

Unveiling DNA sequences: a comparison of machine learning and deep learning techniques for prediction

S. M. Shifana Rayesha¹, Aisha Banu¹, Afzalur Rahman², Sharon Priya¹

¹Department of Computer Science and Engineering, School of Computer Information and Mathematical Sciences,

B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

²School of Commerce and Economics, Presidency University, Bengaluru, India

Article Info

Article history:

Received Nov 18, 2023

Revised Mar 29, 2024

Accepted Apr 17, 2024

Keywords:

Artificial neural network

Decision tree

Fasta

K-nearest neighbors-classification

Random forest

ABSTRACT

DNA is the biological macromolecule unit that carries the information of all protein, amino acid sequences. With the help of this protein sequence, we explore the mutated gene and disease-causing mutated genomic pattern. Currently, the progression of genomic innovation is the source of DNA arrangement information developing at a dangerous rate—external factors have stimulated the volume of research into DNA genomes. Initially, the development process of DNA sequencing is accomplished with the support of the Database, data structures, and sequence similarity. The method is capable of extracting a particular property in DNA. We employ the deep learning algorithm to pull out protein sequences' features. The DNA sequence is classified based on the in-build protein structures extracted into the Fasta file. Therefore, the DNA sequence of E. Coli with 106 data sets and 57 nucleotides is tested experimentally. Finally, we compared the results with the existing decision tree algorithm, k-nearest neighbors (KNN)-classification, random forest, and neural networks. The deep learning algorithm yields higher efficiency of 98% compared to other machine learning algorithms. This highlights the potential of deep learning in genomics research and its ability to yield superior results in classifying DNA sequences.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

S. M. Shifana Rayesha

Department of Computer Science and Engineering, School of Computer Information and Mathematical Sciences

B. S. Abdur Rahman Crescent Institute of Science and Technology

Chennai, India

Email: shifana@crescent.education

1. INTRODUCTION

DNA sequencing is an emerging field in the recent era. With the support of protein information, we inspect the characteristics of species and their genomic sequence. The rapid development of data expansion in maintaining biological and bionics information as well as immense genomic details. We must extract useful information from this vast amount of data, and concurrently, bioinformatics was developed [1]. Bio information comprises biology, mathematics, computer science, and life sciences to explore the biological data, which guides the biological researcher to inspect the bioinformatics data. In particular, the initial procedure is to get data on the metagenomic coding derived from the investigation of genomic DNA grouping [2]. The next step is to simulate the partial features of the protein [3]. Eventually, based on the protein's properties, the investigator develops or discovers the medication.

Metrics show that biological data doubles roughly every two to three years. In 1982, GenBank's first nucleic arrangement data set had just 606 groupings, containing 680,000 nucleotide bases [4]. As of February 2013, it contains 162 million biological arrangement information, with 150 billion nucleotide bases. The most

effective method is to mine data from this massive amount of information and extract the exact information. For complex biological information, it needs to manage the enormous amount of data, we need to retrieve the data, and it should keep the meaning of the data the same [1]. The deep learning technique is employed to fine-tune the enormous amount of information in the data set present in the biological information and ensure the correctness of data. Artificial intelligence (AI) is an emerging field in bio information [5].

DNA is the structural unit of biomolecular information that makes the species unique. It transmits genomic information and facilitates the enhancement of biological processes and the operation of life capacities [6]. DNA sequencing and omics sequencing are powerful techniques used in molecular biology and genetics to analyze and understand the genetic information encoded within an organism's DNA. Omics sequencing refers to the comprehensive analysis of biological molecules (e.g., DNA, RNA, proteins, and metabolites) within a biological sample. Omics technologies allow the systematic study of different molecular components and their interactions within biological systems. Examples of these technologies are metabolomics, transcriptomics, proteomics, and genomics. In the existing methodology, the number of features in an omic dataset considerably exceeds the number of samples, we refer to that dataset as high-dimensional. Biological datasets are notorious for their high dimension and "small n—large p" structure [7]. The omic genomic data of the existing system create data redundancy with the segregation of genomic class and distribution. When the same genomic data appears in multiple places, it is called data redundancy. The same genomic information in two or more tables will lead to data inconsistency. Data redundancy can lead to data inconsistency which causes genomic data to produce useless information and misclassifying data. The Previous prediction technique of antibiotic resistance classes is not flexible, making it difficult to update and remove existing changes. Therefore, the traditional cluster database at high identity with tolerance (CD-HIT) method to predict antibiotic resistance and genomic sequence produces inaccurate results [8]. Also, previous machine learning model yields minimal efficiency because it needs to process a tremendous amount of data.

