MALAYSIAN JOURNAL OF SCIENCE
MJS
ISSN 1394-3065

Regular Issue

Malaysian Journal of Science 45(2): 1-12 (June 2026) https://mjs.um.edu.my

Predicting DNA Sequence Similarity Across Species Using Machine Learning: a K-mer Based Approach

Vinodkumar R. Patil[1a]*, Archana S. Vaidya[2a] and Manisha S. Patil[3b]

Abstract: This study examines the effectiveness of machine learning methods for categorizing DNA sequences across human, chimpanzee, and dog samples. We employed k-mer encoding to renovate DNA structures into numerical representations suitable for machine learning models. Four classifiers, such as naive bayes and weighted naive bayes, random forest, K-nearest neighbors, and decision tree, were applied and evaluated using accuracy, precision, recall, and F1-score. The naive bayes classifier consistently outperformed the others across all three datasets, achieving the highest accuracy in classifying human DNA (98.4%), followed by chimpanzee DNA (91.4%), and exhibiting significantly lower accuracy for dog DNA (68.9%). This performance disparity is attributed to the increasing evolutionary distance from human DNA. Additionally, a weighted naive bayes model that was trained on human data showed very high accuracy in predicting chimpanzee (99.3%) and dog (92.6%) DNA sequences. The results presented here show the significance of taking into account evolutionary relations and dataset features whenever developing and training classification models for genetic sequence analysis. The research extends the present research by evaluating the performance of several different algorithms on separate DNA databases, identifying strengths and weaknesses, and suggesting avenues for future research focusing on advanced feature engineering and algorithm selection for improved cross-species classification.

Keywords: *DNA classification, machine learning, K-mer, naive bayes.*

1. Introduction

DNA is a hereditary molecule found in every living species that determines the basic characteristics of bodies and acts as a genomic scheme for a progressing organism. It is comprised of a phosphate category, a sugar category, and nitrogenous compounds (adenine, cytosine, guanine, and thymine) that make up its double-stranded molecule. DNA analysis is decisive in the recognition of diseases, tracing the extent of infections, solving crimes, and paternity tests. Primers, human nucleotide sequences, are used for DNA synthesis, which is an essential tool in molecular biology (Momenzadeh et al., 2020). These primers have been used for the recognition and detection of parasites, bacteria, and viruses. The PCR technology amplifies the DNA segment of a preexisting virus, making it possible to detect it. DNA analysis is now one of the most popular topics in computational biology. DNA sequencing is a way of ordering nucleotides in a DNA fragment that sometimes exists in double strands (Nayak et al., 2022).

Different techniques have been designed for the application of DNA sequences for species identification purposes in different areas, including the score of correspondence, population genetic knowledge, and, particularly for species organizing features, phylogeny. DNA sequencing research is done through shotgun cloning and walking. Classification is one of the means of character recognition of single or groups of characters, like supermolecule sequences. There are several methods applied to classify these sequences into their particular division, secondary category, or family in order to eliminate alternatives, equalize their value, and finally categorize the sequence. Bioinformatics is the analysis of genetic material sequencing information, which refers to the identification of a DNA sequence's structural order. Categorization breaks up a gene into different regions; selected-effect function genes are categorized into functional DNA and rubbish DNA (Benson et al., 2009).

While unusable DNA loses the unselected-effect activity, healthy DNA possesses it. Efficient DNA sequences can be either precise, with the literal DNA choosing the nucleotide order, or indifferent, with the indifferent DNA choosing the existence or inactivity of an arbitrary pattern of DNA. Junk DNA is part of the organism's fitness and is below the variety neutrality, whereas waste DNA decreases it. Functional categories assigned to DNA regions may change over time. K-mers, substrings of length k, are applied in bioinformatics for sequence analysis and computational genomics. They are made up of nucleotides (A, T, C, and G) and are used to assemble DNA sequences (Benson et al., 2020; Solis-Reyes, 2009; Shadab et al., 2020). We apply several classifiers, among them are the Naive Bayes, Random Forest, K-Nearest Neighbour, and Decision Tree classifiers. The research considers three datasets: human, dog, and chimpanzee, based on protein coding sequences. The datasets are processed using k-mer

Authors information:
[a]Department of Computer Engineering, GES's R. H. Sapat College of Engineering, Management Studies and Research, Nashik, INDIA. E-mail: borsevinodkumar@gmail.com[1]; archana.vaidya@ges-coengg.org[2]
[b]Department of Computer Science and Engineering (Data Science), R. C. Patel Institute of Technology, Shirpur, INDIA. E-mail: manishavpatil2007@gmail.com[3]
*Corresponding Author: borsevinodkumar@gmail.com

Received: March, 2025
Accepted: July, 2025
Published: June, 2026

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

encoding and machine learning classifiers, and classified using scikit-learn to determine gene similarity and predict potential relationships. The dataset has two columns: sequence and class, where class is a pre-determined integer value depending on the protein coding sequence (Solis-Reyes et al., 2009). The research paper seeks to investigate the efficacy of machine learning algorithms in DNA classification of protein sequence data and add to the existing literature by comparing results with prior studies, identifying strengths and weaknesses of various algorithms, and proposing areas for improvement (Shadab et al., 2020).

