IV. International Conference on Electrical, Computer and Energy Technologies (ICECET 2024)
25-27 July 2024, Sydney-Australia

Enhancing Genomic Datasets with cGANs: A Study on Synthetic DNA Sequences for Non-Human Species

Nishat Tasnim
Department of Computer Science
Kennesaw State University
Marietta, Georgia, USA
ntasnim@students.kennesaw.edu

Yong Shi
Department of Computer Science
Kennesaw State University
Marietta, Georgia, USA
yshi5@kennesaw.edu

Kun Suo
Department of Computer Science
Kennesaw State University
Marietta, Georgia, USA
ksuo@kennesaw.edu

Xinyue Zhang
Department of Computer Science
Kennesaw State University
Marietta, Georgia, USA
xzhang48@kennesaw.edu

***Abstract*—DNA sequences, unique to each species, are pivotal in life sciences. Despite their importance, genomic data analysis is hindered by limited availability, uneven distribution, and privacy issues, affecting the efficacy of machine learning in DNA sequence classification. This paper presents a new framework utilizing Conditional Generative Adversarial Networks (cGANs) to synthesize DNA sequences and assess their resemblance to real sequences via supervised learning. This method enriches genomic datasets by modifying existing samples, reducing the need for new acquisitions. The research focuses on the black rat (Rattus rattus) and the chimpanzee (Pan troglodytes). By feeding real, synthetic, and hybrid DNA sequences into classification models, including Support Vector Machine (SVM), Random Forest (RF), and Bagging Decision Trees, the framework illustrates that cGAN-generated synthetic DNA sequences can closely replicate the properties of real sequences, delivering comparable classification outcomes.**

***Index Terms*—generative adversarial networks, deep learning, machine learning, synthetic DNA sequence**

I. Introduction

The field of generative artificial intelligence (GenAI) [1] has profoundly impacted various scientific and technological domains, particularly with the development of Generative Adversarial Networks (GANs) [2], which represent a significant milestone in the last decade. In the last decade, extensive exploration and application of these generative models in various machine-learning contexts have demonstrated notable advancements in genetics, particularly in synthesizing DNA sequences using deep generative models.

Our research utilized deep generative adversarial networks to effectively capture complex patterns in real genomic datasets. This method enabled the creation of high-quality synthetic genomes while mitigating privacy concerns. The study results indicate that these synthetic genomes maintain essential original data features, such as guanine-cytosine (GC) content. We evaluated our model using metrics like accuracy, confusion matrix, and ROC curve, comparing the performance of classifiers on real, synthetic, and hybrid datasets. Applying generative models and synthetic genomes in genetic research offers a means to encapsulate real genomes efficiently.

II. Background and Motivation

DNA is a molecule that carries the biological code that defines each species [3]. DNA has four types of nucleic acid bases: (G) guanine, (A) adenine, (C) cytosine, and (T) thymine. These bases pair as nucleotides GC, AA, TG, and CA in DNA.

Sorting DNA sequences into different groups is very important for computational biology. Using machine learning models to sort DNA sequences can make the process quicker and more efficient than traditional manual methods. However, these models require a substantial amount of data to learn effectively.

Advancements in sequencing technologies and reduced costs have led to a rise in the availability of genetic data. GenBank, a public encyclopedic database, holds 19.6 trillion base pairs derived from more than 2.9 billion nucleotide sequences, representing 504,000 scientifically identified species [4]. The application of these genetic data spans multiple fields, including medicine and evolution. Despite its widespread utility, the cost remains a limiting factor, resulting in an ongoing demand for additional data. Moreover, accessing most data held by governmental and private entities proves challenging due to privacy concerns, impeding scientific pursuits. In this context, generative models, a machine learning method, have potential. Generating synthetic gene expression data presents a promising way to address these challenges.

One use of generative tools in genomics is automatically creating probe sequences that can test the binding of protein-DNA or RNA. Another benefit is to improve genomic sequences to satisfy multiple, possibly conflicting properties, such as producing a specific protein product from the cell while keeping other properties, like GC content, constant.