In recent trends, AI has played a significant role in data analysis and prediction of the DNA chain. It also improves the data processing capacities and produces important bionomics information. In this paper, we focus on DNA sequencing using data mining in deep learning [9]. In this study, we explain that similar arrangement in the premise of sequence in DNA information mining. We have extensively broken down the fundamental interaction of information mining and summed up the calculations usually utilized in AI [10]. At that point, we summed up four common uses of AI in DNA arrangement information: DNA grouping arrangement, order, bunching, and design mining.

2. PROPOSED METHODOLOGY

Early infectious disease identification, distinguishing between infectious and noninfectious pathologies, and successfully treating ensuing symptoms are all critical components in the global fight against antibiotic resistance. AI has the potential to play a pivotal role in combating this problem. One of the key AI techniques that can be employed is the creation of antibiograms and the subsequent development of personalized deep learning-based models for predicting antimicrobial resistance (AMR). These techniques can be particularly valuable in dealing with high-risk infectious pathogens and understanding their susceptibility patterns. The development of a deep learning-based AMR prediction model is central to this approach, and it provides insights into high-potential pathogens with reduced side effects and their corresponding AMR patterns. The Figure 1 illustrates the utilization of deep sequencing AI models as part of this process, highlighting the integration of AI in each phase to address antibiotic resistance effectively.

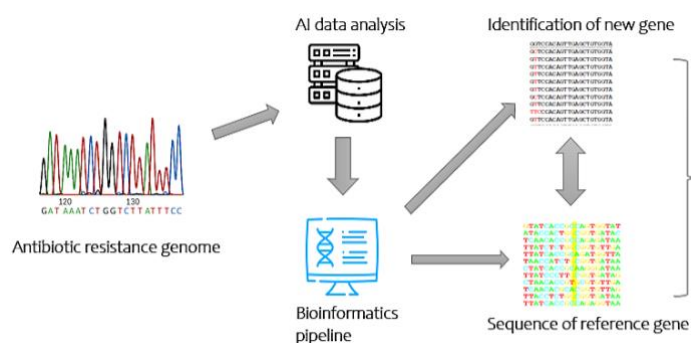


Figure 1. Deep learning antibiotic-resistance gene sequencing model

Existing methods struggle to analyze the vast and diverse data (multi-omics) from various sources like DNA. Deep learning offers a solution by combining this information into a unified space, revealing hidden connections between different data types. This comprehensive view allows researchers to gain a deeper understanding of biological processes. The proposed approach focuses on predicting antibiotics using whole genome sequences from various bacteria. By analyzing this high-density genomic data, the algorithm searches for potential antibiotic-producing genes, paving the way for drug discovery. This method builds upon existing machine and deep learning algorithms to refine the prediction of antibiotic resistance classes.

3. LITERATURE REVIEW

Phumichai *et al.* [6], this journal covers fundamental research on genetic mechanisms in plants and applied research on using genetic knowledge for crop improvement and agricultural innovation. Regarding the particular article by Phumichai *et al.* [6], which delves research on genome-wide association mapping, genomic prediction of features linked to yield, and starch pasting characteristics in cassava, the journal's focus complements the study's goals of enhancing genetic comprehension and utilizing genomic tools for crop enhancement in agriculture [6]. According to López *et al.* [11], genomic selection is a technique employed in plant and animal breeding to forecast the genetic quality of individuals using their DNA markers. Deep learning is a type of machine learning that has shown promise across several industries because of its ability to recognize complicated structures and properties on its own in extensive datasets. The purpose of this work is to explore how deep learning techniques can be applied to enhance genomic selection by increasing prediction models' efficiency and accuracy [11]. Zhang *et al.* [12] discuss the importance of explainability in machine learning models, especially in applications where decisions based on model predictions have significant consequences, such as medical diagnosis or autonomous driving. They propose leveraging uncertainty quantification techniques to not only improve the interpretability of model predictions but also to provide insights into the reliability and robustness of the models [12].