2. Literature Review

Researchers identified new things to implement after evaluating significance, seeking speedy and precise completion. Past scholars contributed a lot to this research area. (Onesime et al., 2021) introduces a new approach to predicting genomic islands (GIs) in bacterial genomes based on seven sequence features extracted with the scikit-learn toolkit. The approach employs a chi-square test to identify significant features, which are subsequently input into a random forest algorithm. The method is appropriate to current approaches in standards of accuracy, precision, recall, and other measures. The article (Alotaibi et al., 2021) discusses the application of six machine learning algorithms (K-nn, RF, SVM, LR, SGD, and GNB) to predict DNA mixture contributors. The provided dataset includes DNA mixtures with a maximum of five contributors. Accuracy, F1-score, recall, and precision have been utilized to measure the metric of each algorithm. Logistic regression (LR) has the best accuracy (95%) for five-contributor mixtures. (Arowolo et al., 2021) Discusses a method for RNA-Seq data classification using genetic algorithm feature selection, decision tree, and K-nearest neighbors (KNN) classifiers. The method is applied to a set of Anopheles gambiae mosquitoes, a malaria vector. The feature selection using GA greatly enhances the performance of the two classifiers, with the decision tree having higher accuracy (98.3%) compared to KNN (88.3%). The paper also discusses literature related to machine learning approaches in the classification of gene expression. (Mathur et al., 2023) Research suggests a DNA sequence classification system for the identification of early diseases, based on NCBI database samples and convolutional neural networks for feature extraction and classification. (Arowolo et al., 2021) Research assesses the application of a genetic algorithm (GA) for feature selection in RNA sequencing data of the Anopheles gambiae malaria vector dataset with 88.3% classification accuracy for KNN and 98.3% for decision tree, showing its viability in bioinformatics. The article (Hamed et al., 2023) explains the application of machine learning models such as random forest, KNN, naïve bayes, decision tree, and SVM in classifying DNA sequences through pattern matching.

The approach (Alshayeji et al., 2023) applies natural language processing and machine learning to automatically detect viruses from human biospecimens at 98.6% classification accuracy, precision, recall, and detection rate. The research employs natural language processing and machine learning to automatically detect viruses in human biospecimens with 98.6% accuracy of classification, precision, recall, and detection rate. The study (Li, F. et al., 2023) introduces an EpiTEAmDNA architecture, a combination of transfer learning and ensemble learning methods, for representing DNA methylation type features in 15 species, outperforming current methods. The study (Yadav, V. et al., 2025) uses genome sequencing to study DNA structure and species similarity between humans, chimpanzees, and dogs, revealing 99.30% and 98.40% similarity, suggesting further research could deepen understanding of life sciences.

3. Methods

The study has two phases of research: preprocessing and postprocessing. The data preprocessing is dealt with in the preprocessing phase, while model learning and framework testing are covered under the article phase. Machine learning and NLP are applied in this research work to process the texting DNA sequence pattern into a series and to evaluate the various ML algorithms.

Data Set

The data file “human_data.txt” was downloaded from the Kaggle database and is used as a public dataset. The data includes two features: DNA sequencing and class. The data size is (4380, 2), consisting of 4380 samples and two features. It contains six gene family classes, including occurrences per class and numeric values. The data is converted from string to numeric classes. In order to demonstrate the occurrences for each class, a count plot was created. Transcription factor (class 6) contained the most data, with 1343 samples, and the Lon channel contained the least class, with 240 samples (Bhushan Bawankar et al., 2024).

Feature Extraction

DNA sequences are written as a set of shorter sub-sequences of length k, known as k-mers. These sub-sequences are similar to words in human language processing (NLP) and help understand DNA sequences and simplify analysis computation. In our model, k-mer encoding generates hexamer “words” with word length adjustable for various situations. In genomics, manipulations like "k-mer counting" are used. Python's natural language processing packages make this process easy. From each pattern string, a function (get_K-mers) is provided to gather all potential overlaid k-mers of a certain length. The data sequences are transformed into repeating k-mers of length six, and repeated operations are performed for every species DNA sequence in the dataset (Hamed et al., 2023).

DNA Classification

This section provides a classification model applied to human, chimpanzee, and dog DNA datasets. We utilize several classifiers; among them are the naive bayes, decision tree classifier, w-naive bayes classifier, random forest, and K-NN. Also, discuss various performance parameters utilized for the evaluation of classifier performance.

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

Naive Bayes Classifier

The bayes principle is applied in the naïve bayes technique, a machine learning statistical methodology, to solve classification issues. It uses an independent variable to establish the range of parameters and requires minimal training data (Peretz et al., 2024). Eq. (1), which employs the Gauss distribution as the decision function, is the most often used naïve bayes classifier (Zuhanda et al., 2025). Eq. (1) utilizes the most probable principle to provide the estimated mean and standard deviation.

$$P(v/y) = \frac{1}{\sqrt{2\pi\sigma^2}}. e^{(v-\mu)^2/2\sigma^2} \quad (1)$$

In order to classify data, naïve bayes considers a Gaussian distribution and computes the mean, variance, and prior probability for each class in training data (Wickramasinghe et al., 2021).

Weighted Naive Bayes Classifier

The NB algorithm assumes condition attributes are independent, which reduces the computational cost. However, practical application varies, and hence, classification accuracy is reduced when the default ownership weight is 1(Xia et al., 2021). In this study, we employ the weighted naïve bayes (WNB) algorithm to assign appropriate attributes based on their classification contribution, preserving computation speed and reducing the impact of attribute conditional independence on classifier performance (Ye, 2021; Wickramasinghe et al., 2021). The weighted NB is calculated using the following Eq. (2).

$$p(C_j/X) = \arg\max p(C_j) \prod_{i=1}^{n} p(A_i / C_j)^{wi} \quad (2)$$

Random Forest Classifier

RF is a technique for classification, an ensemble technique where a number of decision trees are parallel trained with bootstrapping and aggregation of the base learners through bagging. Thus, each decision tree is unequal because of the property of high variance of the decision tree. Aggregation of predictions of the final decision comes from individual base classifiers. The goodness of generalization of the classifiers is also much better and gives higher accuracy in comparison to the random forest, which overcomes overfitting (Onesime et al., 2021; Pazhanikumar et al., 2024). RF classifiers show more robust results in noisy data and are simpler to tune hyperparameters compared to DT classifiers. They give fast computation for the internal measure of variable importance, VIMP, which comes in handy with ranking variables, especially in large multi-dimensional genomic imbalance data sets (Rrmoku et al., 2022).

