International Journal of Basic and Applied Science 11 (2) (2022) 62-70

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Classification of marine mammals based on nucleotide using machine learning

Lukman[1], Tiara Dinda Hapsari[2], Abdul Malik[3], Ester Frescilla Simbolon[4], Ishak Ariawan[5], Nadia Yusuf Istiqomah[6]

[1,2,3,4,5,6]Universitas Pendidikan Indonesia, Indonesia

Article Info

Article history:

Received June 20th, 2022
Revised July 20th, 2022
Accepted September 9th, 2022

Keywords:

Classification
Machine learning
Marine mammals
Nucleotide

ABSTRACT

This study analyzing the nucleotide of marine mammals using machine learning techniques. Analysis on a nucleotide scale in marine mammals can help facilitate the identification process if done properly. Three types of marine mammals used for nucleotide analysis were *Delphinus capensis, Dugong dugon,* and *Orcaella brevirostris*. The solutions offered by machine learning provide a more elegant and effective solution for species identification at the nucleotide scale. This study analyzed the nucleotide s of marine mammals using various classification techniques. Based on this research, it can be concluded that the identification of marine mammals can be done easily based on nucleotide. Different classifiers have been used for analytical purposes such as Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbor, and Multilayer Perceptron. Based on the analysis of the results, it was found that the classification method that had been applied had sufficient performance by being tested on several model performance metrics such as accuracy, precision, recall and f1 score. The study also highlights the best classifiers in the various scenarios and recommendations are given.

Corresponding Author:

Lukman,
Department of Marine Information System,
Universitas Pendidikan Indonesia,
229 Dr. Setiabudi Road, Bandung City, West Java 40154, Indonesia.
Email: oluk@upi.edu

1. INTRODUCTION

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on learning from data [1]. ML focuses on developing systems that can learn independently from the experiences or input they get without the need to be programmed repeatedly by humans [2]. The machine learning process goes through two stages, namely training and testing. Machine learning is believed to be able to help and simplify the way humans work because they can predict with a high degree of accuracy [3]. How machine learning works depends on the presence of data. The more and more quality data there is, the better the work performance of machine learning will be.

Machine learning approach is widely used in solving quite complex problems in various fields. For example, in the field of health, ML can help detect diseases [4], build unmanned aircraft [5] or

transportation fields [6], analyze stock markets [7], as well as identify biota or species in the field of scientific research [2] [8].

Machine learning can be used to find out the genetic similarities or differences of several species by identifying the genetic composition of each species. This study uses the Multinomial Naïve Bayes algorithm, Random Forest, Decision Tree, K-Nearest Neighbor (K-NN), and Multilayer Perceptron (MLP) to identify species of marine mammals and then compare the accuracy of each algorithm model. The development of molecular biology science and knowledge resulted in rapid developments in researching and studying an organism and its benefits for human welfare. The use of molecular techniques to identify an organism has advantages such as being more accurate, faster, and covering all microbes [9].

Effective conservation requires identification of the main drivers of movement including intrinsic traits [10]. The study of genetic diversity in principle aims to examine the genetic composition of individuals within or between populations and to determine the factors that cause the modulation or dynamics of genetic diversity of these populations. In general, genetic diversity in a population can occur because genes undergo mutation, recombination, and the movement of a group of populations from one place to another [11].

Identification of marine biota can be done by looking at existing physical characteristics, such as body size, color or pattern, behavior, and others [12]. Another method to identify species can be done using DNA barcodes with molecular identification methods that are considered precise and accurate [13] [14]. DNA barcoding has become the standard in molecular identification because it has non-recombinant properties and does not depend on environmental conditions, so it is very effective in the management and monitoring of marine biota. In this study we use several machine learning algorithms or models to classify the nucleotide sequences of three marine mammal species. Then the performance of each model is tested to provide the best model recommendation.

2. RESEARCH METHOD

This research was conducted in stages consisting of data collection, data preprocessing, classification, and accuracy test. Research stages can be seen in Figure 1.