Xu *et al.* [13] primary focus is to address the challenge of accurately detecting and segmenting cracks in concrete structures, which is essential for structural health monitoring and maintenance. Conventional techniques for crack identification frequently rely on human error-prone and time-consuming manual inspection. By contrast, the suggested method makes use of deep learning techniques to increase accuracy and automate the procedure [13]. Zhang *et al.* [14] aims to review deep learning approaches being used in the field of omics, which includes genomics, transcriptomics, proteomics, metabolomics, and other high-throughput biological data domains. Omics data is characterized by its large-scale, high-dimensional nature, and deep learning methods have shown promise in extracting meaningful patterns and features from such data for various biomedical applications [14].

Fu *et al.* [8] addresses the computational challenges associated with clustering large volumes of sequence data generated by next-generation sequencing (NGS) technologies. Clustering is an essential step in analyzing NGS data as it helps identify and remove redundancy, reduce computational complexity, and facilitate downstream analysis tasks such as sequence alignment and assembly. This paper concludes by highlighting the practical implications of their work, emphasizing the importance of efficient sequence clustering algorithms in handling the enormous amounts of NGS data produced in contemporary genomic research. The faster CD-HIT algorithm described in this publication is a useful resource for bioinformatics researchers, enabling them to perform fast and scalable clustering of NGS data to support various genomic analysis tasks [8]. Ren *et al.* [15] provide evidence of the way their method works to forecast antibiotic resistance through experimental evaluations using diverse datasets of bacterial genomes with known AMR profiles. They use parameters like sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) to evaluate the effectiveness of their predictive models (AUC), comparing them against baseline methods and existing AMR prediction tools [15].

3.1. Research gap

Several research gaps emerge from the provided references. Omics data analysis requires further research in specific applications and mitigating data limitations. NGS data analysis needs efficient algorithms beyond clustering and scalability for future technologies. AMR prediction can be expanded to other pathogens and incorporate clinical data for a broader understanding. These identified gaps offer opportunities for novel research and advancement in machine learning and deep learning across various fields. This study aims to address the identified research gap by investigating the impact of these challenges on the prediction of antibiotic resistance classes. Specifically, it seeks to develop more flexible and efficient prediction techniques that can effectively handle high-dimensional genomic data, mitigate data redundancy and inconsistency issues, and improve the accuracy of antibiotic resistance predictions.

4. DATA SETS EVALUATION

The evaluation of the dataset is carried out using two distinct methodologies. The first methodology involves ensuring an even distribution of omic sequences. In the second methodology, the focus is on achieving a dissimilar distribution of metagenomic sequences. In cases where a sequence lacks even distribution, the approach involves appending zeros to the end of each sequence until uniform length is achieved. To distribute the sequence evenly we added zero for each line or row of sequence length in metagenomic sequence. The DNA dataset consists of common metagenomic characteristics of DNA preprocessed into numeric as 1, 2, and 3 [11]. The Fasta file format of sequence class types like E. Coli belongs to the gram-positive or gram-negative class. Figure 2 represents the metagenomic class distribution in the DNA sequence of E. Coli. Classes name and metagenomic classes are distributed unevenly in the nucleotide sequences for 57 metagenomic classes. Therefore, finally the DNA sequence is converted to even distribution of metagenomic sequence of class identification follows this experimental method to predict the sequence similarity in the classes.

