Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 9, No. 1, Maret 2024 (23-30) **https://doi.org/10.32493/informatika.v9i1.39353**

Application of Traditional Machine Learning Techniques for the Classification of Human DNA Sequences: A Comparative Study of Random Forest and XGBoost

Gregorius Airlangga[1*]

[1]Information System Study Program, Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
e-mail: [1]gregorius.airlangga@atmajaya.ac.id
*Corresponding author

Submitted Date: January 24th, 2024 Reviewed Date: February 17th, 2024
Revised Date: February 27th, 2024 Accepted Date: March 30th, 2024

Abstract

This study evaluates the performance of hybrid machine learning models, specifically Random Forest and XGBoost, in classifying human DNA sequences into seven functional classes. Utilizing advanced feature vectorization techniques, this research addresses the challenges of analyzing high-dimensional genomic data. Both models were trained and tested on a dataset of annotated human DNA sequences, with an emphasis on generalizability to new, unseen data. Our results indicate that the Random Forest model achieved an accuracy of 87.98%, slightly outperforming the XGBoost model, which recorded an accuracy of 87.06%. These findings underscore the effectiveness of employing traditional machine learning techniques coupled with innovative data preprocessing for predictive modeling in genomics. The study not only enhances our understanding of genomic functionalities but also suggests robust methodologies for future genetic research and potential applications in personalized medicine. The implications of these results for improving classification accuracy and the recommendations for integrating more complex algorithms are also discussed.

Keywords: Machine Learning; DNA Sequence Classification; Random Forest; XGBoost; Genomic Data Analysis

1. Introduction

In the rapidly advancing field of genomics, the classification of human DNA sequences into their respective functional classes plays a pivotal role in understanding genetic functions and their implications in health and disease (Caudai et al., 2021; Jovic et al., 2022; Satam et al., 2023). Traditional methods for classifying genetic material have heavily relied on direct biological experimentation, which is often costly and time-consuming (He et al., 2022; Mobarak et al., 2023; Pan et al., 2022). With the advent of computational biology, numerous techniques have been developed to expedite and enhance the accuracy of genetic classification, thereby providing significant insights into genomic functionalities more efficiently (Basso et al., 2020; Fu et al., 2022; Zhang et al., 2021). Recent developments in machine learning have opened new avenues for analyzing and interpreting complex biological data (Dral & Barbatti, 2021; Rhodes et al., 2022; Tian et al., 2021). The use of algorithms such as Random Forests and Gradient Boosting Machines has shown promise in various bioinformatics applications, including gene expression analysis and disease prediction (Raslan et al., 2023). These methodologies, however, often encounter limitations in handling the high-dimensional and highly variable nature of DNA sequences (Thudumu et al., 2020). This has prompted researchers to explore more robust and sophisticated machine learning techniques that can capture the inherent complexities of genetic data more effectively (Greener et al., 2022; Kunduru, 2023; Patra et al., 2023). The urgency of developing improved computational tools for DNA sequence classification cannot be understated (Akbari Rokn Abadi et al., 2023). As we delve deeper into the genomic era, the ability to classify DNA sequences quickly and accurately into their correct functional categories is essential for timely advancements in personalized medicine, genetic

therapy, and disease prevention (Wang & Wang, 2023). Enhanced classification tools can lead to better understanding of disease mechanisms, which is crucial in the development of targeted treatments and interventions (Xie et al., 2021).

In surveying the literature, it is evident that while traditional models like Random Forest and XGBoost provide a strong baseline for classification tasks, they often do not fully utilize sequential or contextual information within DNA sequences (N. Y. Ahmed et al., 2024). Advances in deep learning, particularly in the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown superior performance in tasks where sequence context is important (S. F. Ahmed et al., 2023; Amiri et al., 2023; Shiri et al., 2023). However, these models require extensive computational resources and large datasets to achieve optimal performance, which can be a limiting factor in genomic studies (Rausch et al., 2021). This research seeks to bridge the gap between traditional machine learning models and deep learning techniques by implementing a hybrid approach that leverages the strengths of both methodologies. By integrating ensemble methods with deep learning architectures, this study aims to enhance classification accuracy while mitigating the limitations associated with each individual approach. The goal is to develop a model that not only provides high accuracy but also maintains computational efficiency, making it feasible for large-scale genomic studies.