III. Related work

Yelmen et al. [5] demonstrate that deep generative models like GANs and RBMs can create high-quality synthetic genomes (AGs) that reflect real genomic datasets' population structure and genetic features. However, these models face

979-8-3503-9591-4/24/$31.00 ©2024 IEEE

limitations, including computational constraints that prevent generating whole genomes, GANs underrepresenting rare alleles, RBMs overfitting, and unresolved privacy concerns regarding the synthetic data.

In another relevant study, Hazra et al. [6] showed that GANs can generate synthetic nucleic acid sequences for the cat genome, with a high mean correlation coefficient of 93.7% to the original data. The model's applicability to a broader range of genomes remains to be determined, and scaling to other types of genomic data presents challenges.

Chen et al. [7] illustrate that the Population-scale Genomic Data Augmentation model (PG-cGAN) successfully generates synthetic HLA genotypes reflecting real genomic data's population structure, variant frequency, and linkage patterns. The model struggles with high-quality, high-dimensional genomic data generation and needs robust evaluation metrics for assessing synthetic data fidelity.

Aswath et al. [8] highlight enhanced DNA sequence classification by combining nature-inspired algorithms with conventional supervised classifiers, resulting in better accuracy and efficiency. It faces challenges in scaling to larger genomic datasets and requires adjustments to algorithm parameters for consistent performance across various data types.

According to a study by Wang et al. [9], the AI-driven generative model effectively designed synthetic promoters with 70.8% functionality in E. coli, showcasing the potential of deep learning in synthesizing biologically active sequences. The model exhibited only moderate success in accurately predicting expression levels from these sequences, underscoring the difficulties in forecasting complex biological behaviors.

In a recent study [10], researchers developed the latent diffusion model DiscDiff for DNA sequence generation, effectively achieving high-quality synthetic sequences that closely mirror natural DNA across multiple metrics. Despite these successes, the study highlights significant challenges in managing high-dimensional genomic data and fully replicating the diversity of natural DNA sequences.

Our research ventures beyond the current state of research in several key aspects:

Innovative Integration of GANs with Supervised Learning: Our research integrates Generative Adversarial Networks with supervised learning to simultaneously generate synthetic DNA sequences and evaluate the generated data through classification models.

Expanding Genomic Research Beyond Human Species: Focusing on non-human species, our study breaks from the usual human-centric genomic research, offering new insights into animal genomics.

Contribution to Data Diversity and Model Robustness: By integrating synthetic DNA sequences into our datasets, we enhance data diversity and strengthen the robustness and accuracy of our computational models, addressing data scarcity and improving the efficacy of genomic analysis.

IV. Method

Our methodology involves data collection, preprocessing for GAN training, and subsequent training of the GAN. The experimental setup for GAN model training is detailed in Table I.

TABLE I.
Training Setup for GAN Model

Parameters	Values
No. of Training Iterations	200
Size of Training Batch	64
Loss Function	Binary Crossentropy
Optimizer	Adam
Learning Rate	0.0002
No. of Black Rat Samples in Real Data	2173
No. of Chimpanzee Samples in Real Data	1472
Targeted No. of Generated Data per Species	1000

The trained GAN model generated 1000 fake sequences for each class. To evaluate the GAN's performance, we conducted a classification task in three steps: (1) using the real dataset, (2) using the GAN-generated dataset, and (3) using a hybrid dataset combining real and GAN-generated data. We then compared the performance of the three classification models to observe differences between classifications based on predicted and real data.

A. Model Architecture

Conditional Generative Adversarial Networks: Generative Adversarial Networks (GANs) comprise two main components: a generator and a discriminator, as shown in Fig. 1.

