JCBD
JOURNAL OF COMPUTERS AND DIGITAL BUSINESS
Volume 4, Number 2, May 2025, pp. 100-105
Homepage: https://jurnal.delitekno.co.id/index.php/jcbd

DNA Sequence Classification Using Machine Learning Models Based on k-mer Features

Afthar Kautsar

Program Studi Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

jjtaryoung21@gmail.com

https://doi.org/10.56427/jcbd.v4i2.762

ARTICLE INFO

Article History
Received: February 7, 2025
Revised: May 13, 2025
Accepted: May 28, 2025

Keywords
DNA Classification
Machine Learning
k-mer
SVM
Random Forest
Deep Neural Network

ABSTRACT

Cell-free DNA (cfDNA) has emerged as a promising biomarker in various clinical applications, particularly in cancer detection, prenatal diagnostics, and disease monitoring. Accurate classification of cfDNA sequences is crucial for improving diagnostic reliability and enabling timely clinical decisions. This study investigates the application of machine learning models—Decision Tree (DT), Support Vector Machine (SVM), and Deep Neural Network (DNN)—for classifying cfDNA sequences using k-mer-based feature extraction, with k set to 3. A total of 3,000 DNA sequences comprising both normal and tumor-derived samples were transformed into numerical feature vectors based on the frequency of 3-mer patterns. The models were trained and evaluated using standard metrics including accuracy, precision, recall, and F1-score. Experimental results demonstrate that the DNN model achieved the highest classification performance, effectively distinguishing between normal and tumor cfDNA. In contrast, the DT and SVM models exhibited relatively lower performance, particularly in identifying normal sequences. The study also addresses challenges such as class imbalance and limitations of simple k-mer representations. These findings highlight the potential of deep learning approaches in improving cfDNA sequence analysis and open avenues for future research using more complex models, larger datasets, and feature engineering techniques to enhance classification accuracy and clinical applicability

JCBD is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

1. Introduction

Cell-free DNA (cfDNA) refers to DNA fragments that circulate in the bloodstream, typically originating from cellular processes such as apoptosis or necrosis. In the context of cancer, cfDNA often contains critical information related to genetic mutations, making it a promising tool for non-invasive diagnostics. In recent years, cfDNA analysis has garnered significant attention in the fields of bioinformatics and precision medicine.

Given the vast volume of genetic data and the high biological complexity, computational approaches such as machine learning (ML) have emerged as effective solutions for extracting mutational patterns from cfDNA data. Models like Random Forest and Support Vector Machine have been widely used for biological data classification due to their robustness in handling high-dimensional datasets. Furthermore, Deep Neural Networks (DNN) have shown increasing potential in genomic data processing, particularly due to their capability to learn complex feature representations.

However, the primary challenge in cfDNA analysis lies in accurately distinguishing between normal and mutated DNA sequences. Therefore, this study aims to evaluate and compare the performance of ML models in classifying cfDNA fragments using a k-mer segmentation approach.

In their study Modern Deep Learning in Bioinformatics, highlighted how advancements in deep learning have significantly transformed bioinformatics analysis, particularly in processing DNA and RNA data[1]. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been proven to recognize patterns in genetic sequence data, enabling more accurate predictions in various bioinformatics applications, including DNA motif identification, genome mapping, as well as gene function classification and annotation. This study demonstrates that the implementation of deep learning not only enhances the accuracy of DNA sequence analysis but also accelerates the identification of genetic patterns, which previously required longer processing times using conventional methods.

In their study Bioinformatics Analysis for Circulating Cell-Free DNA in Cancer, demonstrated how bioinformatics analysis can be utilized for the detection and classification of cell-free DNA (cfDNA) in cancer studies[2]. Cell-free DNA, found in blood plasma, can serve as an important biomarker for early cancer diagnosis. This study applied a machine learning-based approach to analyze DNA sequences, showing that this method can improve sensitivity and specificity in detecting genetic variations associated with cancer. By utilizing machine learning algorithms, cfDNA analysis becomes faster and more accurate compared to conventional methods.