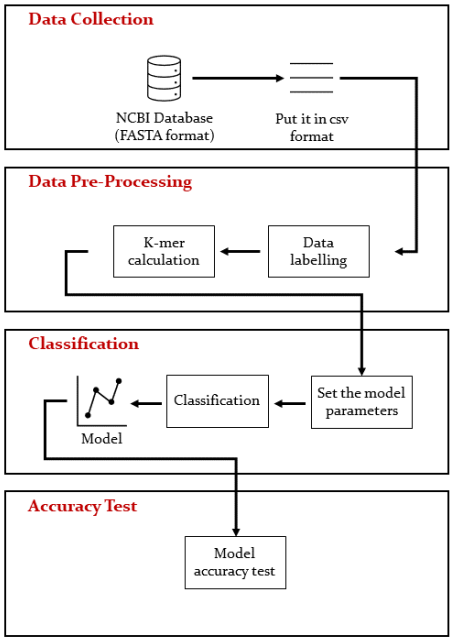


Figure 1. Research stages

2.1 Data Collection

The data used in this study were sourced from the Nucleotide database managed by NCBI (National Center for Biotechnology Information) which was accessed through https://www.ncbi.nlm.nih.gov/. The data collected from this research are nucleotide sequence data from several marine mammal species which are then classified. These species are *Delphinus capensis, Dugong dugon*, and *Orcaella brevirostris*.

2.2 Data Pre-Processing

At this stage, the data is divided into three classes based on each species. Then, the calculation of k-mer needs to be done to classify the nucleotide data based on the model. A k-mer is a sequence of k characters in a string (a nucleotide in a DNA sequence) [15]. It aims to get all k-mer from a sequence starting from the first character. Then move it to one character for the start of the next k-mer and so on. Effectively, this will create an overlapping sequence at position k-1. The perspective provided by the k-mer method is interesting on the complexity of the corresponding species [16].

2.3 Classification

Classification is the process of making decisions from the implementation of the applied algorithm. The grouping of data into classes or objects in the classification process is carried out based on certain attributes. Automatic classification using artificial intelligence technology has a lot of impact on biologists, government and society compared to manual identification by humans who are vulnerable to unexpected things that affect the classification results.

a. Multinomial Naïve Bayes

Multinomial Naïve Bayes is a branch of the Bayes theorem classification method that uses supervised learning with a probabilistic model and is suitable for use in classifying texts or documents [17]. Multinomial Naïve Bayes is able to classify an object whose class is not yet known and has high accuracy and speed results when used into the database [18]. This method is used to find a function model that describes the concept of data and then estimate the class of an object. This algorithm is influenced by the term, which means the number of words that appear in the document and the frequency of occurrence of the words in the document [17]. The formula used in Multinomial Naïve Bayes is as follows:

$$P(tn|c) = \frac{W_{ct}+1}{(\sum_{w'\in VW'ct})+B'} \text{ [17]}$$

W_{ct} is term in category c, then $(\sum_{w'\in VW'ct})$ is total W of all terms in category c, and B' is a unique of W number.

b. Decision Tree

Decision Tree is one of the most used classification methods for decision making by forming branches from each decision [19]. Decision Tree is used to predict the pattern of the data and describe the relationship of the attribute variable x and the target variable y in the form of a tree [20] [21]. This method classifies the given data using the value of the attribute. The classification process of this method is by dividing the data into smaller sub-sections based on the criteria made from each branch. The purest attribute can separate objects according to their class measured from the level of inpurity with the following calculation [22]:

$$entropy(S) = \sum_{i-1}^{m} -p(w_i|S)\,.\,log_2 p(w_i|S) \text{ [22]}$$

S is case set, m is number of data classes, then $p(w_i|S)$ is proportion of class i in all training data processed at node S.