Decision Tree Classifier

Machine learning models can be classified into parametric and nonparametric algorithms. A decision tree model is a nonparametric model based on supervised learning, where the tree-like structure is used to represent the probability of an event based on training data. The model is arranged in a top-down tree structure, with nodes representing features and branches representing possible values (Benson et al., 2009). This process of building DTs from the training data is called DT induction, which can be represented using a conditional control statement for easy decision-making (Peretz et al., 2024)

K-Nearest Neighbour Classifier

The k-NN algorithm is a memory-based method that stores objects during training, requiring less computational effort than deep learning methods. It constructs an approximation of the objective function, which is different for each new data to be stored, which is advantageous when the objective function is complex (Corso et al., 2021). As a progressive method that stores new data only when it becomes accessible, K-NN may be employed to address challenging issues. Still, it cannot express itself precisely, a concise expression of the objects, and depends on the determination of the distance of any object from all training objects. The redundancy in attributes, along with redundant data, weakens the proposed model. This weakness can be compensated for by checking the reduction in the problem space or testing the variation of distance calculations (Wang et al., 2023).

4. Results and Discussion

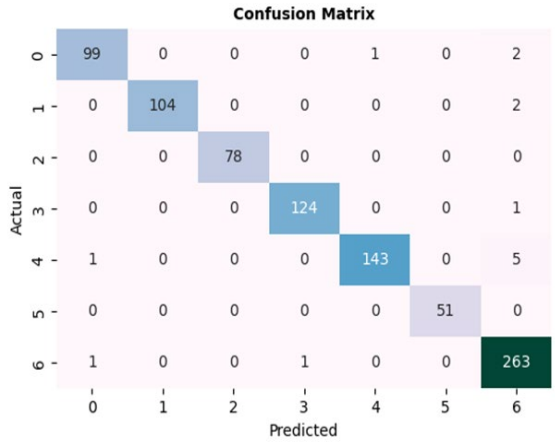
In this stage, we provide an analysis of the classification model performance by our results with human, chimpanzee, and dog DNA datasets.

Table 1. Performance analysis of various classifiers based on human DNA data.

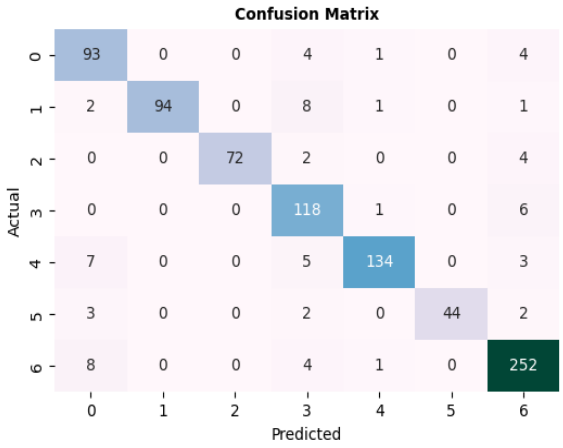
Classifier	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.984	0.984	0.984	0.984
Random Forest	0.918	0.926	0.918	0.919
K-Nearest Neighbour	0.824	0.874	0.824	0.827
Decision Tree	0.807	0.827	0.807	0.813

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

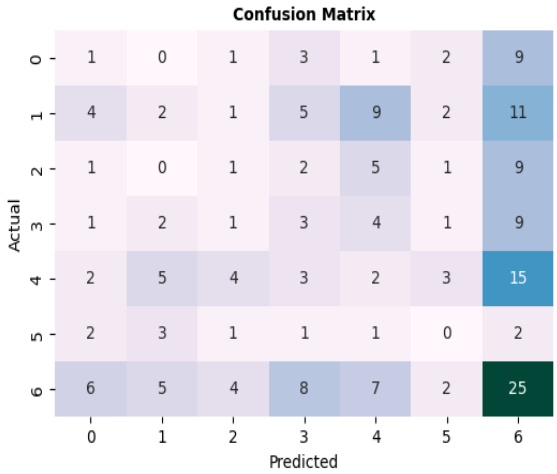
Table 2. Confusion matrix of various classifier based on Human DNA data.



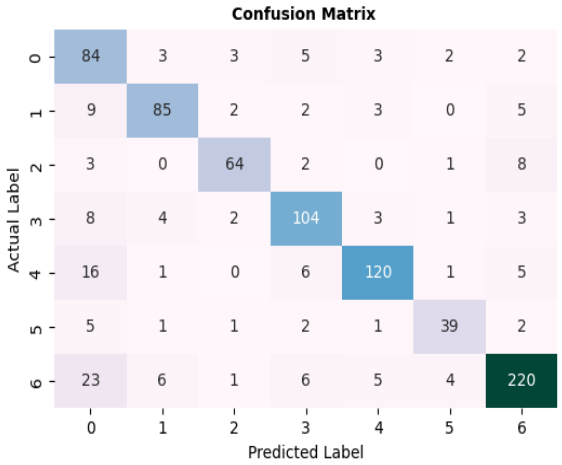
1. Naive bayes classifier



2. Random forest classifier



3. K-nearest neighbour classifier



4. Decision tree classifier

We utilize several classifiers, among them are the naive bayes, decision tree classifiers, k-nearest neighbour, and random forest. As a part of the standard parameters, the test performance for classifiers was also executed in terms of accuracy, precision, recall, and F1 score. We compute the receiver operating characteristic and the Area under the curve values in each classifier. Table 1 shows the actual performance of a classifier, and Table 2 shows the confusion matrix of all classifiers based on human DNA data.

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

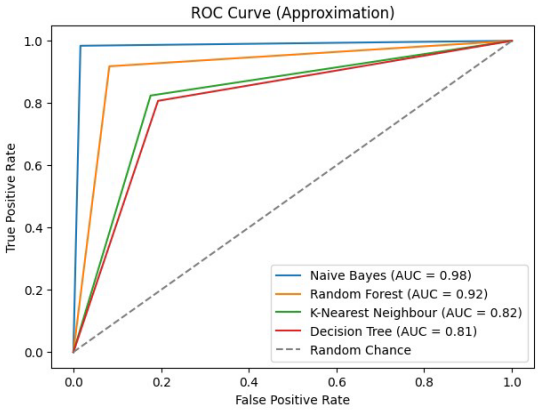


Figure 1. ROC and AUC of different classifier based on human DNA data.