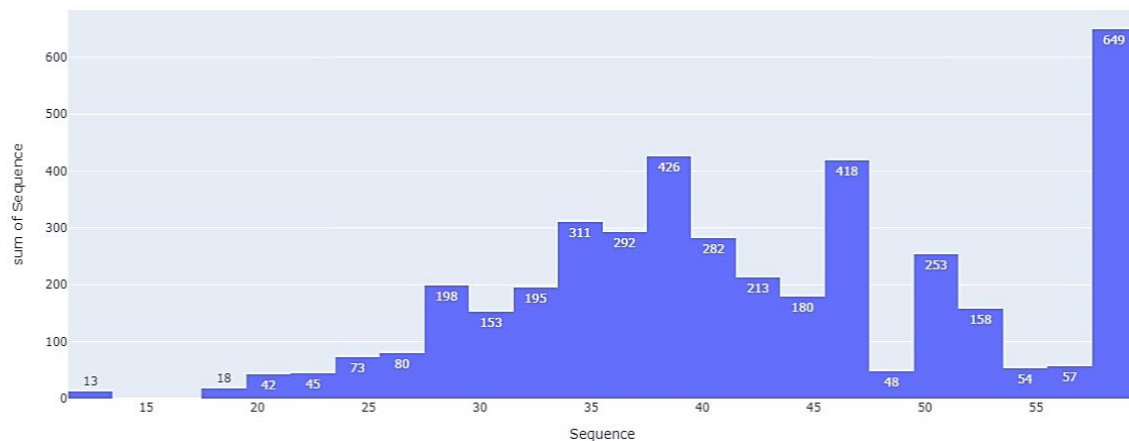


Figure 2. Uneven distribution of metagenomic distribution of DNA omic sequence

The composition of a, c, g, t in DNA is depicted diagrammatically. Figure 3 depicts the composition of adenine(a), cytosine(c), guanine(g) and thymine(t) in metagenomic sequence of Aroh gene and Male G gene of E. Coli. Figure 4 illustrates the metagenomic sequence of E. Coli, with each of the 57 different classes identified by numerical labels, from 1 to 57. The figure showcases the overall composition of adenine (A), cytosine (C), guanine (G), and thymine (T) within these sequences. In Figure 2, represents the specific metagenomic or nucleotide sequences of E. Coli corresponding to these 57 different classes. These sequences are uniquely identified by names such as AMPC, ARAC, AROH, and BIOB.