c. Random Forest

Random Forest works more efficiently on large amounts of training and test data sets. Random Forest is based on the results of the classification and then the result with the highest number of votes is selected. The independent split vector selection method was randomly carried out in tree construction and then all trees were classified [23]. The result

of the final RF classification is the sum of the results of the initial learners or the results of the classification of each tree for each class [8]. Then the results of the tree classification will be added up based on the class. The class with the highest number of classes will be selected as the final classification result. The selection of the final classification in Random Forest is as follows [24]:

$$f(x) = arg^{max}_{y \epsilon Y} \sum^{i}_{j=1} I\left(y = h_j(x)\right) \text{ [24]}$$

Where $f(x)$ is the result of the Rando Forest classification and (x) is the result of the classification of each tree. Meanwhile, $(y = (x))$ is an indicator function that will give a value of 1 if the result of tree classification is the same as class y and 0 otherwise.

d. K-Nearest Neighbor

K-Nearest Neighbor (K-NN) is included in the supervised learning model. K-NN is done to classify new data or objects based on attributes or training samples [25]. K-NN is effectively used in training data that is large and strong against noise. The algorithm of K-NN is very simple, K-NN is determined based on the shortest distance from test data to training data. The way K-NN works is by searching for k-objects in the database that have a measure of the closest distance to the new object is as follows:

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \text{ [25]}$$

Where x_1 is the sample data, x_2 is the test data, i refers to the data variable, d is the distance, and p is the dimension of data.

e. Multilayer Perceptron

Multilayer Perceptron (MLP) is a type of feed-forward neural network with one or more hidden layers. Generally, MLP consists of an input layer which is a collection of neurons to enter data; at least one hidden layer as a computational neuron and an output layer as a storage neuron for computational results [26].

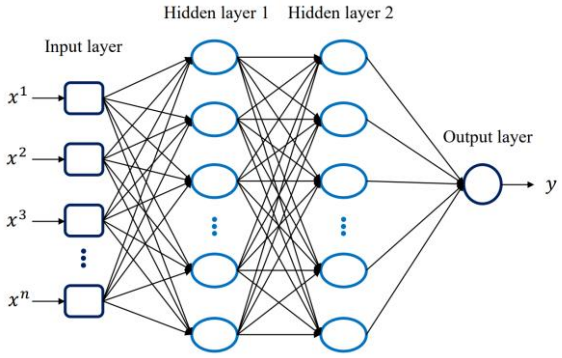


Figure 1. Multilayer perceptron (MLP) architecture [26]

In MLP, there are two important parameters, namely the activation function and the optimization function [27]. The activation function determines the output at a node of some input element. While the optimization function serves to determine the most appropriate weight according to the input and output.

2.4 Accuracy Test

The evaluation criteria for the classification model are seen from the calculation results. In the process of evaluating the capability of the model in generalizing the data, we use several accuracy test methods namely precision, recall, accuracy, and f1-score. The value of accuracy as the main measure of classification is the number of cases correctly classified in the test set divided by the total number of

cases. Precision and recall are measures of the accuracy and completeness of classification in a positive class [28]. The F1-score is a comparison of the average precision and recall.

3. RESULTS AND DISCUSSIONS

Several machine learning models have been applied to marine mammal nucleotide data. Nucleotide data obtained from https://www.ncbi.nlm.nih.gov/ were collected and grouped into three classes, the details of the amount of data used from each class were *Delphinus capensis* 50, *Dugong dugon* 50, and *Orcaella brevirostris* 50. Based on the distribution of the amount of data used per class, it can be seen in Figure 2 that each data for each species is balanced.