Explain that the k-mer-based approach in machine learning can enhance DNA classification accuracy by optimizing genetic data processing [3]. The study A Machine Learning Approach for Classifying LTR-Retrotransposons in Plant Genomes Using k-mer-Based Features develops a k-mer method for classifying LTR-retrotransposons in plant genomes, demonstrating that k-mer-based analysis enables more accurate identification of genetic elements [4]. In their research titled DNA Sequence Classification Using k-mer-Based Vector Representations, add that k-mer-based vector representations can transform DNA sequences into numerical vectors, thereby improving the efficiency of machine learning models in classifying genetic data [5].

The study An Open-Source Machine Learning Tool for Fast and Accurate HIV-1 Subtype Classification Using k-mer-Based Analysis applies a k-mer-based machine learning method for HIV-1 genome subtype classification, proving to be faster and more accurate than conventional methods [6]. In the study Optimization of Support Vector Machine (SVM) for Predicting Mutations in Hepatitis C Virus (HCV) DNA, optimized the SVM algorithm to detect mutations in HCV DNA using various kernels and the Levenshtein Distance method, demonstrating that hyperparameter optimization can improve accuracy up to 99.8% [7]. Meanwhile, in their research Implementation of Deep Neural Network Method on Classification of Type 2 Diabetes Mellitus Disease, implemented the Deep Neural Network (DNN) method for classifying Type 2 diabetes mellitus, showing that hyperparameter tuning enhances model performance, indicating that DNN has the potential to be applied in k-mer-based DNA sequence classification [8].

According to using machine learning and pattern matching techniques for DNA sequence classification can significantly improve both accuracy and efficiency [9]. In their study, the Linear SVM algorithm outperformed other algorithms, such as FLPM and PAPM, in terms of accuracy. Their findings indicate that the choice of pattern length plays a crucial role in both performance and time complexity. The rapid progress in deep learning models has significantly contributed to the field of modern bioinformatics. highlighted the crucial role of deep learning in processing large-scale biological and biomedical data, and discussed its future development in increasingly complex bioinformatics tasks [10]. In line with this, emphasized that [11], although deep learning originates from machine learning, its capabilities have now surpassed traditional methods in handling large and complex omics data, opening up opportunities for the development of more adaptive and efficient algorithms and theories. The foundational concept of deep neural networks (DNN), artificial neural systems with multiple layers designed to mimic the human brain's pattern recognition, was presented by [12]. These networks are widely used in decision-making processes across various fields.

In addition to theoretical applications, DNNs have shown tangible results in medical prediction tasks. For instance, developed a DNN model to detect diabetes with an impressive 88% accuracy [13], leveraging techniques such as data preprocessing, normalization, feature selection, and meticulous hyperparameter tuning. A similar study by [14] the effectiveness of a stacked autoencoder model for diabetes classification, achieving 86.26% accuracy on the Pima Indians dataset, highlighting the success of neural models in healthcare data classification. On a different note, introduced an innovative approach for DNA sequence classification that does not rely on traditional training parameters [15]. Instead, he utilized compression algorithms like Gzip, Brotli, and LZMA to enhance efficiency without sacrificing accuracy. This method provides an alternative that uses fewer computational resources compared to conventional machine learning techniques, marking a significant advancement in genomic data classification. This shift from using deep learning in health data analysis to DNA sequence classification demonstrates the broad potential of modern bioinformatics, spanning disease prediction and genomic analysis across species.

2. Research Methodology

The research methodology consists of several stages, including dataset preparation, preprocessing, model training and testing.

a. Dataset Preparation
The dataset used in this study consists of a DNA sequence comprising 126,033 base pairs (bp). The sequence was fragmented using a k-mer approach with a k value of 3, resulting in a total of 630 DNA fragments.

b. Preprocessing
Each DNA fragment was encoded into a numerical representation based on k-mer frequency. The data was then split into two sets: 80% (504 fragments) for training and 20% (126 fragments) for testing purposes.

c. Model Training and Testing
Three machine learning models were trained and evaluated:

1) Random Forest
2) Support Vector Machine (SVM)
3) Deep Neural Network (DNN)