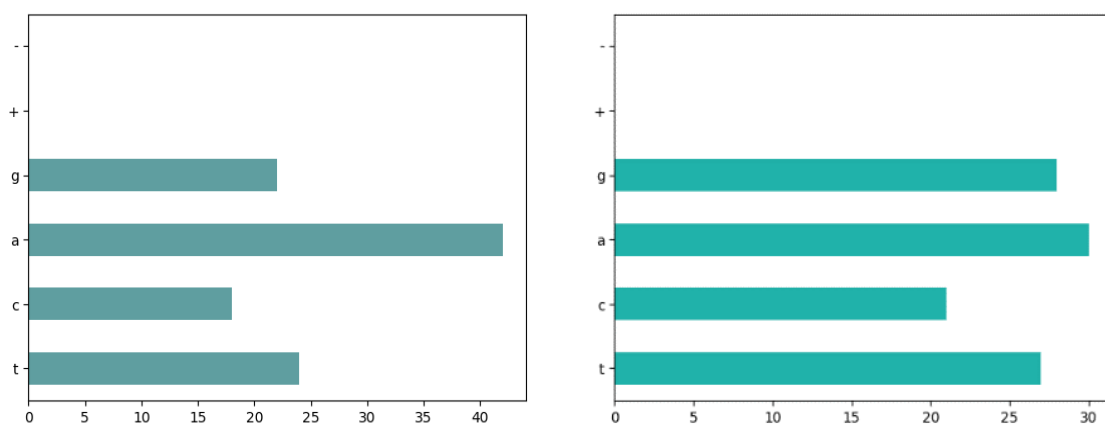


Figure 3. Sample composition of a, c, g, t AROH gene and MALE G gene in E. Coli

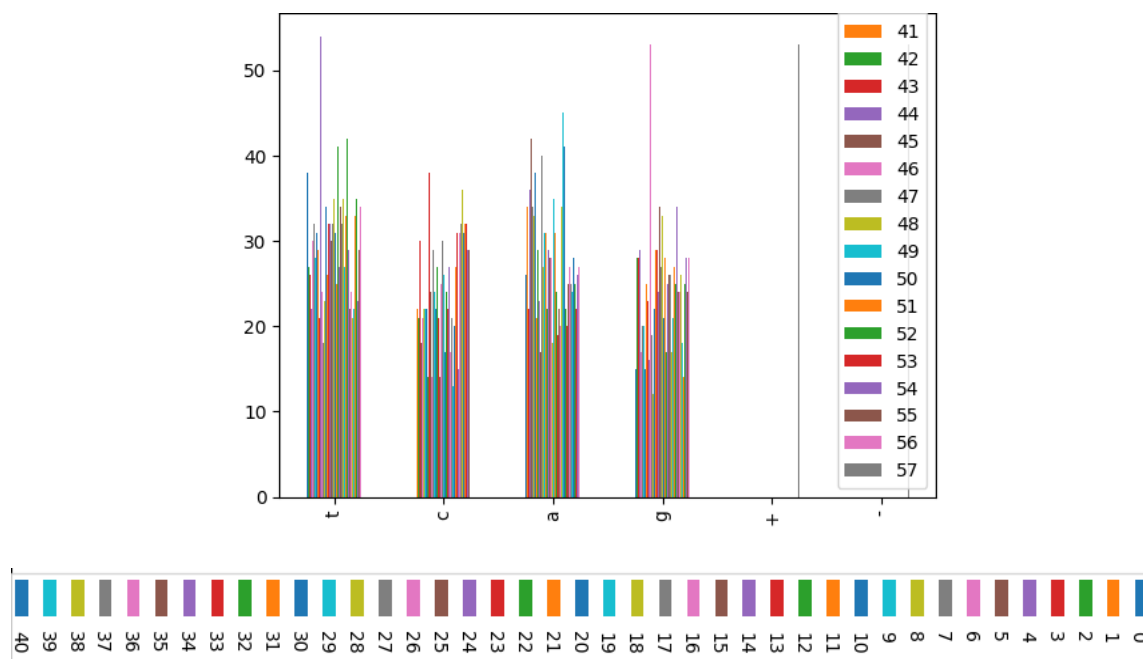


Figure 4. Composition of a, c, g, t in E. Coli of 57 different classes

5. METHOD

5.1. Data preprocessing

The string value is converted into numerical values while refining the DNA sequence. These numerical values are present in the train and test set. The encoding is possible in three methods of sequence encoding [6], i) one-hot encoding, ii) sequential encoding, and iii) k-mers encoding. One hot encoding adds dummy values to the sequence value a, c, g, t as zeros and ones [16]. Sequential encoding assigns random values a, c, g, t. K-mers compose and decompose the nonoverlapping sets of DNA sequences [15]. The strings are off with the values numeric in the test and the training dataset [17]. Then the test and training data set is evaluated with following machine learning and deep learning algorithms.

6. PRE-PROCESSED DATA SET INTO ALGORITHMIC EVALUATION

The pre-processed encoded dataset is implemented in the decision tree, random forest-neural network clustering and in neural network algorithm. The impact of these methods differs depending on the unique characteristics of the datasets being used. Identifying an algorithm that works effectively with the data set structure is essential for obtaining maximum efficiency [18]. The classification algorithms used for the E-Coli dataset are briefly described.

6.1. Decision tree algorithm

It has emerged with the development of the ID3 algorithm. The algorithm employs a divide and conquer strategy. Unlike the ID3 algorithm, this algorithm includes normalization operations. The algorithm calculates the information gain values, which are then utilised as a ratio [19]. At the time of the creation of the decision tree, it is possible to construct lower trees and move them to various levels. In constructing a tree, a single node is identified, and processing is initiated; The leaf node is recognized and represents the class if all of the samples are members of the same class. Therefore, the class of omic sequence is segregated according to the subclass distribution in datasets.

6.2. Random forest algorithm

When numerous decision trees are trained using different training clusters, the results are combined to generate the random forest method. The random forest algorithm generates different sub-training clusters. In the formation of training clusters, preloading is performed [20]. A procedure employing a random selection of properties is utilized to expand the trees. In the algorithm's operation, each node is partitioned with the encoded omic sequence of a similar subclass into branches based on the highest value among randomly selected

values from each node. Derived trees are obtained by selecting variables at random. The random values of derived trees will predict the class accuracy.

6.3. K-nearest neighbors clustering

In recognition of patterns or classification, the k-nearest neighbor algorithm is a method for classifying antibiotic resistance genomic class based on the problem adjacent to training examples. A majority vote of its neighbors classifies an object, with the same antibiotics class assigned to the class most common among its k nearest neighbors [21]. If k is equal to one, the antibiotics class is assigned to the class of its closest neighbor [22]. The class of closest neighbor is identified and predict the sequence.