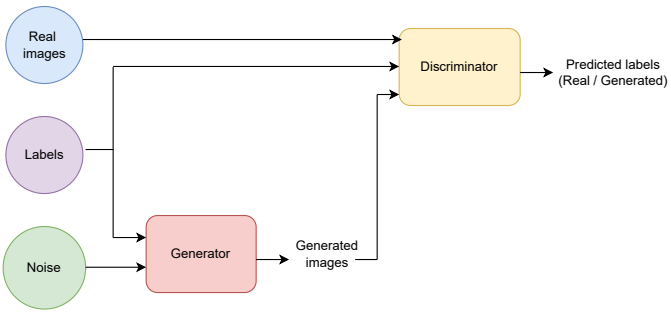

Fig. 1. Conditional GAN architecture.

These components are typically optimized using alternating gradient descent. Like a uniform distribution, the generator aims to sample z from a distribution p_z to approximate the target distribution p_d. The discriminator (D) distinguishes between the generator's output and real data. In a Conditional GAN (cGAN) [11], the generator (G) uses label information and functions within an encoder-decoder framework. The encoder maps the labels to a reduced-dimensional feature space, while the decoder reverses this process by converting the low-dimensional representation back into a high-dimensional space. Let s represent the conditioning label and y, which is

a sample from the target distribution. The adversarial loss can then be formulated as follows:

$$L_{\text{adv}}(G,D) = \mathbb{E}_{s,y\sim p_{\text{data}}(s,y)}[\log D(y|s)] + \mathbb{E}_{s\sim p_{\text{data}}(s),z\sim p_z(z)}[\log(1-D(G(s,z)|s))] \quad (1)$$

This is achieved by addressing a min-max optimization problem.

$$\min_{w_G}\max_{w_D} L_{\text{adv}}(G,D) = \min_{w_G}\max_{w_D} \mathbb{E}_{s,y\sim p_{\text{data}}(s,y)}[\log D(y|s,w_D)] + \mathbb{E}_{s\sim p_{\text{data}}(s),z\sim p_z(z)}[\log(1-D(G(s,z|w_G)|s,w_D))] \quad (2)$$

Here, w_D and w_G represent the parameters of the discriminator and generator, respectively. For simplicity, parameter dependencies and noise variables z are excluded.

Isola et al. [12] and Salimans et al. [13] suggest that incorporating auxiliary loss terms, such as content loss and feature matching, enhances the effectiveness of cGANs compared to standard GAN models. Loss of matched features [13] is defined as:

$$L_f = \sum_{n=1}^{N} \left\| \pi(G(s^{(n)})) - \pi(y^{(n)}) \right\| \quad (3)$$

Here $\pi()$ represents the feature extraction from the discriminator's penultimate layer.

The combined loss function incorporating these improvements for cGAN is outlined below:

$$L = L_{\text{adv}} + \lambda_c \sum_{n=1}^{N} \left\| G(s^{(n)}) - y^{(n)} \right\| + \lambda_\pi L_f \quad (4)$$

In this equation, λ_π and λ_c are adjustable hyperparameters that determine the weight of each term in the total loss function [17].

B. Classifiers

Classification of biological sequences, a prevalent data mining challenge, is complex due to the non-numeric nature of sequence elements, interdependencies among these elements, and varying sequence lengths across samples. This study employs machine learning models to decipher hidden patterns within sequences that correlate with predefined classes. Specifically, we implement three distinct classification algorithms—Support Vector Machine, Random Forest, and Bagging Decision Trees—to effectively classify DNA sequences.

C. Technological Setup

Our experiments utilized PyTorch and TensorFlow version 2.15.0 post1 frameworks, running on a Linux server with multiple Tesla V100-SXM2 32GB GPUs.

D. Dataset

The initial phase of our analysis involved extracting DNA sequences from FASTA files utilizing the BioPython library. The National Centre for Biotechnology Information (NCBI) acquired genomic data for black rats and chimpanzees [14], [15]. They were stored as FASTA files containing categorical representations of ATCG nucleotide characters—each sequence in the dataset comprised over 272 million characters, constituting a substantial and complex dataset. For analysis, these sequences were stored in a Pandas data frame. The dataset demonstrated notable imbalances, which were mitigated through targeted preprocessing techniques.

