation, and energy as well as shape feature extraction namely area, perimeter, and metric eccentricity. The results of feature extraction are then used for training SVM models and PSO-SVM models. This research uses multiclass SVM, namely One Versus Rest (OVR) with C of 10 and RBF kernel. For feature selection using PSO, the parameters used are C1 by 2.0, C2 by 2.0, W by 1.0, the number of particles by 50 and the maximum iteration is 50. Results: In the PSO-SVM model, after feature selection, the number of features used is 9 features from the total feature extraction of 19 features. After testing with confusion matrix, the SVM model gets an accuracy of 92% while the accuracy of the PSO-SVM model is 97%. The test results show that feature selection using PSO can overcome the problems in the SVM algorithm and can improve the performance of the model for the classification of plankton foraminifera fossils. Originality / state of the art: the use of the PSO method for the selection of relevant features in the results of GLCM feature extraction and shape feature extraction with SVM classification algorithm.

Keywords: SVM, GLCM, Classification, Fossils

1. Introduction

Fossilized plankton foraminifera are very important in paleontological and geological studies because they can provide valuable information about the history of the marine environment and past climate change. Research on fossilized plankton foraminifera can provide insight into sea surface temperature, salinity, nutrient levels, and other environmental factors at specific times in the past (Darling and Wade 2008). In the upstream industry, especially in oil and gas exploration, many foraminifera micropaleontology laboratories have been developed as the main supporting facilities (Sukandarrumidi, et

A. Putro Suryotomo and H. Cahya Rustamaji (eds.), *Proceedings of the 2023 1st International Conference on Advanced Informatics and Intelligent Information Systems (ICAI3S 2023)*,
Advances in Intelligent Systems Research 181,
https://doi.org/10.2991/978-94-6463-366-5_4

al. 2020). The Micropaleontology Laboratory located at UPN "Veteran" Yogyakarta analyzes fossil foraminifera plankton for learning media and also research materials for geology majors. To determine the type of a plankton foraminimera, you must see its shape and physical characteristics under a microscope, then identify the characteristics and shape of the fossil by asking a fossil expert or reading a reference book. This takes a long time and requires special expertise in determining the type of foraminifera fossil. Experts in the micropaleontology laboratory also make many mistakes and take a longer time when determining the species of plankton foraminifera. In the research of Mitra et al, with the title Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance, a comparison of foraminifera fossil classification between machine learning performance and classification performance by plankton experts was carried out. The result is that machine learning performance is more accurate with an average F1 score of 81% while experts only get an average F1 score of 61% [2]. Therefore, it is necessary to develop a system that is able to classify the type of foraminifera plankton, so that the type of foraminifera plankton can be known. Research related to classification using image data or images has been done before using various methods. In Budianto's research in 2018, Comparison of KNN and SVM Methods in Motorized Vehicle Plate Recognition obtained the results of vehicle plate recognition testing accuracy with the Support Vector Machine method with 95% accuracy. While using the KNN method to get 80% test accuracy [3]. In Sethy's research in 2019, Identification of Diseases in Rice Leaves using the PSO-based SVM method with feature extraction in the form of Correlation, Entropy, Variance, Homogeneity, Contrast, Energy and Mean. In this study, the PSO algorithm was used to optimize features in the SVM algorithm, and it was proven that PSO-based SVM could be a promising technique for classification and identification of rice leaf diseases with an accuracy of 97.91% [4]. In Novichasari's research in 2018, with the title PSO-SVM for Clove Leaf Classification Based on Morphological Feature Shape, Color and Texture of Leaf Surface GLCM with feature extraction of four morphological features of feature shape, three color features, and ten texture features resulting in an accuracy of 90.5% [5]. In Ramdani's research in 2021 with the title Optimizing Face Recognition Based on Linear Discriminant Analysis and K-Nearest Neighbor Using Particle Swarm Optimization resulted in an accuracy of 71.67% where testing by applying PSO feature selection got better results than testing without PSO feature selection. This is evident from the test results of applying PSO feature selection being able to increase accuracy by 1.67% [6]. In Hsiang's research in 2019, model training was carried out for the classification of foraminifera fossil image data using the CNN method. The study has concluded that the CNN method has an accuracy of 78.96% with VGG16 architecture and overfitting [7]. In Kour's 2019 research, Segmentation and Classification of Jambu, Jamun, Mango, Grape, Apple, Tomato, and Arjun plants using the PSO-based SVM method with LAB color feature extraction and LBP texture feature extraction resulted in an accuracy of 95.23% [8].