6.4. Artificial neural network

This technique optimizes input attributes, weights, and network topology when implementing artificial neural networks (ANN) [23]. The encoded genomic sequence of a feedforward network with n inputs and m outputs. Information always flows from the input layer to the output layer (hence, unidirectional). To generalize the solutions generated by its outputs, A neural network is trained by taking the necessary coordinated procedures to adjust the synaptic weights and thresholds of its neurons. The sequence of stages used to train the network is known as the learning algorithm [24]. Consequently, the network can extract distinguishing characteristics of the same genomic sub-sets of E. coli from the dataset acquired during its execution.

7. RESULTS AND DISCUSSION

In this paper DNA sequence of the E. Coil 106 dataset of 57 sequential nucleotides is extracted. The test and train set into 89492, 237518 [25]. Then the model is trained and validated in a random forest, decision tree, and KNN algorithm. The model's main aim is mathematically mapping the data from input and output. Therefore, the parameters present in the algorithm learn the DNA sequence. The machine learning algorithms yield minimal accuracy in predicting sequences compared to the deep learning algorithm. From Table 1 and the machine learning algorithm prediction for the decision tree is 31.82%, the random forest is 88.99%, and KNN classifier is 62.57%. From Figure 3 the ANN accuracy in predicting the DNA sequence is 98% for same dataset is comparatively high. Therefore, the ANN learns in detail about the encoded data compared to the other machine learning algorithms. This is due to the sizeable omic sequence, and the series is challenging to read the genomic sequence. Learning rate accuracy for machine learning algorithm as shown in Table 2. From the provided information, it seems like a comparison has been made between different machine learning algorithms, specifically KNN classifier, decision tree, random forest, and neural network, in predicting antibiotic resistance in genomes. The interpretation of results is as following.

Table 1. Machine learning algorithm training and testing accuracy

Algorithm	Training accuracy (%)	Test accuracy (%)
Random forest	85.71	88.99
Decision tree	54.76	31.82
KNN classifier	100	62.57
ANN		98

Table 2. Learning rate accuracy for machine learning algorithm

Algorithm	Train accuracy (%)	Testing accuracy (%)
Random forest	90	84
Decision tree	50	47
KNN classifier	97	90

7.1. Learning curve comparison

Figures 5(a) to 5(c) illustrate the learning curve rates for KNN classifier, decision tree, and random forest models compared to a neural network. When compared to the other models, the neural network's learning curve rate seems to be higher, suggesting that it learns more quickly or performs better in terms of prediction accuracy. This suggests that the neural network is more efficient at capturing the intricate features present in the omic sequences, leading to superior predictive performance.