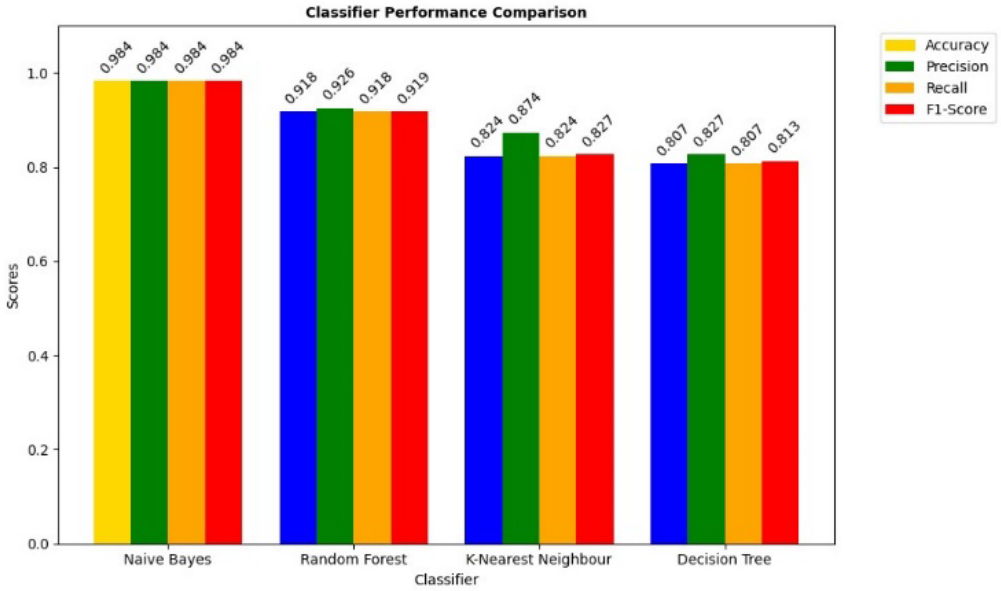


Figure 2. Comparison of various classifier metrics of human DNA data.

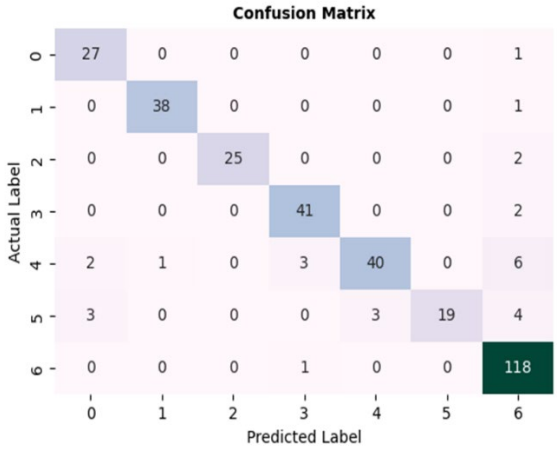
The naive bayes classifier performed better with an accuracy, precision, recall, and F1-score of 98.4 %. Random forest classifier performed fairly well with an accuracy of 91.8%. K-nearest neighbour and decision tree classifiers performed poorly, with an accuracy of 82.4% and 80.7 % respectively, as shown in Table 1.

Table 3. Performance analysis of various classifiers based on chimpanzee DNA data.

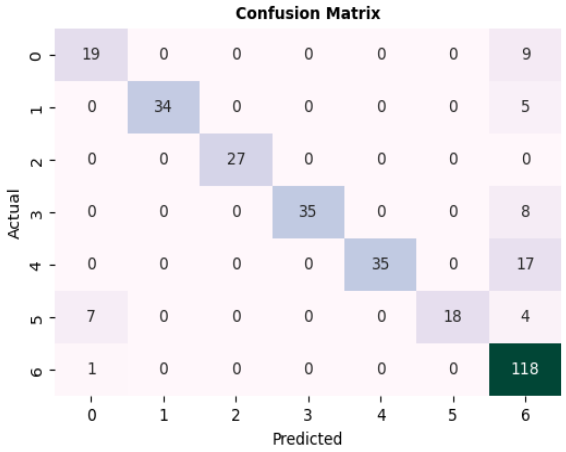
Classifier	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.914	0.920	0.914	0.911
Random Forest	0.843	0.874	0.843	0.841
Decision Tree	0.769	0.767	0.769	0.766
K-Nearest Neighbour	0.682	0.852	0.682	0.705

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

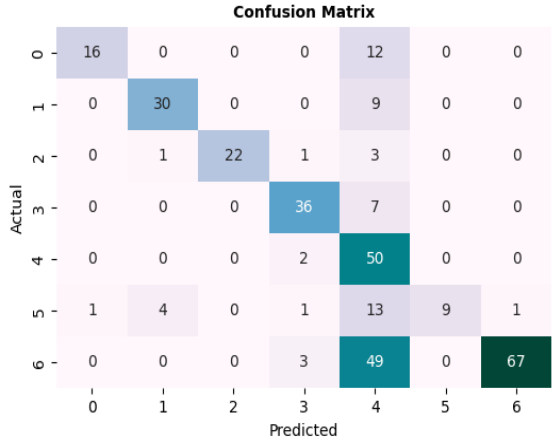
Table 4. Confusion matrix of various classifiers based on chimpanzee DNA data.