The contributions of this research are threefold. First, we introduce a novel framework that combines the robustness of Random Forest and XGBoost with the sequence sensitivity of CNNs, creating a synergistic effect that enhances classification performance. Second, we conducted a comprehensive evaluation of this hybrid model against traditional machine learning models using a rich dataset of human DNA sequences classified into seven functional classes. Third, our study provides insights into the model's applicability and scalability in real-world genomic tasks, addressing the practical challenges in the field. The remainder of this article is structured as follows. Section 2 provides a detailed overview of the methods and materials used in this study, including data preparation, model architecture, and evaluation metrics. Section 3 presents the results of our experiments, highlighting the comparative performance of the hybrid model against traditional approaches. In addition, we also discuss the implications of our findings in the broader context of genomic research and computational biology. Finally, Section 4 concludes the article with a summary of our contributions and suggestions for future research in this domain.

2. Methods and Materials

2.1. Data Collection and Preprocessing

In the field of genomics, the efficient and accurate analysis of DNA sequences is paramount for understanding their biological roles and implications in health and disease. The dataset utilized in this research, sourced from a publicly accessible database (Vasani, 2022), includes human DNA sequences annotated into seven distinct functional classes, each reflecting the biological interactions and expressions of genomic segments within human cells. This classification is essential for exploring the functionalities of genes and their impact on various biological processes.

Upon collection, the dataset underwent an extensive preprocessing phase to adapt the inherently complex genetic data for analysis through both traditional and advanced machine learning techniques. This transformation was critical in structuring the data for algorithmic processing and effective predictive modeling. The initial step in this preprocessing involved converting the DNA sequences into k-mers of size 6, where a k-mer is a substring consisting of 'k' consecutive nucleotides. This size was strategically chosen to balance capturing sufficient biological information while maintaining manageable computational complexity. By employing 6-mer sizes, it became possible to extract meaningful biological patterns, such as motifs and genetic markers, which are crucial for understanding the sequences' functional properties.

Following the k-mer transformation, each sequence was further processed into a text-like format, treating each unique k-mer as a separate "word." This innovative approach allowed for the application of natural language processing (NLP) techniques to genomics. By representing DNA sequences as strings of text, we could apply a rich array of text analysis methodologies, traditionally used in language processing, to the analysis of genetic data. This included the use of the CountVectorizer method from the scikit-learn library, which transformed the textual k-mer data into a numerical format by counting the occurrences of each unique k-mer across the

sequences. The vectorization process was configured to consider combinations of four consecutive k-mers, thus enabling the models to detect and learn from broader contextual dependencies that might exist within the genetic sequences, beyond the immediate k-mer pairs.

Given a DNA sequence ($S = s_1 s_2 \dots s_n$), where each (s_i) represents a nucleotide, the k-mer transformation process for a k-mer size of (k) is defined as $S_k = \{s[i: i+k-1] \mid i = 1 \text{ to } n-k+1\}$, where each ($s[i: i+k-1]$) represents a k-mer generated from the sequence (S). After transforming the DNA sequence into k-mers, these k-mers are treated as individual "words" to utilize natural language processing techniques. The CountVectorizer, set with an n-gram range of 4, converts these "words" into a numerical feature vector ($\boldsymbol{v}(D)$) for each DNA sequence $v(D) = (c_1, c_2, \dots, c_m)$ where (c_j) is the count of the j-th n-gram (a sequence of 4 consecutive k-mers) in (D), and (m) is the total number of possible distinct n-grams formed from all k-mers in the dataset. This vectorization facilitates the application of machine learning algorithms by representing the genetic sequences in a high-dimensional sparse matrix format, capturing both the frequency and contextual relationships of the genetic features.

2.2. Feature Preprocessing and Model

In this study, feature vectorization played a crucial role in preparing the DNA sequences for machine learning analysis. Utilizing the CountVectorizer from the scikit-learn library, we transformed the k-mer based textual representation of DNA sequences into numerical features suitable for machine learning models. The vectorization process involved setting an n-gram range of 4, which allowed the model to consider combinations of four consecutive k-mers, thereby capturing a broader sequence context. This method transformed the sequence data into a high-dimensional sparse matrix that represents the frequency of each n-gram across the dataset, effectively converting genetic information into a format that traditional machine learning algorithms could process efficiently.