E. Data Preprocessing for GAN model

A regular expression is utilized to remove non-DNA characters, ensuring the purity of the DNA sequences, while all characters are converted to uppercase to maintain uniformity across the dataset.

DNA sequences are standardized to a consistent length of 1000 base pairs, with sequences shorter than this length padded with 'N' characters and those exceeding it truncated, ensuring dataset consistency and facilitating further analysis.

DNA sequences in text format are preprocessed into a numerical format for compatibility with machine learning algorithms. One-hot encoding converts each DNA base into numerical values, as Fig. 2 shows.

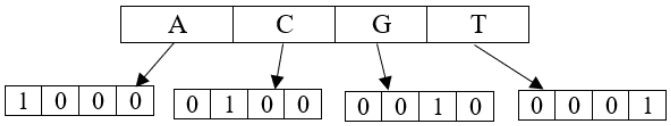

Fig. 2. Sequence data encoding using label binarizer.

An 80% to 20% division is employed for partitioning the data into training and testing sets, adhering to the widely accepted global split ratio standard.

Fig. 3. A single DNA sequence from the real black rat dataset.

Fig. 4. A single DNA sequence from the generated black rat dataset.

F. Data Preprocessing for ML Classification

The datasets for the rat and chimpanzee species are merged, with class labels designated as 0 for rats and 1 for chimpanzees. Each class (real, GAN-generated, hybrid) consists of 2000 samples for Black Rats and Chimpanzees, making a balanced dataset. To mitigate potential bias due to the order of data entry, the dataset is randomized, ensuring that the classification model does not inadvertently learn patterns influenced by the sequence of the data.

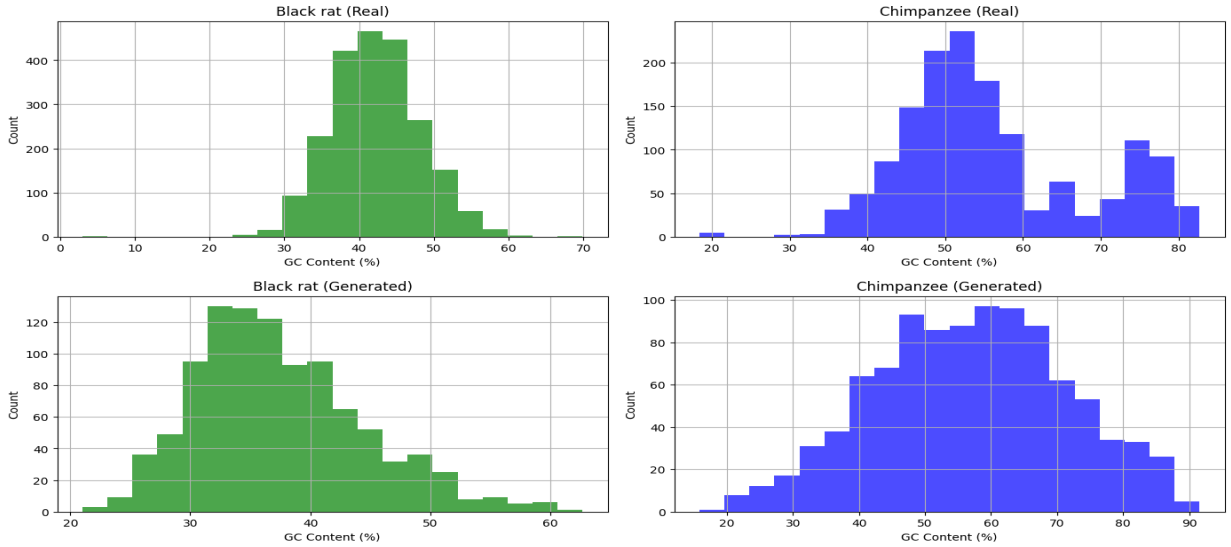

Fig. 5. Comparative analysis of GC content distribution.

