GR-504

SYNTHETIC DNA SEQUENCE GENERATION AND CLASSIFICATION FOR SPECIES DISCRIMINATION

Abstract

Generative models have recently achieved breakthroughs in various domains due to enhancements in machine learning algorithms and increased computational power. However, genomic data still faces challenges, such as small data sizes, imbalances, and biases caused by the rarity of diseases, the cost of tests, and concerns about privacy and security. These challenges affect the performance of machine learning classifiers, especially when dealing with imbalanced data, where some data is more prevalent than others. To overcome the constraints associated with genomic data, we propose the creation of a Generative Adversarial Network (GAN). The key objective of this GAN is to augment both the quantity and diversity of genomic data by modifying existing samples within the data set instead of exclusively depending on acquiring new samples. This study presents two methods: the initial one entails generating synthetic DNA sequences using a GAN, whereas the second one employs a Random Forest Classifier, Support Vector Machine (SVM), and Logistic Regression with k-mer encoding for the classification of DNA sequences. The Conditional GAN architecture is employed in crafting the GAN model, which undergoes training with the available data. Assessing the proposed model using various machine learning algorithms reveals that the SVM linear classifier attains the highest accuracy and F1 score among the algorithms tested. This outcome suggests that synthetic data is more effective than original data in improving the performance of classification models. The study underscores the potential of synthetic data in addressing challenges associated with genomic data.

Keywords: Generative Adversarial Networks, Random Forest, Support Vector Machine, Logistic Regression, DNA sequence classification.

Introduction

This work aimed to classify DNA sequences effectively using machine learning techniques.

Fig 1. A single DNA sequence of Black rat data

We used three different supervised learning methods to perform the classification task. The results showed moderate accuracy levels with these methods. One of the main challenges was the need for more sufficient actual data. We observed that the classifier needed help identifying the minority class because it had less data than the majority class. The imbalanced data problem is a well-known issue where the standard classifier tends to favor the majority class, leading to neglect of the minority class. In response to the imbalanced data problem, synthetic data was generated. The realm of generative artificial intelligence (GenAI), particularly generative deep learning, is revolutionizing various scientific and technological fields. An essential advancement in this area is the creation of GANs. Over the past decade, generative models have been explored and applied across many machine-learning domains. Additionally, there have been applications in genetics, exemplified by a study focused on using deep generative models to generate DNA sequences. Ethical and logistical constraints frequently limit data collection's size, diversity, and speed.

A model was trained to understand the intricate distributions of real genomic data sets using deep generative adversarial networks. This enabled the generation of new, high-quality synthetic genomes with minimal privacy loss. Our study demonstrates that these synthetic genomes maintain characteristics of the source dataset, including GC content.

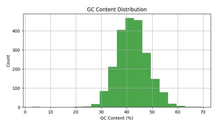

Fig 2. GC content of Black rat data

Generative models and synthetic genomes can prove beneficial in genetic studies by offering a rich yet concise representation of existing genomes and providing high-quality, easily accessible, and anonymous alternatives for private databases.

Materials and Methods

The GAN model design comprises three main components: the generator, the discriminator, and the overall GAN architecture. After data preprocessing, the sequences are split into training and testing sets, including cleaning, one-hot encoding, and GC content calculation.

Generative Adversarial Networks:

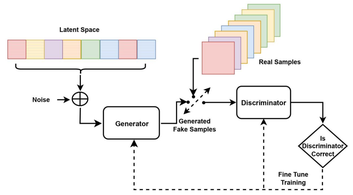

Fig 3. The Structure of GAN

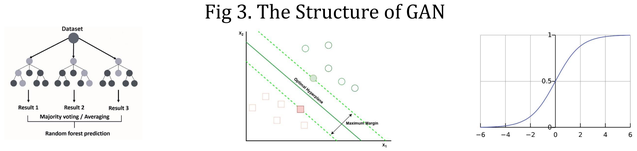

Fig 4. Random Forest Classifier, Support Vector Machine (SVM), and Logistic Regression

Results

Table 1: Experimental values setup

Parameter	Black rat	Human
Learning rate	0.0002	0.0001
Batch size	64	64
Epochs	200	300
Number of real data	2173	705
Number of synthetic data	827	2295

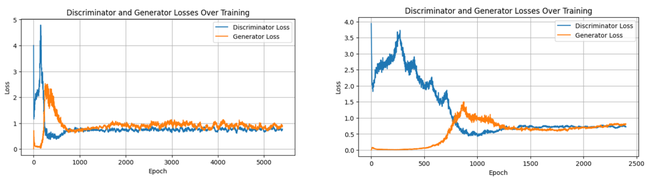

Fig 5. Discriminator and Generator Losses Over Training for Black Rat & Human

In the case of the rat dataset, the GAN training exhibited a relatively lower discriminator loss of 0.805 and a generator loss of 0.909 after 200 epochs, indicating a good convergence. The average test loss of 0.746 for rat sequences suggests that the generated synthetic sequences align well with the actual rat DNA data. Conversely, for the human dataset, the GAN training showed a higher discriminator loss of 1.209 but a lower generator loss of 0.560 after 300 epochs, indicating some challenges in convergence. However, the average test loss of 0.750 for human sequences is comparable to the rat dataset. These results demonstrate that the GAN-based synthetic sequence generation process can produce sequences that resemble real DNA data for rats and humans.