Based on the explanation above, it can be concluded that the Particle Swarm Optimization-based Support Vector Machine algorithm can classify image data well. Therefore, in this study, feature selection will be carried out on the SVM algorithm using the

Particle Swarm Optimization algorithm for the classification of plankton foraminifera microfossil varieties.

2. Research Method

Research methods are steps and concepts in order to obtain data that has been processed into clearer, more accurate and detailed information. Starting from the stages of data collection, data preprocessing, feature extraction, model building and model testing. The research stages are depicted in Figure 1.

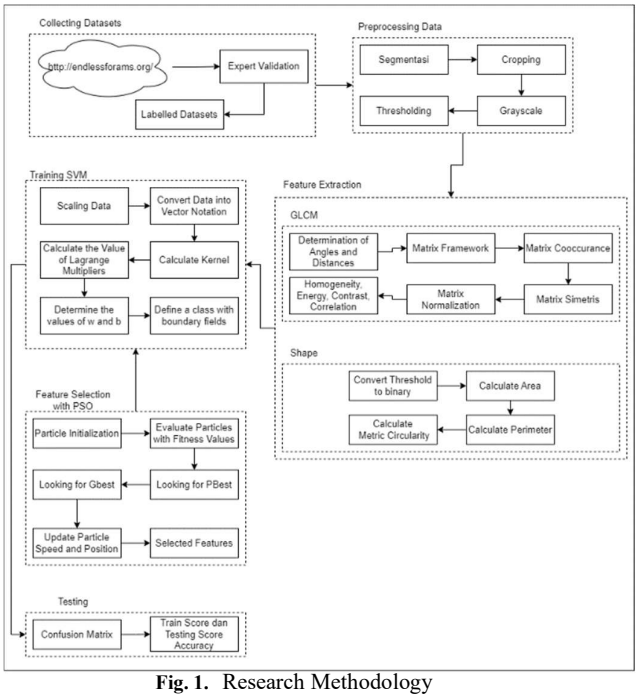

Fig. 1. Research Methodology

Fig. 2. Dataset collection was done by downloading images from the website http://endlessforams.org/ with the classes Globigerinoides sacculifer, Globigerina praebulloides, Globorotalia menardi, and Orbulina universa. The data was then saved into a folder with the name "raw data". Furthermore, the data was validated by a plankton foraminifera fossil expert, namely Dr. Ir. Siti Umiyatun Choiriah, M.T. who is a geology lecturer at the National Development University "Veteran" Yogyakarta by conducting interviews related to the validity of previously downloaded datasets. Examples of datasets that have been validated by experts can be seen in Figure 2 below.

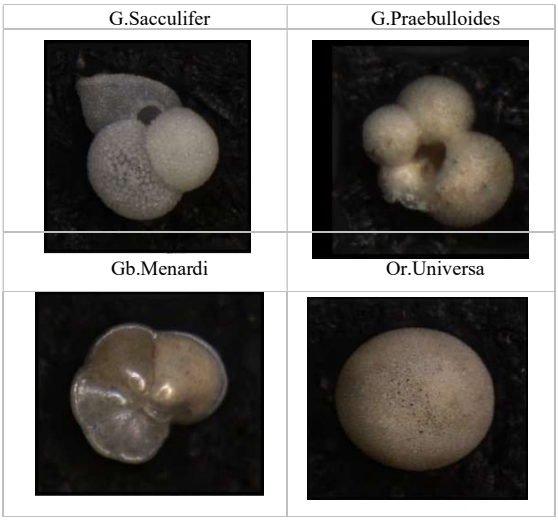

Fig 2 Data Set

2.1 Data Processing

The pre-processing stage is carried out to prepare the data so that it is easier to process before becoming a model. The pre-processing stages carried out in this study include segmentation, cropping, grayscale and thresholding.