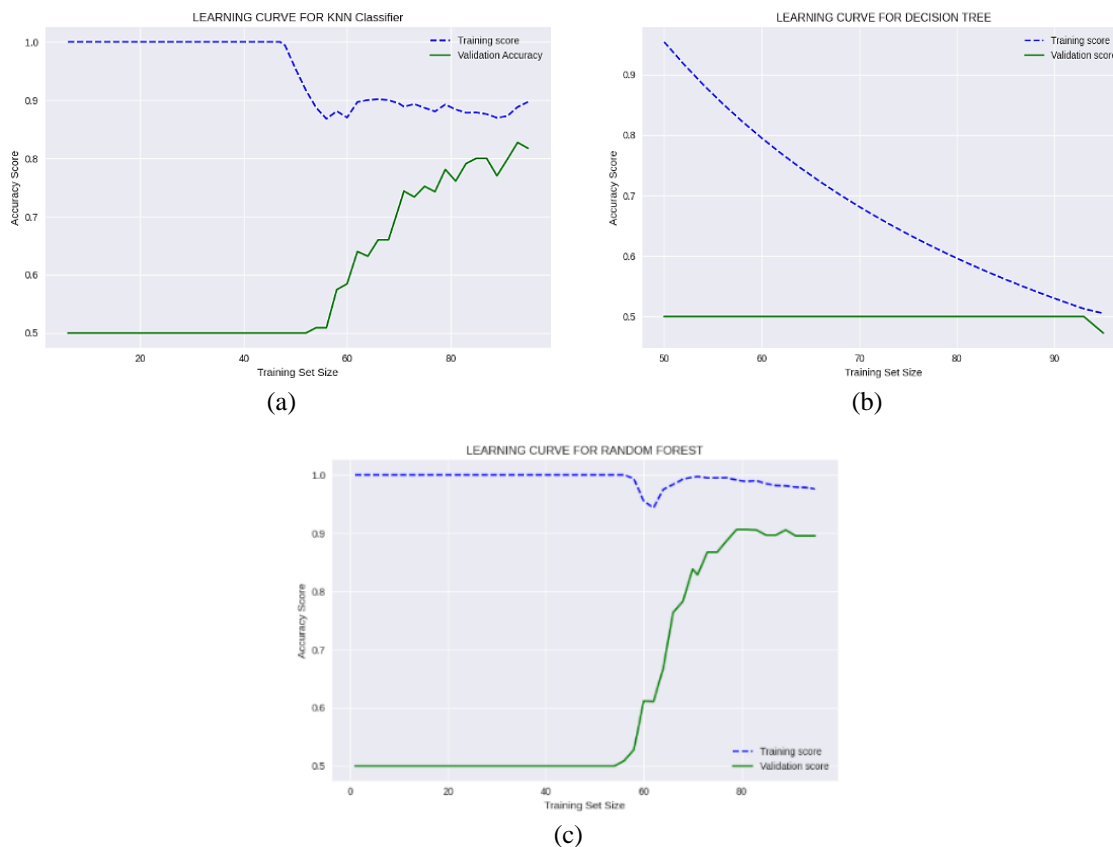


Figure 5. The learning curve for (a) KNN classifier, (b) decision tree, and (c) random forest

7.2. Model loss function

Figure 6 shows the loss function in the ANN, which quantifies the difference between predicted and actual values. A loss value of 0.1533 shows that this is the average deviation between the model's predicted and actual values. According to the statement, the accuracy of predicting genomic sequences is enhanced when the number of hidden layers in a neural network increases, as seen by a decrease in loss percentage. The methodology employed seems suitable for predicting omic sequences in new species and potentially applicable in drug discovery, implying the versatility and reliability of the neural network approach.

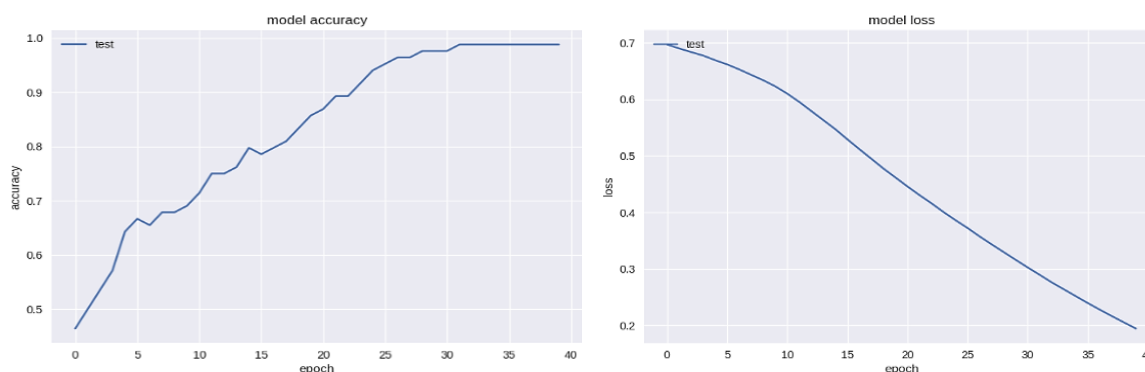


Figure 6. Model accuracy and model loss in ANN

7.3. Algorithm implemented to predict the model

The test and train set of data is 89492, 237518. The training and test set data to the machine learning algorithms. Therefore, the following machine learning algorithm is implemented in this paper: i) random forest ii) K-NN clustering, and iii) decision tree.