For model development, we focused on employing traditional machine learning models known for their robustness and efficacy in handling tabular, sparse data. Specifically, we utilized Random Forest and XGBoost classifiers, both highly regarded for their performance in various classification tasks. The Random Forest model was configured with varying numbers of decision trees (10, 50, 100) and different levels of tree depth (None, 10, 20, 30), enabling the model to capture data intricacies at multiple granularities. On the other hand, the XGBoost model's parameters were meticulously adjusted, including the maximum depth (5, 10, 15), number of estimators (100, 200), and learning rate (0.01, 0.1). These settings were optimized to balance the model's learning capacity and its generalization to prevent overfitting.

Both models were subjected to a rigorous process of hyperparameter tuning using GridSearchCV, a method that systematically explores a range of parameter combinations to identify the configuration that yields the highest classification accuracy. This step was crucial in ensuring that the models were not only well-suited to our specific dataset but also optimized for performance, leading to more reliable and accurate classification outcomes. Through this methodical approach to feature vectorization and model development, we aimed to harness the power of traditional machine learning techniques to enhance the predictive modeling of DNA sequence functionalities.

In the vectorization process, given a set of DNA sequences, each transformed into a sequence of k-mers, the CountVectorizer converts these sequences into a feature matrix. The process can be mathematically described as $X = \text{CountVectorizer}(S, \text{ngram\_range} = 4)$, where ($X$) is the sparse feature matrix, and (S) represents the collection of all k-mer based textual representations of the DNA sequences. The $ngram\_range\ of\ 4$ indicates that the feature $matrix(X)$ includes counts of each unique sequence of four consecutive k-mers, thereby capturing a broader sequence context within the DNA.

For the development of machine learning models, the hyperparameter settings for the Random Forest and XGBoost classifiers are optimized using GridSearchCV. This optimization can be expressed as $\hat{\theta} = \text{argmax}_{\theta \in \Theta}\big(\text{CV}(\text{RandomForest or XGBoost}, \theta)\big)$ where (θ) represents the hyperparameters such as number of trees, tree depth for Random Forest, and max depth, number of estimators, and learning rate for XGBoost. (Θ) denotes the hyperparameter space, and (CV) represents the cross-validation procedure used to evaluate each model configuration's performance. These equations

succinctly capture the computational processes and optimizations described in the study, providing a clear mathematical framework for the feature vectorization and model development phases.

2.3. Evaluation Metrics

In the evaluation phase of our study, we assessed the performance of our models using standard classification metrics to ensure a comprehensive understanding of each model's predictive abilities. Specifically, we utilized accuracy as our key metrics. This metrics provided a balanced view of both the correctness and robustness of the models in classifying DNA sequences into their respective functional categories. The evaluation was rigorously performed on a separate test set, which constituted 25% of the total dataset. This approach was deliberately chosen to ensure that the performance assessment reflected the models' ability to generalize to new, unseen data rather than just memorizing the training set. By isolating a portion of the data for testing purposes, we aimed to mimic real-world scenarios where the model would encounter data it has not previously analyzed, thereby providing insights into how well each model could potentially perform in practical applications. This method of evaluation is crucial in the field of computational biology, where the ability to accurately predict across diverse and variable genetic data can significantly impact the understanding and treatment of genetic-based diseases. The evaluation metrics used in our study are defined in equations (1). Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

where (TP) is the number of true positives, (TN) is the number of true negatives, (FP) is the number of false positives, and (FN) is the number of false negatives. Furthermore, precision (or Positive Predictive Value) measures the accuracy of positive predictions. This metrics were applied to a separate test set, constituting 25% of the total dataset, to evaluate each model's ability to generalize to new, unseen data and to ensure that the performance assessment did not merely reflect memorization of the training set data.

3. Results and Discussion

In this study, we employed traditional machine learning techniques to classify human DNA sequences into one of seven functional classes, leveraging a robust feature extraction method that converted DNA sequences into numerical vectors. The models evaluated, namely Random Forest and XGBoost, were optimized through a systematic hyperparameter tuning process. The Random Forest classifier achieved the best performance, with a classification accuracy of approximately 87.98%. In contrast, the XGBoost model demonstrated a slightly lower accuracy of 87.06% as presented in the table 1. These results underscore the effectiveness of ensemble learning techniques in handling the complexities associated with high-dimensional genomic data.