2.2 Image Segmentation

Image segmentation is done with the aim of separating objects (foreground) from the background so that object values can be extracted appropriately. In this segmentation is done using the features of the remove.bg website by using the website's third party API.

2.3 Cropping Image

Image cropping is the process of cutting or removing parts of the image that are irrelevant or unnecessary to focus the object to be classified. By cropping the image to focus the object, it is expected to improve the model's ability to recognize and classify the object with higher accuracy.

2.4 GLCM Feature Extraction

1. *Contrast*

$$Contrast = \sum_{i=0}^{G} \sum_{j=1}^{G} (i-j)^2 \, p(i,j) \quad (1)$$

2. *Homogeneity*

$$Hom = \sum_{i=0} \sum_{j} \frac{p(i,j)}{1+|(i-j)|} \quad (2)$$

3. *Correlation*

$$Correlation = \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{(i-\mu_{it})(j-\mu_{jt})(GLCM(I,J))}{(\sigma_{it}\sigma_{jt})} \quad (3)$$

4. *Energy*

$$Energy = \sum_{i} \sum_{j} p(i,j)^2 \quad (4)$$

2.5 Shape Feature Extraction

Fossils can be distinguished from the shape of the chamber, the number of spheres in the fossil and the shape of the fossil itself so that this research will use shape feature extraction. The shape feature extraction that will be used in this research is as follows:

1. Area

$$Area = Ai = \sum_{r=0}^{height-1} \sum_{c=0}^{width} Ii(rc) \quad (5)$$

2. Perimeter (batas objek)

$$Perimeter = (P_{ij}, X\,edge[P] = i, Y\,edge[P] = j \quad (6)$$

3. *Metric Circularity*

$$Metric = 4\pi \frac{A}{C^2} \quad (7)$$

2.6 SVM Model Building

After obtaining the extraction results from the feature calculation process, the next step is to use these results into the classification process using Multiclass Support Vector Machine using the one versus rest or one versus all method based on previous research, one versus all is good at classifying multilabel classes [9]. The first step of the OVA approach process is to divide the classification problem and training process into N models according to the number of categories from the dataset until the hyperplane of each category is obtained.

2.7 PSO-SVM Modeling

Feature optimization is done to find the most relevant features so that it can produce a better SVM model. Broadly speaking, this process consists of several stages, namely particle initialization, particle evaluation by comparing the fitness value of each particle, finding Personal Best (Pbest), finding Global Best (Gbest) and updating the speed and position of each particle. For feature selection using PSO, the parameters used are C1 of 2.0, C2 of 2.0, W of 1.0, number of particles of 50 and maximum iterations of 50. The stages of feature selection using PSO can be seen in Figure 3.