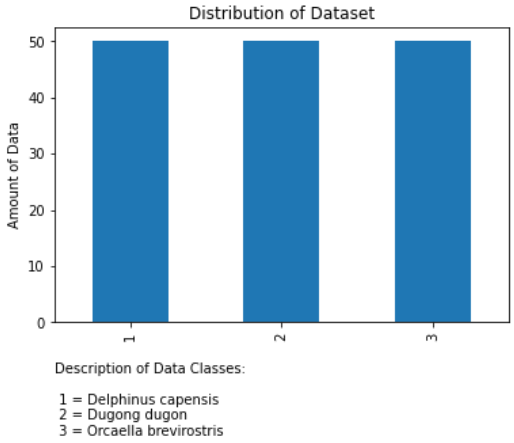


Figure 2. Data distribution of each class

3.1 Calculation of K-mer

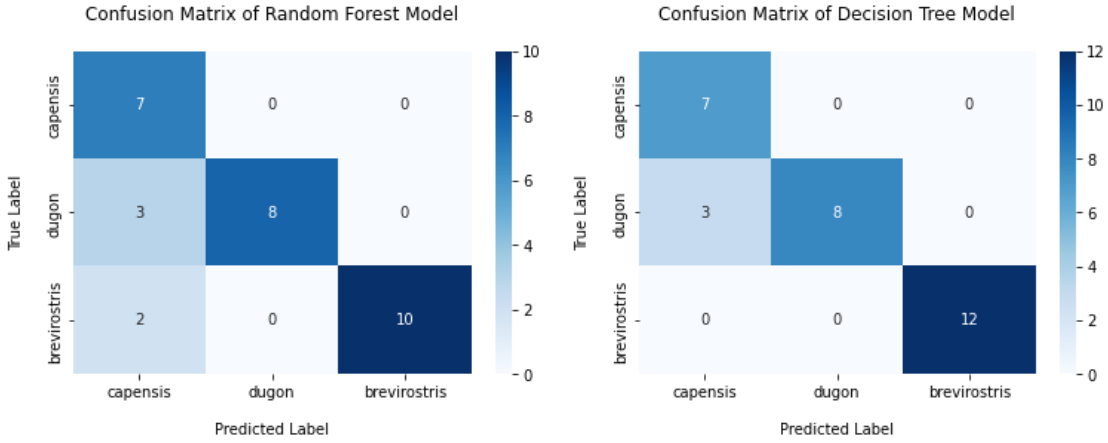
The process of processing nucleotide data before entering the classification stage is the k-mer calculation process. K-mer aims to break the sequence of nucleotide data then separated by overlapping. The term k-mer usually refers to all possible substrings of length k contained in a string. In this study, the sample taken was Delphinus capensis with GenBank id MZ401208.1, the sequence data "TCAAGGAAGAGA" was split into 6 hexamers into: 'TCAAGG', 'CAAGGA', 'AAGGAA', 'AGGAAG', 'GGAAGA', 'GAAGAG', 'AAGAGA'. The data that has been converted using the k-mer calculation is then converted into lowercase letters. Then the word length and the amount of overlap were determined empirically. This process aims to manipulate the data to calculate the probability of occurrence of certain k-mer sequences.

3.2 Classification

The classification process with the machine learning method is carried out using several packages available on Scikit-Learn. There are several steps in creating models using the packages available in Scikit-Learn. First, define a package to call the library that will be used in the process of making the classification model. The second is to set the model parameters. The Random Forest parameter is set with max_depth = 2 and random_state=0. The Decision Tree parameter is set to random_state=3 and max_depth=42. Multinomial Naïve Bayes parameter with alpha = 0.1, multinomial type selection because the dataset used is a multinomial distribution for each feature. The K-Nearest Neighbor (KNN) parameter is set to n_neighbors=3. The Multilayer Perceptron (MLP) parameter is set with solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5,2), random_state=1. Third, modeling based on reference labels for classification and training data. The last step is the analysis and evaluation of the prediction results.

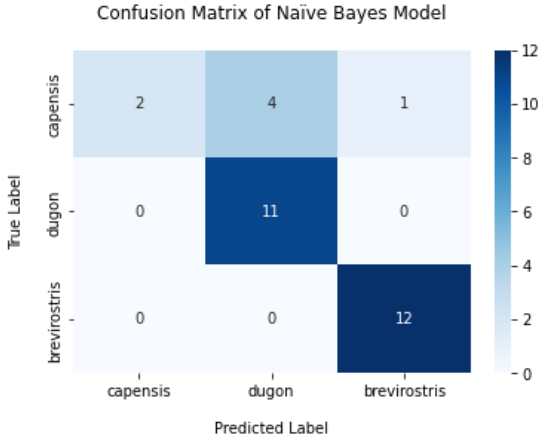
The classification process is carried out by implementing a dataset that becomes training data on a predetermined model. Then the model that has been trained is tested by testing data. This process

is used to analyze the model's capability to accommodate data outside of training data. Confusion matrix represents predictions and actual conditions of the data generated by models. The results of the experiment using test data on each model can be seen in the confusion matrix in Figure 3 and the percentage of predicted labels based on actual labels in Figure 4.