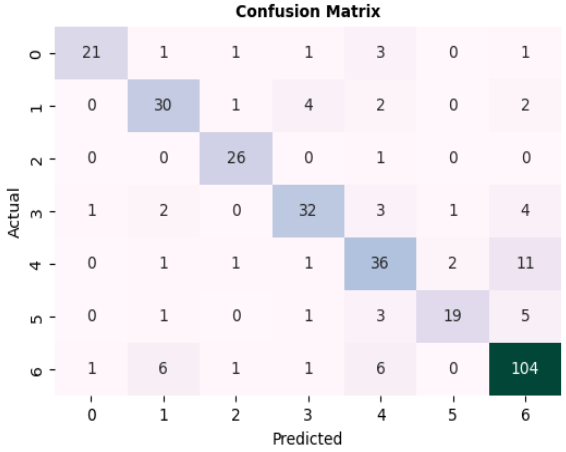
1. Naive bayes classifier



2. Random forest classifier



3. K-nearest neighbour classifier



4. Decision tree classifier

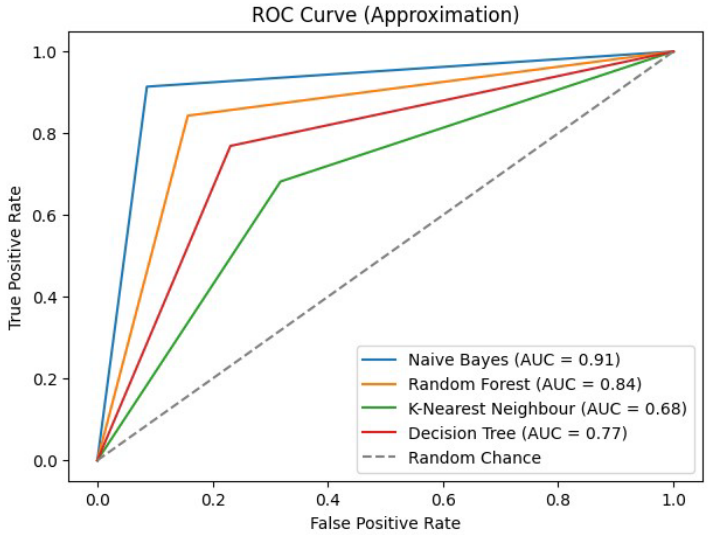


Figure 3. ROC and AUC of different classifiers based on chimpanzee DNA data.

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

These results indicate that, because of its ability to classify high-dimensional data and its underlying assumptions that align with the data's structure, the naive bayes classifier is the most appropriate for classifying human DNA data. Figures 1 and 2 are the graphical representations of ROC curves and comparison of different classifier metrics, respectively shown in Table 1.

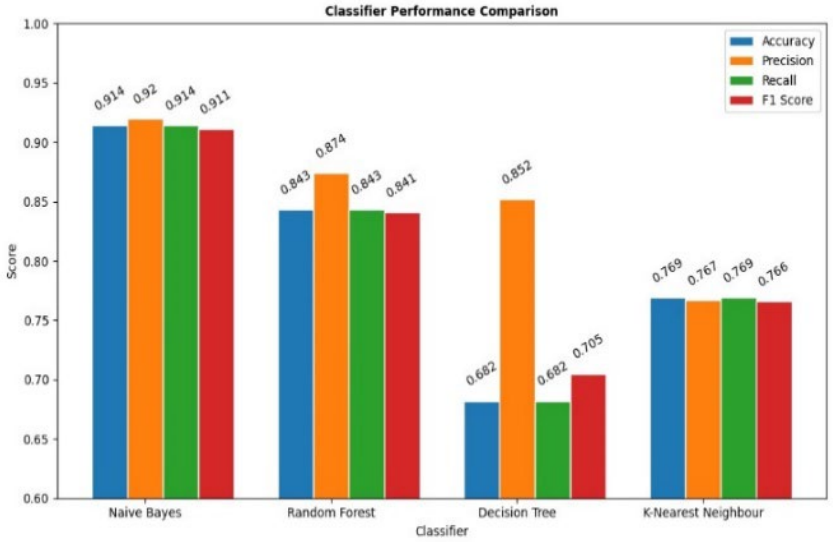


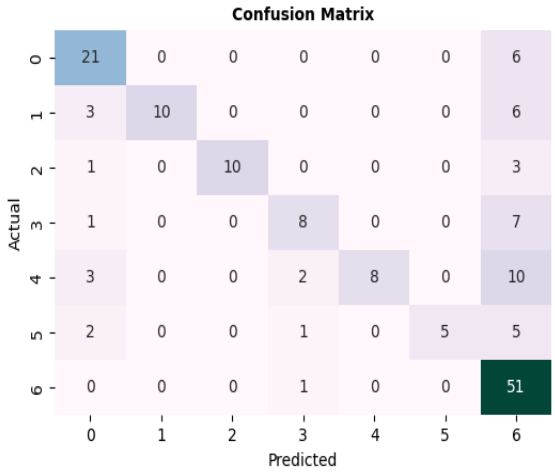
Figure 4. Comparison of various classifier metrics of chimpanzee DNA data.

Table 5. Performance analysis of various classifiers based on dog DNA Data.

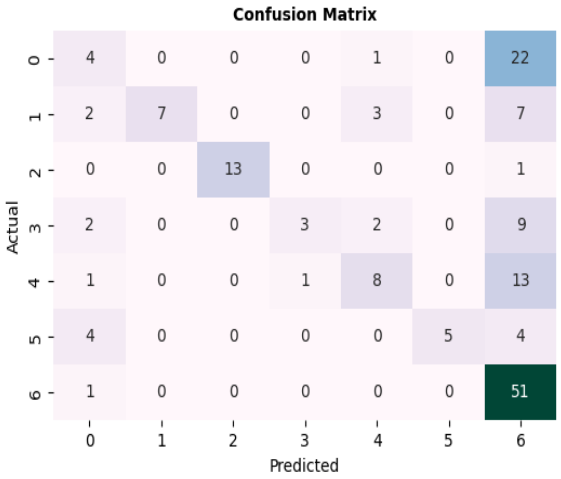
Classifier	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.689	0.781	0.689	0.673
Decision Tree	0.506	0.500	0.506	0.496
Random Forest	0.573	0.670	0.571	0.541
K-Nearest Neighbour	0.323	0.406	0.323	0.22

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

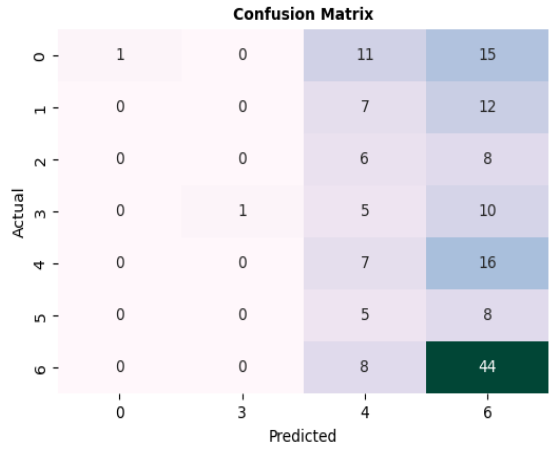
Table 6. Confusion matrix of various classifiers based on dog DNA Data.