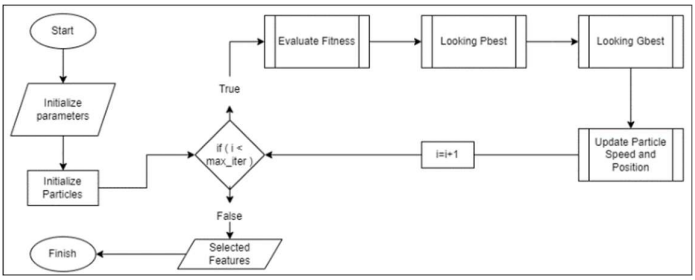

Fig. 3. Data Set

3. Results and Discussion

Results and discussion contain research results and discussion related to the results of the study. Each table image displayed must be accompanied by an explanation so that the reader can understand the contents of the image or table. Explanations related to the data presented must be conveyed in this section with the aim of clarifying the usefulness of the data in the study. In the preprocessing stage, this stage starts by removing the background from the dataset. Background removal is done using the python language. In the background removal process in this study using a third party API from the remove.bg website by generating API_KEY for the third party API configuration. After removing the background, the next stage is cropping the data according to the size of the object, cropping the data is done using the PIL library in the python language. Cropping is done with a box system, so that the object will be right in the middle of the image and the image only contains objects with a little remaining space. The result of background removal and cropping can be seen in Figure 4. The cropped image is then converted into a gray image using the PIL library and converted into a numpy array so that it can be read in the GLCM library in python. The image data that has been preprocessed will be continued with feature extraction using texture features, namely GLCM and shape features.

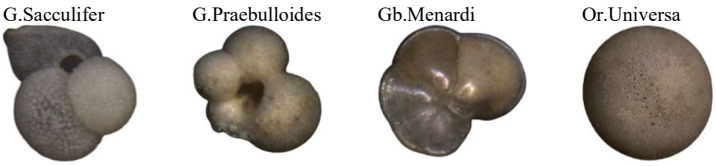

Fig 4. Background Removal and Cropping Result

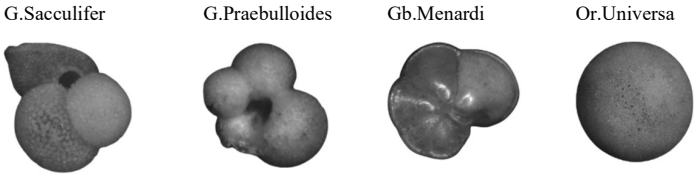

Fig. 5. Grayscaling Result

The first feature extraction is texture feature extraction using GLCM with greycomatrix and greycoprops libraries. The extraction of texture features includes homogeneity, contrast, correlation, and energy with a distance of 2 pixels and angles of 0°, 45°, 90°, and 135°. The process to obtain the GLCM texture extraction value is first to find the initial matrix value of the input image, then find the transpose matrix, then make a symmetrical matrix, and normalize the matrix for the calculation process of each GLCM feature. After calculating the four GLCM features, 16 values of each feature and each angle will be obtained. Examples of GLCM feature values from each fossil class can be seen in Table 1 to Table 5.

Table 1. GLCM Feature Value of Globigerina praebulloides Class

	0	**45**	**90**	**135**
Homogeneity	0.401573789	0.442793142	0.400651582	0.442529351
Contrast	41.87988854	25.05251526	43.04103554	27.2994887
Correlation	0.988385578	0.993044694	0.993044694	0.992420869
Energy	0.232018745	0.233144686	0.230502329	0.233192315

Table 2. GLCM Feature Value of Globigerinoides sacculifer Class

	0	**45**	**90**	**135**
Homogeneity	0.431179102	0.431179102	0.430810812	0.463621222
Contrast	72.87788195	42.79249413	77.50432041	50.66586015
Correlation	0.985468169	0.991469057	0.98456185	0.989899454
Energy	0.287808845	0.290669767	0.288488525	0.289749961

Table 3. GLCM Feature Value of Globorotalia menardi Class

	0	45	90	135
Homogeneity	0.425242694	0.450515717	0.415724459	0.457388326
Contrast	77.31746868	45.5622635	85.76827117	51.88654541
Correlation	0.982379456	0.989609897	0.980438767	0.988167696
Energy	0.284821875	0.285788138	0.283512895	0.286074719

Table 4. GLCM Feature Value of Orbunila Universa Class

	0	45	90	135
Homogeneity	0.342018587	0.373547087	0.34132116	0.371261894
Contrast	76.71486272	47.91138781	74.26730733	49.84606309
Correlation	0.981750982	0.988598164	0.982335792	0.988137755
Energy	0.204705229	0.206325871	0.204972599	0.20626581

The next feature extraction is shape feature extraction. The shape features used are area, perimeter, and circularity metrics. Before shape feature extraction, the grayscale image is thresholding by converting the image pixels to a value of 0 or 255 based on the specified threshold. This stage uses the cv2 library found in python.