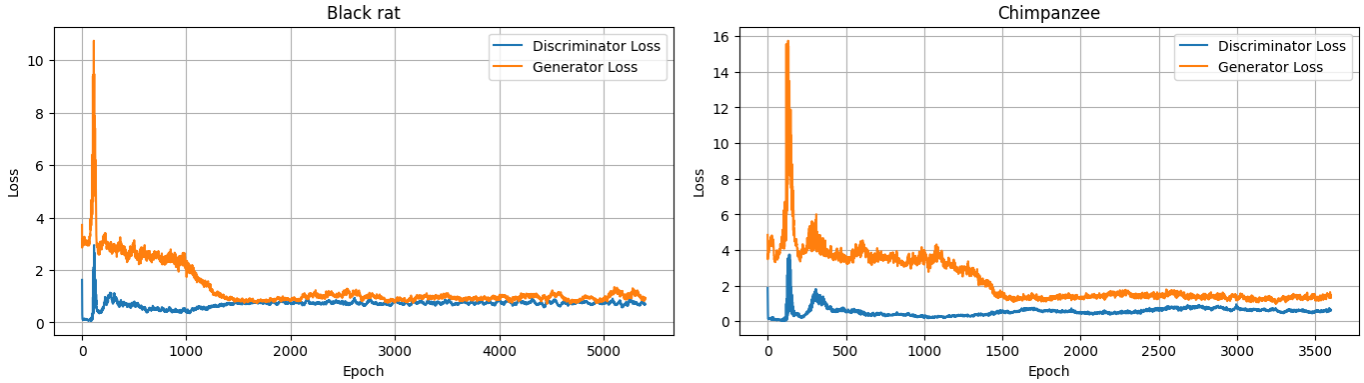

Fig. 6. Graphical representation of the GAN losses over the training period.

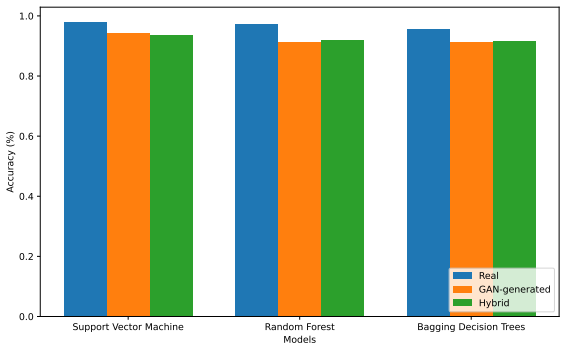

Fig. 7. Detailed analysis of the classifier performance across different datasets.

K-mer features are derived from DNA sequences utilizing the CountVectorizer technique, which converts these sequences into a numerical format amenable to machine learning algorithms. Subsequently, the extracted features are normalized via the StandardScaler method to guarantee that each feature uniformly influences the model training process, thereby mitigating biases associated with features of varying scales.

V. Experimental Results and Evaluation

Our generator accurately identifies base pair relationships, showing that adenine and guanine, both purine bases with double-ring structures, do not pair together. In DNA, bases form hydrogen bonds only with their complements: adenine pairs with thymine, and guanine pairs with cytosine [16]. Fig. 3 and Fig. 4 compare the generated sequences to the real ones, highlighting their similarity.

Our study calculates the GC content for each DNA sequence by quantifying the percentage of Guanine (G) and Cytosine (C) bases. Fig. 5 presents the GC content distribution for real and generated data from the two species.