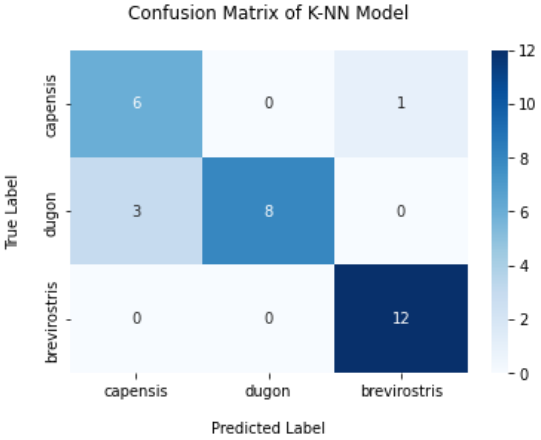


a. Random Forest

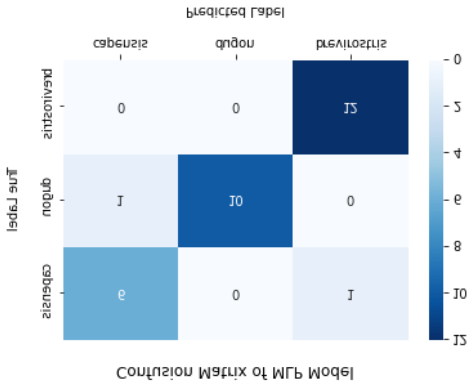
b. Decision Tree



c. Naïve Bayes



d. K-Nearest Neighbor



e. Multilayer Perceptron

Figure 3. Confusion matrix of models

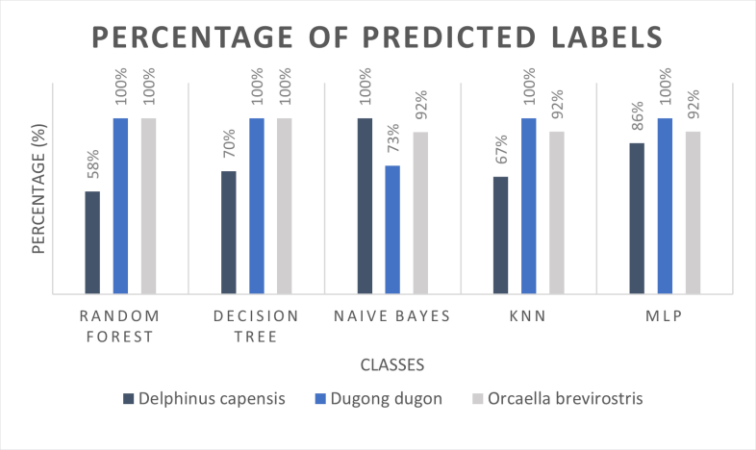


Figure 4. Percentage of predicted labels based on actual labels

3.3 Accuracy Test

After the Confusion Matrix is obtained as a reference to evaluate the performance of the models. Based on the confusion matrix, we can determine Accuracy, Precision, Recall and F1-score. The results of the model accuracy test that have been carried out can be seen in table 1.