Table 5. Shape Feature Extraction Result

area	perimeter	metric	Kelas
45849.5	833.7371	0.828869603	Globigerina Praebulloides
65683.5	1020.389	0.792747013	Globigerinoides Sacculifer
86676.5	1102.4722	0.896140151	Globorotalia menardii
63414	1013.9453	0.775114675	Orbulina Universa

3.1 SVM model training

The first step in SVM model training is to initialize the required parameters. This research uses the SVM method with multiclass One Versus Rest (OVR) with C of 10 and RBF kernel. Furthermore, model training is carried out using training data that has been scaled, namely X_train and Y_train. Model training uses the library from sklearn.svm, namely SVC. After training the model with the SVM method, the next step is to test using test data that has been scaled, namely X_test. The next stage is testing the confusion matrix. This stage uses the library from sklearn, namely confusion_matrix, ConfusionMatrixDisplay, accuracy_score, and classification_report. The confusion matrix calculation is done by comparing the actual label Y_test_label with the predicted label Y_pred_label. The results of testing with confusion matrix that has been displayed with a plot can be seen in table 6.

Table 6. Confusion Matrix Model SVM

		Prediction Class				Total Actual Class Data
		GS	GP	GM	U	
Actual Class	G.Sacculifer (GS)	28	2	0	0	30
	G.Praebulloides (GP)	3	27	0	0	30
	Gb.Menardi (GM)	5	0	25	0	30
	Or.Universa (U)	0	0	0	30	30
Total Prediction Class Data		36	29	25	30	120

After the values of TP, FP, and TP are found as in table 7, then calculations are made to dance accuracy, precision, and recall. The results of the calculation of accuracy, precision, and recall values can be seen in Table 7 below.

Table 7. Calculation Results of Accuracy, Precision, and Recall SVM Model

No	Class	TP	FP	FN	Accuracy	Precision	Recall
1	G.Sacculifer	28	8	2	-	0.78	0.93
2	G.Praebulloides	27	2	3	-	0.93	0.90
3	Gb.Menardi	25	0	5	-	1.00	0.83
4	Or.Universa	30	0	0	-	1.00	1.00
Total					0.92	0.92	0.92

3.2 PSO-SVM Model Training

Training the PSO-SVM model uses the same data as training the SVM model, namely data that has been scaled. The first step at this stage is to initialize the parameters needed for the PSO algorithm. The next step is to create a fitness function to evaluate the fitness of the PSO algorithm with the selected_features parameter taken from initialization. Then selected_features will be converted to bolean to get the feature index from the dataset. If there is no true value in the selected_features variable, the function will return -1.0. The selected features will be used to train the SVM model with RBF kernel, C of 10 and OVR multiclass. After training the SVM model, testing is then carried out using test data to get the accuracy of the SVM model which functions as fitness in the PSO algorithm. The next step is to perform feature selection using the PSO algorithm with the pyswarm library. PSO performs 50 iterations with 50 particles so that each iteration there are 50 combinations of features resulting from updating the speed and position of the particles so that Gbest is obtained. The next step is to display the results of the feature selection.

Table 8. Feature Selection Result

Homogeneity 0	Homogeneity 45	Homogeneity 90	Homogeneity 135	Energy 0	Energy 45	Energy 90	Energy 135	Contrast 0
0	0.91442106	0.019914 02	1	0.2965790 6	0	1	0.943502 9	0

Table 9. Feature Selection Results (Continued 1)

Contrast 45	Contrast 90	Contrast 135	Correlation 0	Correlation 45	Correlation 90	Correlation ...	area	perimeter	metric	
0	0	0	0	0	0.1108 0298	-	0.619 76724			0.75143 024

In Table 6 and Table 7, the number 0 indicates that the feature is not selected by PSO, while the number more than 0 indicates the feature is selected by PSO. From Table 8 and Table 9, from a total of 19 features extracted, only 9 features were selected by the PSO algorithm for SVM model training. After feature selection using PSO, the next stage is the creation of SVM models that have been selected using PSO. SVM models that have been trained, then tested with testing data using confusion matrix. After testing the model using testing data and confusion matrix, calculations are then carried out to find accuracy, precision, and recall.