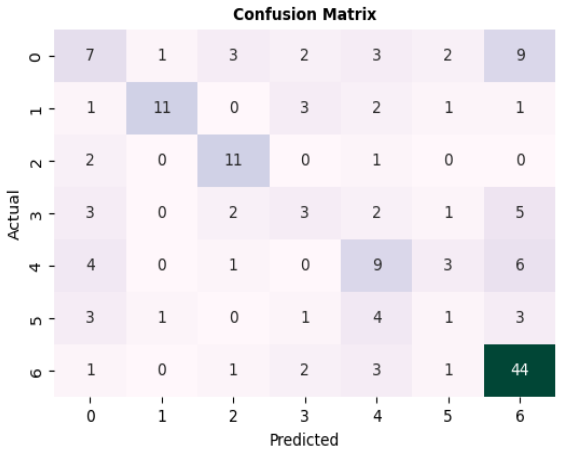
1. Naive bayes classifier



2. Random forest classifier



3. K-nearest neighbour classifier



4. Decision tree classifier

The performance of the classifiers on chimpanzee DNA data was generally lower compared to human DNA data. The Naive Bayes classifier was once more the best, at 91.4% as illustrated in Table 3. Random forest at 84.3%, the decision tree at 76.9%, and K-nearest neighbour at 68.2% had lower accuracies as indicated in Table 3. It could be the differences in the genetic structure and complexity between human and chimpanzee DNA that result in reduced performance in comparison with human data. Figure 3 and Figure 4 illustrate the ROC curves and the comparison of the classifier metrics for the chimpanzee DNA data shown in Table 3. Table 4 shows the confusion matrix of all classifiers.

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

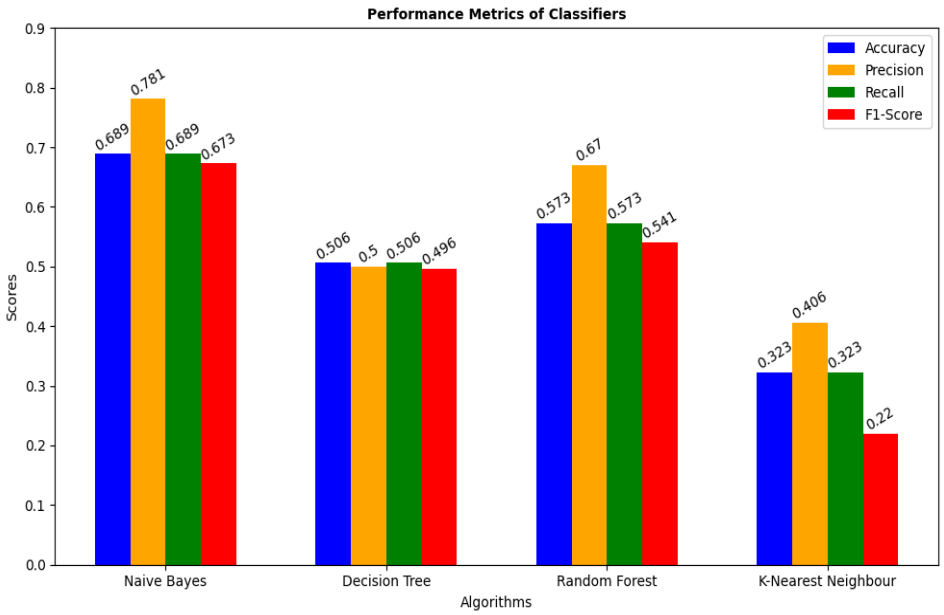


Figure 5. Comparison of various classifier metrics of dog DNA data.

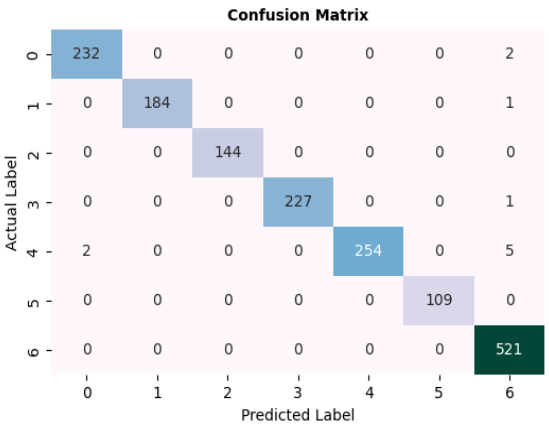
The classifiers performed substantially lower in dog DNA data compared to both human and chimpanzee data. The highest accuracy was observed for the naive bayes classifier at 68.9%, followed by the random forest classifier at 57.3% as shown in Table 5. The Decision Tree and K-Nearest Neighbour classifiers had poor accuracies of 50.6% and 32.3%, respectively, as shown in Table 5. This can be due to major differences in genetic structure and evolutionary distance between dogs and humans. Figures 5 and 6 compare the classifier metrics and ROC curves for dog DNA data shown in Table 5, and Table 6 shows the performance comparison and confusion matrix of all classifiers.

We used the trained weighted naive bayes human model from the human DNA data to predict chimpanzee and dog DNA types in order to further investigate the generalizability of our models. The model obtained high accuracy in predicting chimpanzee DNA at 99.3% and dog DNA at 92.6% Table 7. High accuracy in predicting chimpanzee DNA may indicate some similarity in genetic features between humans and chimpanzees that the model can effectively capture. Lower accuracy in predicting dog DNA is consistent with previous findings indicating greater genetic divergence between dogs and humans. Performance comparison of the Weighted Naive Bayes model for the three species shown in Figure 7. Table 7 and Table 8 show the Prediction metrics and confusion matrix of all classifiers

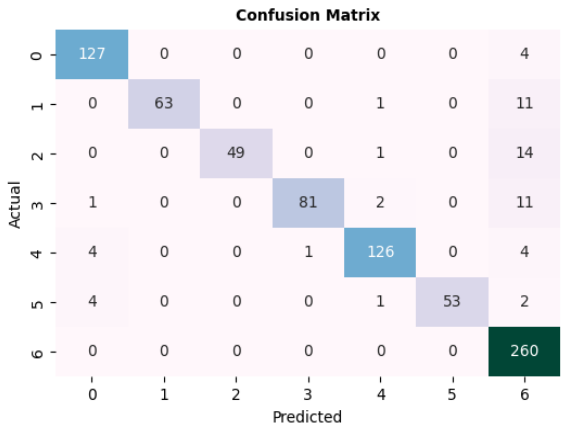
Table 7. Prediction metrics using a weighted naive bayes classifier (human model).