Table II reveals that the discriminator differentiates between real and synthetic data for chimpanzees more effectively than for black rats. Although the generator performs better with black rat sequences, the model fits the chimpanzee data

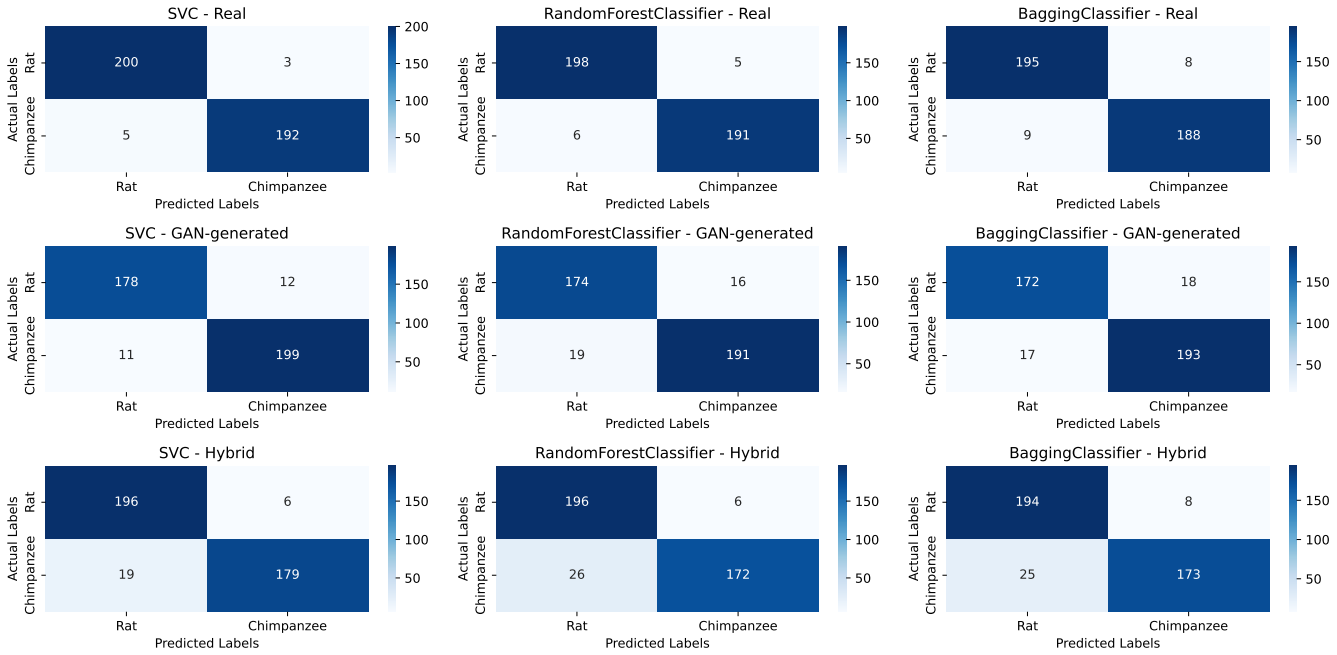

Fig. 8. Confusion matrices of classification models used for evaluation.

slightly better, as the average test loss indicates. This suggests a balanced adversarial dynamic where the generator, despite some challenges, produces more accurate representations of real data for chimpanzees. This improved fit may result from the model's enhanced error correction during training, adjusting more effectively to the complexities of the chimpanzee data.

TABLE II.
GAN TRAINING RESULTS

Species	Disc. Loss	Gen. Loss	Average Test Loss
Black Rat	0.7011	1.3344	0.7289
Chimpanzee	0.5480	2.3279	0.6863

Fig. 6 displays a declining trend in loss values, suggesting enhancements in the generator's ability to produce realistic synthetic DNA sequences and improvements in the discriminator's capability to differentiate between real and synthetic data.

Despite decreasing scores compared to the real data, the Generated and Hybrid results remain robust indicators of effective model performance, as illustrated in Fig. 7. For the Support Vector Machine (SVM), the scores of 0.9425 in the Generated and 0.9375 in the Hybrid categories showcase high predictive accuracy, aligning closely with the near-perfect score of 0.98 in the Real data. Similarly, the Random Forest (RF) model scores of 0.9125 and 0.92 in Generated and Hybrid are commendable, considering the Real score of 0.9725. The Bagging Decision Trees model also performs well, with scores of 0.9125 in Generated and 0.9175 in Hybrid, which are slightly lower but close to the Real score of 0.9575. These results highlight that Generated and Hybrid models maintain substantial efficacy and are only marginally behind the Real model's outcomes, suggesting effective learning and generalization capabilities in varied scenarios.