The results from the classification models applied to synthetic data demonstrate their effectiveness in distinguishing between two classes of sequences. In the case of Random Forest, it achieved an accuracy of 90.83%, with a confusion matrix showing 603 true negatives, 487 true positives, 18 false positives, and 92 false negatives. This results in a well-balanced F1-score of 0.90 for class 1 and 0.92 for class 0, indicating good precision and recall. The SVM model outperformed the Random Forest, achieving an accuracy of 92.42% with a confusion matrix showing 609 true negatives, 500 true positives, 12 false positives, and 79 false negatives. Both classes' precision, recall, and F1 scores are also notably high. Lastly, the Logistic Regression model attained an accuracy of 87.75%, with a confusion matrix reflecting 551 true negatives, 502 true positives, 70 false positives, and 77 false negatives. The F1 scores for class 0 and class 1 are relatively balanced at 0.88 and 0.87, indicating a good overall performance.

Table 2: Real data results

Real Data					
Methods	Accuracy (%)	Precision	Recall	F1-Score	Support
Random Forest	89.41	0.90	0.89	0.88	576
SVM	89.93	0.90	0.90	0.89	576
Logistic Regression	86.63	0.87	0.87	0.87	576

Table 3: Real and Synthetic data results

Real + Synthetic Data					
Methods	Accuracy (%)	Precision	Recall	F1-Score	Support
Random Forest	90.83	0.91	0.91	0.91	1200
SVM	92.42	0.93	0.92	0.92	1200
Logistic Regression	87.75	0.88	0.88	0.88	1200

These results illustrate that the synthetic data generated by the GAN model can be effectively classified by these machine learning algorithms, with SVM achieving the highest accuracy. In contrast, when the models were tested on actual data without synthetic data, their performance decreased, as evidenced by lower accuracies, imbalanced confusion matrices, and lower F1 scores. This underscores the importance of using synthetic data for training and testing when actual data may need to be more balanced.

Conclusions

This study introduced a generative adversarial network (GAN) framework to create synthetic DNA sequences for rat and human genomes. The GAN models were designed to capture authentic DNA characteristics, incorporating GC content distribution. Evaluation results indicated the successful generation of synthetic sequences closely resembling real DNA. The seamless integration of synthetic sequences into existing datasets expanded the genomic data diversity.

To refine and broaden this approach, future research can focus on optimizing GAN model architecture and hyperparameters for improved sequence quality and diversity. Exploring the impact of different target sequence lengths and GC content constraints would enhance model flexibility. Further assessments of synthetic sequences in downstream applications, such as training machine-earning models for DNA sequence classification, can validate their effectiveness across various classifiers and datasets. Additionally, incorporating other genomic aspects like structural features or epigenetic patterns into GAN models could advance comprehensive synthetic genomic data generation. This study lays a foundation for robust synthetic DNA sequences with potential applications in genomics and bioinformatics.

Acknowledgments

Dr. Yong Shi, Associate Professor, Department of Computer Science, Kennesaw State University

Contact Information

Nishat Tasnim

Dr. Yong Shi yshi5@kennesaw.edu

References

[1] Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., ... & Jay, F. (2021). Creating artificial human genomes using generative neural networks. PLoS genetics, 17(2), e1009303.

[2] Hazra, D., Kim, M. R., & Byun, Y. C. (2022). Generative adversarial networks for creating synthetic nucleic acid sequences of cat genome. International Journal of Molecular Sciences, 23(7), 3701.

[3] Killoran, N., Lee, L. J., Delong, A., Duvenaud, D., & Frey, B. J. (2017). Generating and designing DNA with deep generative models. arXiv preprint arXiv:1712.06148.

[4] Sarkar, S., Mridha, K., Ghosh, A., & Shaw, R. N. (2022). Machine Learning in Bioinformatics: New Technique for DNA Sequencing Classification. In Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022 (pp. 335-355). Singapore: Springer Nature Singapore.

[5] Ao, C., Jiao, S., Wang, Y., Yu, L., & Zou, Q. (2022). Biological sequence classification: A review on data and general methods. Research, 2022, 0011.

[6] D'amico, S., Dall'Olio, D., Sala, C., Dall'Olio, L., Sauta, E., Zampini, M., ... & Della Porta, M. G. (2023). Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. JCO Clinical Cancer Informatics, 7, e2300021.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author: Nishat Tasnim
Advisor: Dr. Yong Shi