DNA Type	Accuracy	Precision	Recall	F1 Score
Chimpanzee	0.993	0.994	0.993	0.993
Dog	0.926	0.934	0.926	0.925
Human	0.996	0.996	0.996	0.996

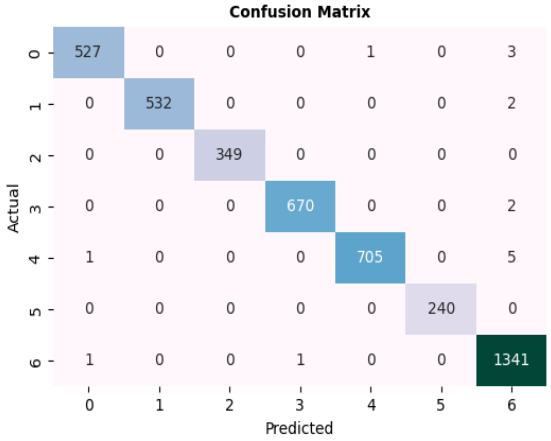
DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)



1. Chimpanzee DNA



2. Dog DNA



3. Human DNA

Table 8. Confusion matrix of weighted naive bayes classifier (human model).

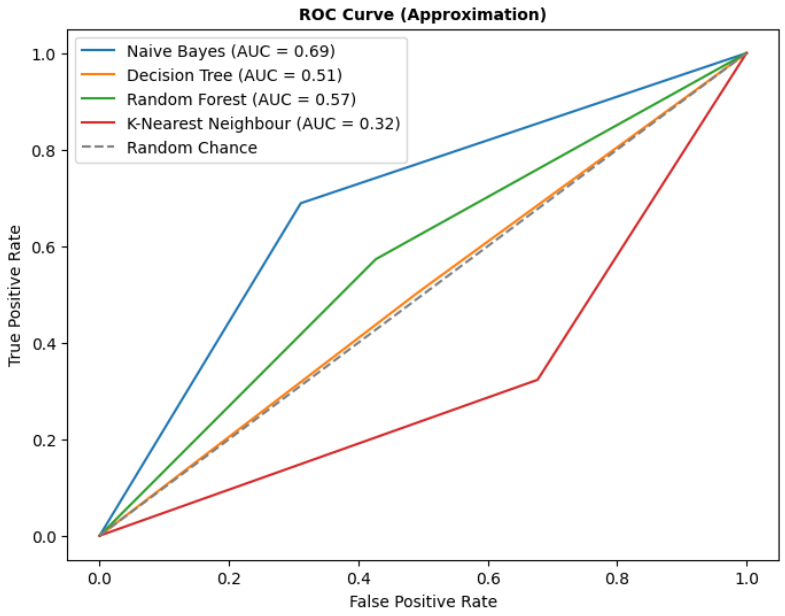


Figure 6. ROC and AUC of different classifiers based on dog DNA data.

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

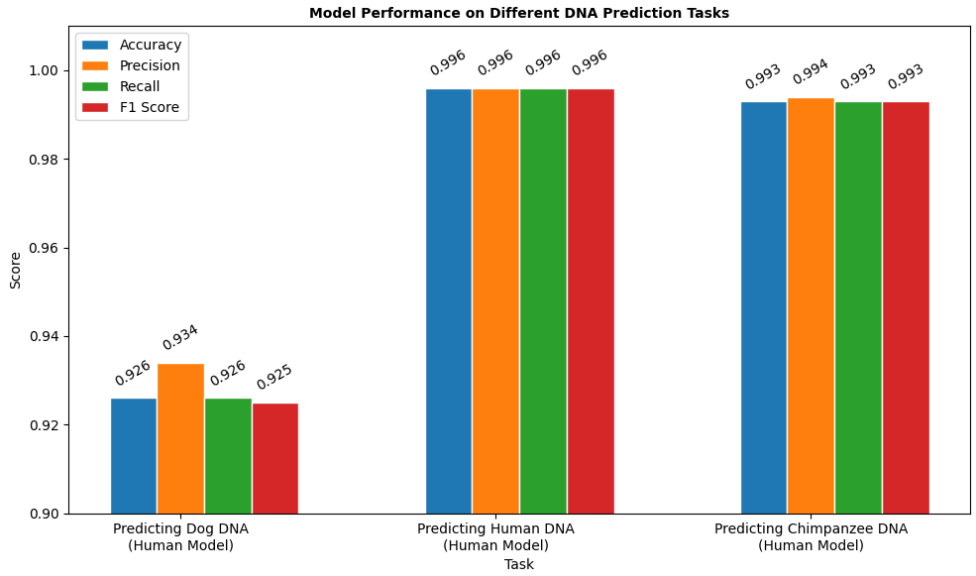


Figure 7. Performance comparison of human, chimpanzee, and dog DNA data using weighted naive byes (human model).

4. Conclusion

In conclusion, our results show that the weighted naive bayes classifier outperforms other classifiers for all three datasets. However, the performance is significantly reduced with an increase in evolutionary distance from human DNA. This shows the need to pay special attention to the characteristics of the DNA data while choosing and training classification models. Future research may include more advanced feature engineering techniques and other classification algorithms that may improve performance on more distantly related species.

7. References

Alotaibi, H., Alsolami, F., & Mehmood, R. (2021). DNA profiling: An investigation of six machine learning algorithms for estimating the number of contributors in DNA mixtures. *International Journal of Advanced Computer Science and Applications, 12*(11).