Dataset Preparation
Preprocessing
Model Training and Testing

Figure 1. Research Methodology

3. Results and Discussion

The evaluation results revealed that the Random Forest model achieved the highest accuracy at 71%, followed by SVM with 70%, and DNN with 67%. Despite these differences, all three models demonstrated a similar pattern: they performed better in classifying mutated DNA fragments than in recognizing normal ones. For instance, Random Forest achieved a recall of 0.99 for the mutation class but only 0.08 for the normal class. This imbalance is likely due to the skewed class distribution in the dataset (88 mutation vs. 38 normal) and the possibly higher complexity of features in normal sequences, which the models failed to capture effectively. These findings highlight the need for better dataset balancing techniques and more advanced feature extraction strategies. While the DNN model showed potential as a deep learning-based approach, it may require further optimization in terms of architecture and hyperparameter tuning. On the other hand, although SVM failed to correctly predict any normal DNA fragments (recall = 0), it demonstrated high precision for the mutation class. Future work should focus on employing sampling methods such as SMOTE or ADASYN to balance the training data and applying techniques like sequence embeddings, attention mechanisms, or transformer-based models to better capture subtle patterns in normal DNA. These improvements may enhance generalization and classification robustness across diverse cfDNA samples.

Model performance was assessed using accuracy, precision, recall, and F1-score metrics.

```
DNA Length: 126,033 base pairs (bp)
Total Fragments: 630
Total Data Samples: 630
Training Data: 504 samples
Testing Data: 126 samples
```

a. Random Forest

Table 1. Model Evaluation Random Forest

	precision	**recall**	**f1-score**	**support**
normal	0.75	0.08	0.14	38
mutation	0.71	0.99	0.83	88
accuracy			0.71	126
marco avg	0.73	0.53	0.49	126
weighted avg	0.72	0.71	0.62	126

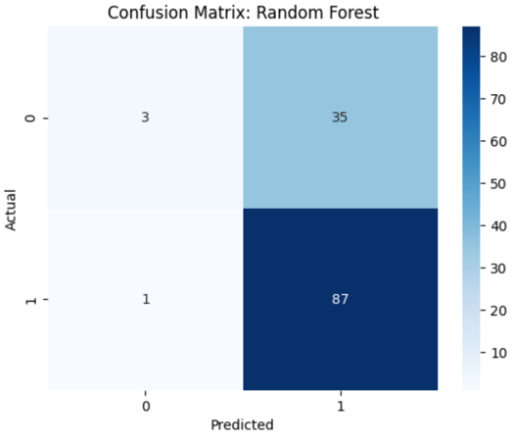

Figure 2. Confusion Matrix Random Forest

Confusion matrix of the Random Forest model showing high accuracy in predicting mutation class (87 true positives), but poor performance in classifying normal DNA (only 3 true negatives).

b. Support Vector Machine (SVM)

Table 2. Model Evaluation Support Vector Machine

	precision	recall	f1-score	support
normal	0.00	0.00	0.00	38
mutation	0.70	1.00	0.82	88
accuracy			0.70	126
marco avg	0.35	0.50	0.41	126
weighted avg	0.49	0.70	0.57	126

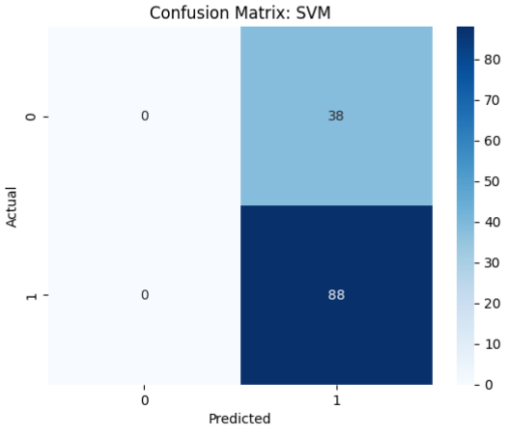

Figure 3. Confusion Matrix Support Vector Machine

Confusion matrix of the SVM model showing perfect prediction for mutation class (88 true positives) but complete failure in identifying normal DNA (0 true negatives, all 38 misclassified).

c. Deep Neural Network (DNN)

Table 3. Model Evaluation Deep Neural Network

	precision	recall	f1-score	support
normal	0.38	0.13	0.20	38
mutation	0.71	0.91	0.80	88
accuracy			0.67	126
marco avg	0.55	0.52	0.50	126
weighted avg	0.61	0.67	0.62	126