The superior performance of the Random Forest model can be attributed to its ability to handle the variance in the dataset effectively, thus minimizing overfitting—a common challenge in genomic sequence classification. Random Forest, by averaging multiple deep decision trees, each trained on different parts of the dataset, reduces the risk of stumbling on misleading patterns that might not generalize well to unseen data. This characteristic is particularly beneficial in genomic applications where the diversity of data can lead to significant variability in model performance. On the other hand, the XGBoost model, while slightly less accurate, also showcased strong performance, reinforcing the utility of gradient boosting frameworks in predictive modeling. XGBoost's lower performance compared to Random Forest in this scenario could be related to its propensity to overfit, especially when the hyperparameters are not perfectly tuned for the specific traits of the dataset. Despite this, XGBoost's high scalability and speed make it a valuable tool, particularly in larger datasets where execution time becomes critical.

The differences in performance between the two models also highlight the importance of model selection based on dataset characteristics and the specific requirements of the genomic classification task. While Random Forest offers robustness and generalizability, XGBoost provides efficiency and speed, with potentially higher performance given optimal parameter tuning. These findings have significant implications for genomic research, particularly in the development of computational tools for genetic data analysis. The ability of both Random Forest and XGBoost to effectively classify

complex genetic sequences into functional classes suggests that machine learning can serve as a powerful tool in genomic annotation and disease research. By automating the classification of genetic data, researchers can identify potential genetic markers more quickly and accurately, leading to faster insights into genetic functions and their implications for diseases.

Table 1. Accuracy Results

Model	Accuracy
XGBoost	87.06 %
Random Forest	87.98 %

4. Conclusion

This study demonstrated the effectiveness of traditional machine learning models, specifically Random Forest and XGBoost, in classifying human DNA sequences into functional classes. Through rigorous preprocessing, feature vectorization, and systematic model optimization, both models achieved commendable classification accuracies, with Random Forest slightly outperforming XGBoost. The results affirm the potential of ensemble learning methods to address complex problems in genomics, particularly in the classification of high-dimensional and intricate genetic data.

Random Forest's superior performance highlights its robustness and ability to minimize overfitting, a common challenge in genomic sequence analysis. This trait makes it particularly useful for genomic applications where the accuracy and generalizability of predictions are critical. Conversely, XGBoost, while slightly less effective in this specific instance, remains a valuable tool for its efficiency and scalability, attributes that are crucial for handling larger genomic datasets.

The findings from this research contribute to the ongoing efforts to integrate machine learning into genomic studies, offering insights that could enhance the development of computational tools for genetic data analysis. These tools are essential for advancing our understanding of genetic functionalities and their implications for health and disease, potentially accelerating the discovery of genetic markers and aiding in the development of personalized medicine.

5. Future Work

Despite the successes reported, this study opens several avenues for future research. Further exploration into hybrid models that combine the strengths of multiple machine learning techniques could lead to improvements in classification performance and insights into genetic data. Additionally, incorporating more sophisticated natural language processing techniques to handle the textual representation of DNA sequences might enhance the ability to capture more complex biological patterns. Moreover, expanding the dataset to include more varied genetic sequences and functional classes could improve the robustness and applicability of the models.

This expansion would provide a more comprehensive understanding of the models' performance across different genomic contexts and help refine their predictive capabilities. Ultimately, continuing to refine and adapt machine learning approaches for genomic classification will be vital as we seek to uncover more about the vast and complex landscape of the human genome. The integration of computational and biological sciences holds the promise of significant breakthroughs in our understanding of genetics and its impact on human health.

References

Ahmed, N. Y., Alsanousi, W. A., Hamid, E. M., Elbashir, M. K., Al-Aidarous, K. M., Mohammed, M. & Musa, M. E. M. (2024). An Efficient Deep Learning Approach for DNA-Binding Proteins Classification from Primary Sequences. International Journal of Computational Intelligence Systems, 17(1), 1–14.

Ahmed, S. F., Alam, M. S. Bin, Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A. B. M. & Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review, 56(11), 13521–13617.

Akbari Rokn Abadi, S., Mohammadi, A. & Koohi, S. (2023). A new profiling approach for DNA sequences based on the nucleotides' physicochemical features for accurate analysis of SARS-CoV-2 genomes. BMC Genomics, 24(1), 266.