Alshayeji, M. H., Sindhu, S. C., & Abed, S. (2023). Viral genome prediction from raw human DNA sequence samples by combining natural language processing and machine learning techniques. *Expert Systems with Applications*, 218, 119641. https://doi.org/10.1016/j.eswa.2023.119641.

Arowolo, M. O., Adebiyi, M., Adebiyi, A. A., & OKesola, J. O. (2021). Predicting RNA-Seq data using genetic algorithm and ensemble classification algorithms. Indonesian Journal of *Electrical Engineering and Computer Science*, 21(2), 1073-1081.

Arowolo, M. O., Adebiyi, M. O., & Adebiyi, A. A. (2021). A genetic algorithm approach for predicting ribonucleic acid sequencing data classification using KNN and decision tree. *TELKOMNIKA, Telecommunication Computing Electronics and Control*, 19(1), 310-316.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2009). GenBank. *Nucleic Acids Research, 37*(suppl_1), D26-D31.

Beskopylny, A. N., Stel'makh, S. A., Shcherban', E. M., Mailyan, L. R., Meskhi, B., Razveeva, I., & Beskopylny, N. (2022). Concrete strength prediction using machine learning methods CatBoost, k-nearest neighbors, support vector regression. *Applied Sciences, 12*(21), 10864.

Bhushan Bawankar. (2024). Analysis of machine learning approaches for dna sequencing and classification: An optimized approach. *Communications on Applied Nonlinear Analysis, 31*(2s), 436–453. https://doi.org/10.52783/cana.v31.659.

Coscia, A., Dentamaro, V., Galantucci, S., Maci, A., & Pirlo, G. (2024). Automatic decision tree-based NIDPS ruleset generation for DoS/DDoS attacks. *Journal of Information Security and Applications, 82*, 103736.

Corso, M. P., Perez, F. L., Stefenon, S. F., Yow, K. C., García Ovejero, R., & Leithardt, V. R. Q. (2021). Classification of contaminated insulators using k-nearest neighbors based on computer vision. *Computers*, 10(9), 112.

Gao, L., Li, D., Liu, X., & Liu, G. (2022). Enhanced chiller faults detection and isolation method based on independent component analysis and k-nearest neighbors classifier. *Building and Environment*, 216, 109010.

Hamed, B. A., Ibrahim, O. A. S., & Abd El-Hafeez, T. (2023). Optimizing classification efficiency with machine learning techniques for pattern matching. *Journal of Big Data*, 10(1), 124.

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)

Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, 77(5), 5198-5219.

Li, F., Liu, S., Li, K., Zhang, Y., Duan, M., Yao, Z., Zhu, G., Guo, Y., Wang, Y., Huang, L., & Zhou, F. (2023). EpiTEAmDNA: Sequence feature representation via transfer learning and ensemble learning for identifying multiple DNA epigenetic modification types across species. *Computers in Biology and Medicine*, 160, 107030. https://doi.org/10.1016/j.compbiomed.2023.107030

Mathur, G., Pandey, A., & Goyal, S. (2023). A comprehensive tool for rapid and accurate prediction of disease using DNA sequence classifier. *Journal of Ambient Intelligence and Humanized Computing*, 14(10), 13869-13885.

Momenzadeh, M., Sehhati, M., & Rabbani, H. (2020). Using hidden Markov model to predict recurrence of breast cancer based on sequential patterns in gene expression profiles. *Journal of Biomedical Informatics*, 111, 103570.

Nayak, J., Mishra, M., Naik, B., Swapnarekha, H., Cengiz, K., & Shanmuganathan, V. (2022). An impact study of COVID-19 on six different industries: Automobile, energy and power, agriculture, education, travel and tourism and consumer electronics. *Expert Systems*, 39(3), e12677.

Onesime, M., Yang, Z., & Dai, Q. (2021). Genomic Island Prediction via Chi-Square Test and Random Forest Algorithm. *Computational and Mathematical Methods in Medicine*, 2021(1), 9969751.

Pazhanikumar, K., & KuzhalVoiMozhi, S. N. (2024). Remote sensing image classification using modified random forest with empirical loss function through crowd-sourced data. *Multimedia Tools and Applications*, 83(18), 53899-53921.

Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier–An ensemble procedure for recall and precision enrichment. *Engineering Applications of Artificial Intelligence*, 136, 108972.

Rrmoku, K., Selimi, B., & Ahmedi, L. (2022). Application of trust in recommender systems—utilizing naive Bayes classifier. *Computation*, 10(1), 6.

Shadab, S., Khan, M. T. A., Neezi, N. A., Adilina, S., & Shatabda, S. (2020). DeepDBP: deep neural networks for identification of DNA-binding proteins. *Informatics in Medicine Unlocked*, 19, 100318.

Solis-Reyes, S., Avino, M., Poon, A., & Kari, L. (2018). An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PloS One*, 13(11), e0206409.

Wang, A. X., Chukova, S. S., & Nguyen, B. P. (2023). Ensemble k-nearest neighbors based on centroid displacement. *Information Sciences*, 629, 313-323.

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, 25(3), 2277-2293.

Xia, X., & Yan, J. (2021). Construction of music teaching evaluation model based on weighted naïve bayes. *Scientific Programming*, 2021(1), 7196197.

Yadav, V., Kumar, V., Gupta, S., & Kaushik, V. D. (Eds.). (2025). Computational intelligence and its applications. *BENTHAM SCIENCE PUBLISHERS*. https://doi.org/10.2174/9789815313321125010101.

Ye, Y. (2024, May). Design and Implementation of an English Mobile Learning System Based on Weighted Naive Bayes. In 2024 *5th International Conference on Big Data and Informatization Education (ICBDIE 2024)* (pp. 187-196). Atlantis Press.

Zuhanda, M. K., Permata, L., & Ongko, E. (2025). Impact of Adaptive Synthetic on Naïve Bayes Accuracy in Imbalanced Anemia Detection Datasets. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi),* 9(1), 85-93.

DOI:https//doi.org/10.22452/mjs.vol45no2.1
Malaysian Journal of Science 45(2): 1-12 (June 2026)