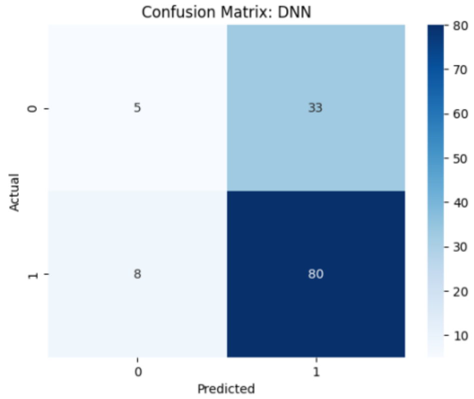

Figure 4. Confusion Matrix Deep Neural Network

d. Model Accuracy Summary

The evaluation results indicate that the Random Forest model achieved the highest classification accuracy at 71%, followed closely by the Support Vector Machine (SVM) with 70%, and the Deep Neural Network (DNN) with 67%. Despite these differences in overall accuracy, all three models showed a common tendency: they performed significantly better in identifying mutated cfDNA fragments than normal ones.

For instance, Random Forest achieved a recall of 0.99 for the mutation class, but only 0.08 for the normal class, revealing a strong classification imbalance. This discrepancy suggests that while the models are effective in detecting patterns associated with mutations, they struggle to generalize to the more complex or underrepresented normal DNA sequences. These results underscore the need for improved class balancing and feature representation in future work.

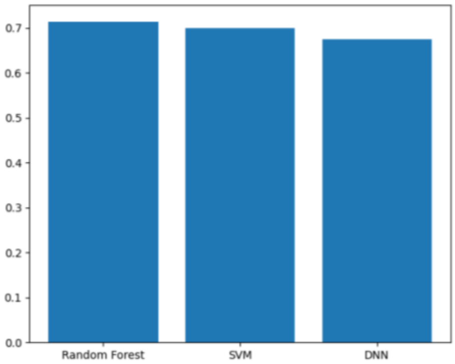

Figure 5. Accuracy Model

Distribution of the Top 20 Most Frequent k-mers in cfDNA Fragments

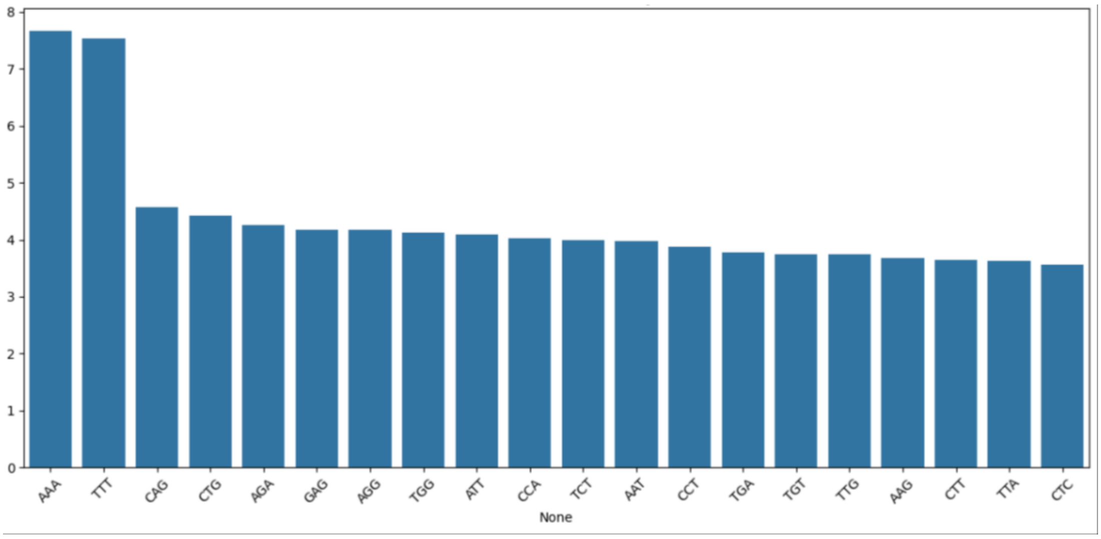

Figure 6. Distribution k-mers

This bar chart illustrates the frequency of the 20 most common k-mers (k=3) identified from the DNA sequence, highlighting dominant patterns that may contribute to mutation classification.