Fig. 8 illustrates that in real data scenarios, SVC and Random Forest perform excellently with few errors, while the Bagging Classifier is slightly less accurate with rat classifications. In GAN-generated data, SVC faces difficulties, particularly with rats, while Random Forest remains fairly accurate despite some errors, and Bagging Classifier excels, especially with chimpanzees. In hybrid data, the SVC and the Bagging Classifier achieve high accuracy with minimal errors, but the Random Forest struggles with rat classifications. Overall, the Bagging Classifier consistently performs consistently across all data types, demonstrating its versatility, while SVC is optimal in less noisy settings. Random Forest is reliable but more prone to hybrid and synthetic scenario errors.

Regarding the ROC curve, the Real dataset performs best across all classifiers. The GAN-generated dataset performs slightly lower but still well. The Hybrid dataset shows the lowest performance yet remains strong. Both GAN-generated and Hybrid datasets demonstrate robust classification capabilities and potential for generating high-quality data, as shown in Fig. 9.

VI. LIMITATIONS

While applicable in many scenarios, synthetic data may only partially capture the complex variability and diversity

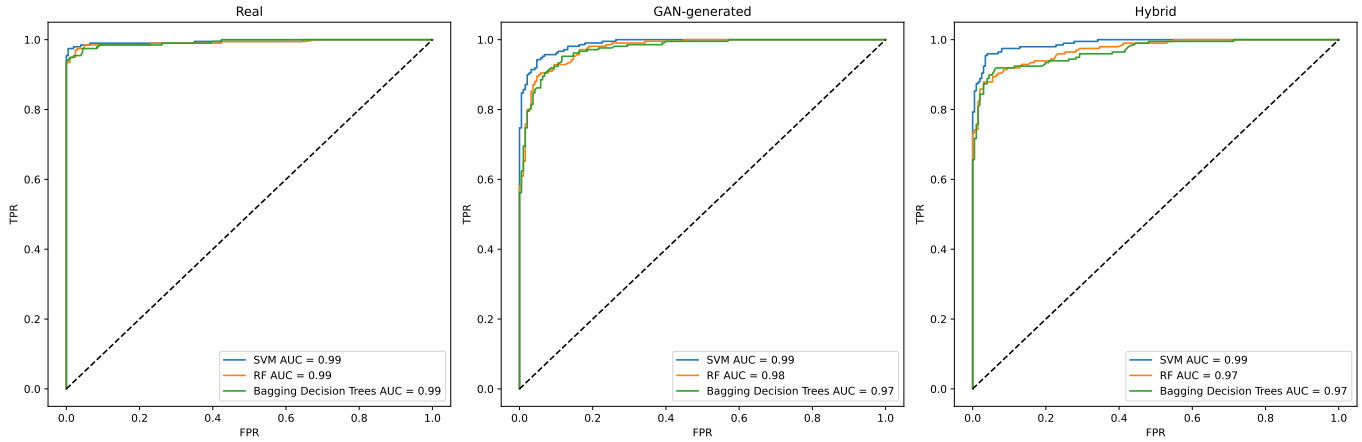

Fig. 9. Receiver Operating Characteristic (ROC) analysis for classification models.

inherent in genomic sequences, which can limit model accuracy. Although these data can mimic basic genomic properties such as GC content, they often need more of the deeper biological complexities found in natural sequences, which restricts their effectiveness in improving classification models. Furthermore, integrating synthetic data into training datasets may not introduce new or unique patterns, potentially leading to a plateau in learning improvements, as models might not gain additional predictive power from this data.