Table 1. Model accuracy results

Model	Accuracy	Precision	Recall	F1-score	Average (%)
Random Forest	0.833	0.903	0.833	0.844	85.32%
Decision Tree	0.900	0.930	0.900	0.901	90.77%
Naïve Bayes	0.833	0.871	0.833	0.798	83.37%
K-NN	0.867	0.891	0.867	0.868	87.32%
MLP	0.933	0.936	0.933	0.933	93.37%

The results of the model accuracy test can be used as an evaluation reference. Table 1 shows the results of the model accuracy test using several model performance matrices. The results of the acquisition of an accuracy score show that overall, the tested models can study the data well. The model accuracy with the highest average is the MLP model with an accuracy gain of 93.37%. The model with the lowest average accuracy is Naïve Bayes with an accuracy of 83.37%. The lowest model accuracy is the F1-score accuracy test method is Naïve Bayes with an accuracy of 79.80%. The highest model accuracy is the Precision accuracy test method is MLP with an accuracy gain of 93.60%. This shows that the MLP model can study the data better than other models.

4. CONCLUSION

This study uses machine learning techniques to predict marine mammal species using different classification techniques. By using nucleotide data, this research predicts and compares several machine learning models or algorithms to classify the three species into the target class. Then several accuracy test methods are used to analyze the model's capability in studying the data. It was found that the MLP model gave the best performance for the classification of marine mammal species. The MLP model is proven to have good performance with an accuracy gain of 93.37%. The model with the lowest average accuracy is Naïve Bayes with an accuracy of 83.37%. Overall, all models tested using marine mammal nucleotide data performed well. Based on the research results, the MLP model is highly recommended as a nucleotide data classification model.

REFERENCES

[1] I. Cholissodin, A. A. Soebroto, U. Hasanah, and Y. I. Febiola, "AI, Machine Learning & Deep Learning." Fakultas Ilmu Komputer, Universitas Brawijaya, Malang, 2020.

[2] Z. B. S. Azminuddin I. S. Azis, *Machine Learning & Reasoning Fuzzy Logic Algoritma, Manual, Matlab, & Rapid Miner*. 2019.

[3] Y. F. Kao and R. Venkatachalam, "Human and Machine Learning," *Computational Economics*, vol. 57, no. 3, 2021, doi: 10.1007/s10614-018-9803-z.

[4] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *International Journal of Computer Applications*, vol. 17, no. 8, 2011, doi: 10.5120/2237-2860.

[5] K. Y. Li *et al.*, "An automated machine learning framework in unmanned aircraft systems: New insights into agricultural management practices recognition approaches," *Remote Sensing*, vol. 13, no. 16, 2021, doi: 10.3390/rs13163190.

[6] A. T. W. Min, R. Sagarna, A. Gupta, Y. S. Ong, and C. K. Goh, "Knowledge Transfer Through Machine Learning in Aircraft Design," *IEEE Computational Intelligence Magazine*, vol. 12, no. 4, 2017, doi: 10.1109/MCI.2017.2742781.

[7] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, vol. 53, no. 4, 2020, doi: 10.1007/s10462-019-09754-z.

[8] I. Ariawan, A. A. Rosalia, L. Anzani, L. Lukman, and others, "IDENTIFIKASI SPESIES MANGROVE MENGGUNAKAN ALGORITME RANDOM FOREST," *Jurnal Kemaritiman: Indonesian Journal of Maritime*, vol. 2, no. 2, pp. 118–128, 2021.

[9] D. Suryanto, "Melihat keanekaragaman organisme melalui beberapa teknik genetika molekuler," *Universitas Sumatera Utara. USU Digital Library. Medan*, 2003.

[10] A. M. M. Sequeira *et al.*, "Convergence of marine megafauna movement patterns in coastal and open oceans," *Proc Natl Acad Sci U S A*, vol. 115, no. 12, 2018, doi: 10.1073/pnas.1716137115.

[11] R. Sanjuán and P. Domingo-Calap, "Genetic Diversity and Evolution of Viral Populations," in *Encyclopedia of Virology*, 2021. doi: 10.1016/b978-0-12-809633-8.20958-8.

[12] D. Dharmadi, R. Faizah, and N. N. Wiadnyana, "FREKUENSI PEMUNCULAN, TINGKAH LAKU, DAN DISTRIBUSI MAMALIA LAUT DI LAUT SAWU, NUSA TENGGARA TIMUR," *BAWAL Widya Riset Perikanan Tangkap*, vol. 3, no. 3, 2017, doi: 10.15578/bawal.3.3.2010.209-216.