Table 10. PSO-SVM Model Testing Results with Confusion Matrix

		Prediction Class				Total Actual Class Data
		G.Sacculifer	G.Praebulloides	Gb.Menardii	Or.Universa	
Actual Class	G.Sacculifer	28	2	0	0	30
	G.Praebulloides	1	29	0	0	30
	Gb.Menardi	1	0	29	0	30
	Or.Universa	0	0	0	30	30
Total Prediction Class Data		30	31	29	30	120

The next stage calculated the accuracy, precision and recall of the PSO-SVM model using the results of model testing in table 11 The calculation is done using formulas. The results of the calculation are then displayed as in table 11 below.

Table 11. Calculation Results of Accuracy, Precision, and Recall of PSO-SVM Model

No	Class	TP	FP	FN	Accuracy	Precision	Recall
	G.Sacculifer	28	2	2	-	0.93	0.93
	G.Praebulloides	29	2	1	-	0.94	0.97
	Gb.Menardi	29	0	1	-	1.00	0.97
	Or.Universa	30	0	0	-	1.00	1.00
Total					0.97	0.97	0.97

The PSO-SVM model has an accuracy of 97% on the overall model in classifying all test samples from all classes. The average precision and recall of the PSO-SVM model is 0.97.

4. Results and Discussion

The results of feature extraction are then used for model training using SVM and PSO-SVM methods. The PSO algorithm is used for selecting relevant features for the SVM model. This research compares SVM models without feature selection and SVM

models with PSO feature selection. Then testing using confusion matrix on each model and calculated accuracy, precision and recal. After testing, it was found that the SVM model that has been selected using PSO has increased accuracy by 5% from the SVM method without PSO feature selection. Graphs regarding the comparison of the accuracy of SVM models with feature selection and without feature selection can be seen in Figure 6 below.

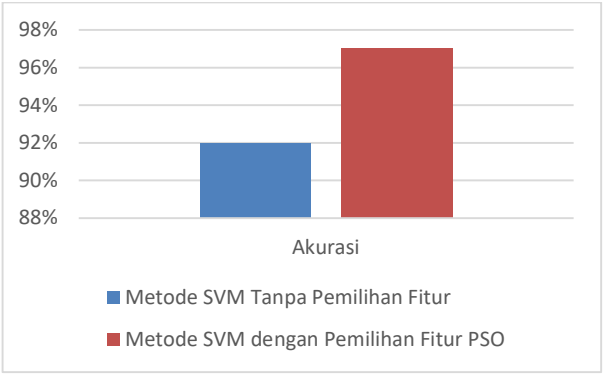

Fig. 6. Model Accuracy Comparison

Based on the results of the Support Vector Machine (SVM) Algorithm Feature Optimization Using Particle Swarm Optimization (PSO) for Classification of Plankton Foraminifera Fossil Types that have been carried out, the system that has been made can classify foraminifera fossil species well as evidenced by the comparison between machine performance and the performance of micropaleontology laboratory assistants. With the same data, the system can determine the fossil species correctly and produce 100% accuracy for 19 test data. Meanwhile, the micropaleontology laboratory assistant can determine 15 species only and get an accuracy of 78.95% with 4 errors on data with the sacculifer and praebulloides classes. The Support Vector Machine method with feature selection using Particle Swarm Optimization can recognize foraminifera fossil images well and can perform classification with test results of 97% accuracy, 97% precision and 97% recall with 280 train data and 120 testing data. This study uses 4 classes of plankton foraminifera fossils with a total of 400 data. Particle Swarm Optimization used for feature selection has an influence on the accuracy, precision and recall of the Support Vector Machine method in classifying plankton foraminifera fossil data, where the test results show an increase in accuracy, precision and recall by 5%.