VII. Conclusion and Future Work

In conclusion, our research on using GANs and supervised learning for DNA sequence analysis shows notable success in generating synthetic sequences for black rats and chimpanzees. These synthetic sequences, nearly indistinguishable from real data, have significantly enhanced the effectiveness of classification models. Our work demonstrates the substantial potential of GANs in overcoming the challenges of data availability and privacy in genomic studies.

Future endeavors should focus on refining the GAN architecture and hyperparameters for better quality and varied synthetic DNA sequences. Expanding the research to cover more species and incorporating other genomic elements like structural features will broaden the scope and application of this technology, paving the way for more profound impacts in genomics and bioinformatics.

Acknowledgment

We thank the anonymous reviewers for their valuable feedback on this manuscript. Partial funding for this research was provided by the U.S. National Science Foundation under grants CPS-2103459 and SHF-2210744.

References

[1] J. KAPLAN, Generative Artificial Intelligence. Oxford University Press, 2024.

[2] M. Ghayoumi, Generative Adversarial Networks in Practice. CRC Press, 2023.

[3] R. J. Roberts, “Nucleic acid — chemical compound,” Encyclopædia Britannica. 2020. Available: https://www.britannica.com/science/nucleic-acid

[4] E. W. Sayers et al., “GenBank 2023 update,” Nucleic Acids Research, Nov. 2022, doi: https://doi.org/10.1093/nar/gkac1012.

[5] B. Yelmen et al., “Creating artificial human genomes using generative neural networks,” PLOS Genetics, vol. 17, no. 2, p. e1009303, Feb. 2021, doi: https://doi.org/10.1371/journal.pgen.1009303.

[6] D. Hazra, M.-R. Kim, and Y.-C. Byun, “Generative Adversarial Networks for Creating Synthetic Nucleic Acid Sequences of Cat Genome,” International Journal of Molecular Sciences, vol. 23, no. 7, p. 3701, Mar. 2022, doi: https://doi.org/10.3390/ijms23073701.

[7] J. Chen, Mohammad Erfan Mowlaei, and X. Shi, “Population-scale Genomic Data Augmentation Based on Conditional Generative Adversarial Networks,” Sep. 2020, doi: https://doi.org/10.1145/3388440.3412475.

[8] S. Aswath, CH.Mohan Sai Kumar, V.Hima Deepthi, S.Imran Javeed, and SVN. Rupesh, “DNA Sequence Classification with Improved Performance of Supervised Classifiers using Nature Inspired Algorithms,” 2022 2nd International Conference on Intelligent Technologies (CONIT), Jun. 2022, doi: https://doi.org/10.1109/conit55038.2022.9848025.

[9] Y. Wang, H. Wang, L. Wei, S. Li, L. Liu, and X. Wang, “Synthetic promoter design in Escherichia coli based on a deep generative network,” Nucleic Acids Research, May 2020, doi: https://doi.org/10.1093/nar/gkaa325.

[10] Z. Li et al., “Latent Diffusion Model for DNA Sequence Generation,” arXiv (Cornell University), Oct. 2023, doi: https://doi.org/10.48550/arxiv.2310.06150.

[11] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” arXiv.org, Nov. 06, 2014. https://arxiv.org/abs/1411.1784v1

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” Nov. 2016, doi: https://doi.org/10.48550/arxiv.1611.07004.

[13] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,” arXiv.org, Jun. 10, 2016. https://arxiv.org/abs/1606.03498v1

[14] “Rattus rattus,” NCBI. https://tinyurl.com/4t8bdky8 (accessed Jun. 18, 2024).

[15] “Pan troglodytes,” NCBI. https://tinyurl.com/3ywwxm96 (accessed Jun. 18, 2024).

[16] “DNA Base Pairings: Why Adenine & Guanine Don’t Mix,” Physics Forums: Science Discussion, Homework Help, Articles, Apr. 25, 2005. https://tinyurl.com/5n9355xr (accessed Jun. 18, 2024).

[17] G. G. Chrysos, J. Kossaifi, and Stefanos Zafeiriou, “Robust Conditional Generative Adversarial Networks,” arXiv (Cornell University), Sep. 2018.