4. Conclusion

This study confirms that Random Forest, SVM, and DNN are viable machine learning models for classifying cell-free DNA (cfDNA) fragments based on k-mer features. Among the models evaluated, Random Forest achieved the highest overall accuracy. However, all models exhibited a noticeable bias toward accurately classifying mutated DNA, while struggling with normal sequences, likely due to class imbalance and feature complexity. These findings highlight the importance of improving data representation and distribution. Future research should investigate advanced data balancing techniques, such as SMOTE, as well as refined feature engineering or embedding-based representations to enhance classification robustness and model generalizability across diverse cfDNA datasets.

References

[1] X. G. Haoyang Li, Shuye Tian, Yu Li, Qiming Fang, Renbo Tan, Yijie Pan, Chao Huang, Ying Xu, "Modern deep learning in bioinformatics," *J. Mol. Cell Biol.*, vol. 12, no. 11, pp. 823–827, 2020.

[2] L. W. y Chiang-Ching Huang, Meijun Du, "Bioinformatics Analysis for Circulating Cell-Free DNA in Cancer," *Cancers (Basel)*, vol. 11, no. 6, pp. 1–15, 2019.

[3] S. Juneja, A. Dhankhar, A. Juneja, and S. Bali, "An Approach to DNA Sequence Classification Through Machine Learning: DNA Sequencing, K Mer Counting, Thresholding, Sequence Analysis," *Int. J. Reliab. Qual. E-Healthcare*, vol. 11, no. 2, pp. 1–15, 2022.

[4] G. I. Simon Orozco-Arias, Mariana S Candamil-Cortés, Paula A Jaimes, Johan S Piña, Reinel Tabares-Soto, Romain Guyot, "K-mer-based machine learning method to classify LTR-retrotransposons in plant genomes," 2021.

[5] Ü. M. Akkaya and H. Kalkan, "Classification of DNA Sequences with k-mers Based Vector Representations," 2021.

[6] A. Fiannaca, M. La Rosa, R. Rizzo, and A. Urso, "A k-mer-based barcode DNA classification methodology based on spectral representation and a neural gas network," *Inst. High-Performance Comput. Networking, Natl. Res. Counc. Italy, Viale delle Sci. Ed. 11, 90128 Palermo, Italy*, 2015.

[7] T. A. S. B. A. Kindhi and M. H. Purnomo, "Optimasi Support Vector Machine (SVM) untuk memprediksi adanya mutasi pada DNA Hepatitis C Virus (HCV)," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 3, pp. 1–6, 2018.

[8] D. H. G. M. Rizky, A. Pramuntadi, W. D. Prastowo, "Implementation of Deep Neural Network Method on Classification of Type 2 Diabetes Mellitus Disease," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 1043–1050, 2024.

[9] O. A. S. I. & Belal A. Hamed and T. A. El-Hafeez, "Optimizing classification efficiency with machine learning techniques for pattern matching," *J. Big Data*, vol. 10, no. 124.

[10] J. A. Malik YOUSEF, "Deep learning in bioinformatics," *Sci. Technol. Res. Counc. Turkey*, 2024.

[11] Z. Binhua Tang and A. K. Pan, Kang Yin, "Recent Advances of Deep Learning in Bioinformatics and Computational Biology," *Natl. Cent. Biotechnol. Inf.*, 2019.

[12] T. L. Kyongsik Yun, Alexander Huyen, "Deep Neural Networks for Pattern Recognition," 2018.

[13] M. Raheem, "Deep Neural Network to Predict Diabetes: A Data Science Approach," *Int. J. Recent Technol. Eng.*, vol. 9, no. 6, pp. 1–5, 2021.

[14] K. K, D. R. Edla, and V. Kuppili, "stacked autoencoders in deep neural networks," *Dep. Comput. Sci. Eng. Natl. Inst. Technol. Goa, India*, 2019.

[15] S. Ozan, "DNA Sequence Classification with Compressors," *digiMOST GmbH,* 2024.