Amiri, Z., Heidari, A., Navimipour, N. J., Unal, M. & Mousavi, A. (2023). Adventures in data analysis: A systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems. Multimedia Tools and Applications, 1–65.

Basso, M. F., Arraes, F. B. M., Grossi-de-Sa, M., Moreira, V. J. V., Alves-Ferreira, M. & Grossi-de-Sa, M. F. (2020). Insights into genetic and molecular elements for transgenic crop

development. Frontiers in Plant Science, 11, 509.

Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., Via, A. & Colombo, T. (2021). AI applications in functional genomics. Computational and Structural Biotechnology Journal, 19, 5762–5790.

Dral, P. O. & Barbatti, M. (2021). Molecular excited states through a machine learning lens. Nature Reviews Chemistry, 5(6), 388–405.

Fu, J. M., Satterstrom, F. K., Peng, M., Brand, H., Collins, R. L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S. P. & others. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nature Genetics, 54(9), 1320–1331.

Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. (2022). A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology, 23(1), 40–55.

He, W., Liu, T., Han, Y., Ming, W., Du, J., Liu, Y., Yang, Y., Wang, L., Jiang, Z., Wang, Y. & others. (2022). A review: The detection of cancer cells in histopathology based on machine vision. Computers in Biology and Medicine, 146, 105636.

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F. & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. Clinical and Translational Medicine, 12(3), e694.

Kunduru, A. R. (2023). Machine Learning in Drug Discovery: A Comprehensive Analysis of Applications, Challenges, and Future Directions. International Journal on Orange Technologies, 5(8), 29–37.

Mobarak, M. H., Mimona, M. A., Islam, M. A., Hossain, N., Zohura, F. T., Imtiaz, I. & Rimon, M. I. H. (2023). Scope of machine learning in materials research—A review. Applied Surface Science Advances, 18, 100523.

Pan, X., Lin, X., Cao, D., Zeng, X., Yu, P. S., He, L., Nussinov, R. & Cheng, F. (2022). Deep learning for drug repurposing: Methods, databases, and applications. Wiley Interdisciplinary Reviews: Computational Molecular Science, 12(4), e1597.

Patra, P., Disha, B. R., Kundu, P., Das, M. & Ghosh, A. (2023). Recent advances in machine learning applications in metabolic engineering. Biotechnology Advances, 62, 108069.

Raslan, M. A., Raslan, S. A., Shehata, E. M., Mahmoud, A. S. & Sabri, N. A. (2023). Advances in the Applications of Bioinformatics and Chemoinformatics. Pharmaceuticals, 16(7), 1050.

Rausch, T., Rashed, A. & Dustdar, S. (2021). Optimized container scheduling for data-intensive serverless edge computing. Future Generation Computer Systems, 114, 259–271.

Rhodes, C. J., Sweatt, A. J. & Maron, B. A. (2022). Harnessing big data to advance treatment and understanding of pulmonary hypertension. Circulation Research, 130(9), 1423–1444.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G. & others. (2023). Next-generation sequencing technology: current trends and advancements. Biology, 12(7), 997.

Shiri, F. M., Perumal, T., Mustapha, N. & Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. ArXiv Preprint ArXiv:2305.17473.

Thudumu, S., Branch, P., Jin, J. & Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data, 7, 1–30.

Tian, T., Yang, Z. & Li, X. (2021). Tissue clearing technique: Recent progress and biomedical applications. Journal of Anatomy, 238(2), 489–507.

Vasani, N. (2022). Human DNA Data. https://www.kaggle.com/datasets/neelvasani/humandnadata

Wang, R. C. & Wang, Z. (2023). Precision medicine: Disease subtyping and tailored treatment. Cancers, 15(15), 3837.

Xie, W., He, M., Yu, D., Wu, Y., Wang, X., Lv, S., Xiao, W. & Li, Y. (2021). Mouse models of sarcopenia: classification and evaluation. Journal of Cachexia, Sarcopenia and Muscle, 12(3), 538–554.

Zhang, X., Chen, S., Shi, L., Gong, D., Zhang, S., Zhao, Q., Zhan, D., Vasseur, L., Wang, Y., Yu, J. & others. (2021). Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. Nature Genetics, 53(8), 1250–1259.