[13] Z. Gao, Y. Liu, X. Wang, X. Wei, and J. Han, "DNA Mini-Barcoding: A Derived Barcoding Method for Herbal Molecular Identification," *Frontiers in Plant Science*, vol. 10. 2019. doi: 10.3389/fpls.2019.00987.

[14] A. R. Simbolon and L. P. Aji, "IDENTIFIKASI MOLEKULAR DAN STRUKTUR FILOGENETIK MOLUSKA (GASTROPODA DAN BIVALVIA) DI PERAIRAN BIAK, PAPUA," *BAWAL Widya Riset Perikanan Tangkap*, vol. 13, no. 1, pp. 11–21, 2021.

[15] G. Rizk, D. Lavenier, and R. Chikhi, "DSK: K-mer counting with very low memory usage," *Bioinformatics*, vol. 29, no. 5, 2013, doi: 10.1093/bioinformatics/btt020.

[16] B. Chor, D. Horn, N. Goldman, Y. Levy, and T. Massingham, "Genomic DNA k-mer spectra: Models and modalities," *Genome Biology*, vol. 10, no. 10, 2009, doi: 10.1186/gb-2009-10-10-r108.

[17] F. K. S. Dewi, "KLASIFIKASI BERITA MENGGUNAKAN METODE MULTINOMIAL NAÏVE BAYES," *SCAN - Jurnal Teknologi Informasi dan Komunikasi*, vol. 16, no. 3, 2021, doi: 10.33005/scan.v16i3.2870.

[18] M. T. H. Bunga, B. S, Djahi, and Y. Y. Nabuasa, "Multinomial Naive Bayes Untuk Klasifikasi Status Kredit Mitra Binaan Di Pt . Angkasa Pura I Program Kemitraan," *J-Icon*, vol. 6, no. 2, 2018.

[19] D. Sartika and D. I. Sensuse, "Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian," *Jurnal Teknik Informatika Dan Sistem Informasi*, vol. 1, no. 2, 2017.

[20] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Archives of Psychiatry*, vol. 27, no. 2, 2015, doi: 10.11919/j.issn.1002-0829.215044.

[21] N. Ye, *Data mining Theories, Algorithms and Examples*, vol. 16, no. 4. 1997.
[22] R. A. Putranto, T. Wuryandari, and Sudarno, "Perbandingan Analisis Klasifikasi antara Decision Tree dan Support Vector Machine Multiclass untuk Penentuan Jurusan pada Siswa SMA," *Jurnal Gaussian*, vol. 4, no. 4, 2015.
[23] H. Santoso, "Performa Random Forest Group untuk Klasifikasi Penyakit Busuk Pangkal Batang yang Disebabkan oleh Ganoderma boninense pada Perkebunan Kelapa Sawit," *Jurnal Penelitian Kelapa Sawit*, vol. 28, no. 3, pp. 133–146, 2020.
[24] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble machine learning*, Springer, 2012, pp. 157–175.
[25] M. Lestari, "Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) untuk Mendeteksi Penyakit Jantung," *Faktor Exacta*, vol. 7, no. September 2010, 2014.
[26] A. S. Shirazi and I. Frigaard, "Slurrynet: Predicting critical velocities and frictional pressure drops in oilfield suspension flows," *Energies (Basel)*, vol. 14, no. 5, 2021, doi: 10.3390/en14051263.
[27] W. Castro, J. Oblitas, R. Santa-Cruz, and H. Avila-George, "Multilayer perceptron architecture optimization using parallel computing techniques," *PLoS ONE*, vol. 12, no. 12, 2017, doi: 10.1371/journal.pone.0189369.
[28] B. Juba and H. S. Le, "Precision-Recall versus accuracy and the role of large data sets," 2019. doi: 10.1609/aaai.v33i01.33014039.