The results obtained in the Support Vector Machine method without feature selection get an accuracy of 92%, precision of 92% and recall of 92%, while the results obtained from the Support Vector Machine method with feature selection using PSO get an accuracy of 97%, precision of 97% and recall of 97%. The suggestions that can be used in future research for the development of the Support Vector Machine model in classifying plankton foraminifera fossils using other feature selection algorithms such as Information Gain, Forward Selection or Backward Elimination and compared with this research so that it can be seen which feature selection algorithm is better for

the Support Vector Machine method. Future research can also add datasets of foraminifera fossil images so that the Support Vector Machine model can perform better classification.

Acknowledgment. We would like to thank the lab of fossils in UPN Veteran Yogyakarta for providing access to the data required in this study. I would also like to thank all the researchers who assisted in the data analysis process and provided valuable feedback.

References

1. K. F. Darling and C. M. Wade, "The genetic diversity of planktic foraminifera and the global distribution of ribosomal RNA genotypes," *Mar. Micropaleontol.*, vol. 67, no. 3–4, pp. 216–238, 2008, doi: 10.1016/j.marmicro.2008.01.009.
2. R. Mitra *et al.*, "Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance," *Mar. Micropaleontol.*, vol. 147, pp. 16–24, Mar. 2019, doi: 10.1016/j.marmicro.2019.01.005.
3. A. Budianto, R. Ariyuana, and D. Maryono, "PERBANDINGAN K-NEAREST NEIGHBOR (KNN) DAN SUPPORT VECTOR MACHINE (SVM) DALAM PENGENALAN KARAKTER PLAT KENDARAAN BERMOTOR," *J. Ilm. Pendidik. Tek. Dan Kejuru.*, vol. 11, no. 1, p. 27, Nov. 2019, doi: 10.20961/jiptek.v11i1.18018.
4. P. Kumar Sethy, A. Rath, and N. Kanta Barpanda, "Detection & Identification of Rice Leaf Diseases using Multiclass SVM and Particle Swarm Optimization Technique Special Session Chair : Smart City with Emerging Technologies View project Gene Expression Feature Extarction and Classification View project," 2019. [Online]. Available: https://www.researchgate.net/publication/332978495
5. S. Ika Novichasari and Y. Romando Sipayung, "PSO-SVM Untuk Klasifikasi Daun Cengkeh Berdasarkan Morfologi Bentuk Ciri, Warna dan Tekstur GLCM Permukaan Daun," 2018.
6. M. H. Ramdani, G. Pasek, S. Wijaya, and R. Dwiyansaputra, "Optimalisasi Pengenalan Wajah Berbasis Linear Discriminant Analysis Dan K-Nearest Neighbor Menggunakan Particle Swarm Optimization (Optimization Of Face Recognition Based On Linear Discriminant Analysis And K-Nearest Neighbor Using Particle Swarm Optimiza," 2021. [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA/
7. A. Y. Hsiang *et al.*, "Endless Forams: >34,000 Modern Planktonic Foraminiferal Images for Taxonomic Training and Automated Species Recognition Using Convolutional Neural Networks," *Paleoceanogr. Paleoclimatology*, vol. 34, no. 7, pp. 1157–1177, 2019, doi: 10.1029/2019PA003612.
8. V. P. Kour and S. Arora, "Particle Swarm Optimization Based Support Vector Machine (P-SVM) for the Segmentation and Classification of Plants," *IEEE Access*, vol. 7, pp. 29374–29385, 2019, doi: 10.1109/ACCESS.2019.2901900.
9. E. Permata, D. Aribowo, and A. Maulana, "Klasifikasi Parasit Malaria Plasmodium Vivax Pada Citra Sel Darah Merah Menggunakan Metode Support Vector Machine One Against All," 2014.