7.3.1. Decision tree

The decision tree calculates the weight of numerical values [26]. The mathematical modeling of decision tree is given in (1). The model predicts the accuracy rate by adding weights and comparing the test and trained datasets. The equation provided describes the mathematical modeling of a decision tree for predicting the accuracy rate by calculating the weight of numerical values.

$$n I_j = K_j Y_j - K_{\text{leftnode}(j)} Y_{\text{leftnode}(j)} - K_{\text{Rightnode}(j)} Y_{\text{Rightnode}(j)} \quad (1)$$

$n I_j$ is root node weight of numerical values j ; C_j is impurities present in the DNA sequence; W is weights of numerical DNA sequence; K_j represents the number of samples (instances) associated with the root node; Y_j represents the impurity measure (e.g., entropy or Gini impurity) of the root node; $K_{\text{leftnode}(j)}$ represents the number of samples associated with the left child node of the root node; $Y_{\text{leftnode}(j)}$ represents the impurity measure of the left child node of the root node; $K_{\text{Rightnode}(j)}$ represents the number of samples associated with the right child node of the root node; and $Y_{\text{Rightnode}(j)}$ represents the impurity measure of the right child node of the root node.

The equation computes the weight of the root node $n I_j$ by subtracting the weighted impurities of the left and right child nodes from the weighted impurity of the root node. This process helps determine the best split at each node of the decision tree based on the impurity measures of the child nodes. The weights W of numerical DNA sequences are used to calculate the impurity measure Y_j , which reflects the purity of the data at each node. In summary, (1) provides a mathematical framework for decision tree modeling, where the accuracy rate is predicted by considering the impurity measures and weights associated with different numerical values in the dataset.

7.3.2. Random forest

The random forest calculates the Weight by normalizing the all the decision tree values. The mathematical modeling of random forest tree is [27], [28] given in (2). The data points of one hot encoding identity are examined to determine the class of antibiotics. If instances i (encoding data points) are both at the same terminal node of a given tree, their proximity is enhanced by one. The distance matrix, which measures the degree of similarity between encoded data points of the genomic sequence, is produced by adding all terminal nodes in a forest [28].

$$G_i = 1 - \sum_{i=1}^C (p_i)^2 \quad (2)$$

where p_i represents the relative frequency of the class you are observing in the dataset and C represents the number of classes.

7.3.3. K-nearest neighbors-clustering

KNN predicts the data by finding the similitude between the data points in data set [29]. To obtain the best fit, take the mean of all the data points and use the Euclidean distance to calculate the gaps between each data point. The mathematical modeling of KNN classification is given in (3):

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \quad (3)$$

To find the Euclidean distance between the sample vectors and the centroids is parallelizable by separating the data into distinct subgroups and clustering samples in each 57-nucleotide subset independently (by the mapper). We divide the sample vectors into subgroups, compute the sum of vectors in each subset of 57 nucleotides in parallel, and eventually, the reducer adds up the partial sums and computes the new centroids [30].

7.3.4. Artificial neural network

The human brain's neurons inspire the ANN. The elements of ANN are nodes and their processing elements. These nodes are interconnected to form layers, by Adding all the layers and weight to give the predicted values [31]. The mathematical formula of ANN is:

$$A_j(t) = f(a_j(t), p_j(t), \phi_j) \quad (4)$$

where $A_j(t)$ is *activation function*, ϕ_j is optional threshold (learning rate), and $p_j(t)$ is neuron with j label. The activation function used in the algorithm is sigmoid because if the class accuracy is greater than 0.5% in the prediction it belongs to same antibiotic genomic class. If it is less than 0.5% the class of antibiotic resistance genome belongs to different class.

8. KEY FINDINGS

Traditional methods for analyzing DNA rely on sequence similarity, but deep learning offers a more powerful approach. By extracting complex features from DNA sequences, deep learning models achieved 98% accuracy in predicting antibiotic resistance genes in *E. coli* data, significantly exceeding the performance of simpler algorithms (31.82% - 88.99%). This ability to learn intricate patterns makes deep learning a valuable tool for drug discovery and analyzing DNA in new species.

9. LIMITATIONS OF THE STUDY

Deep learning thrives on large datasets. Consider how to address situations where limited data is available. Techniques like data augmentation (artificially creating new data points) can be explored. Deep learning models can be "black boxes," making it difficult to understand why they make certain predictions. Research methods that explain model reasoning are an active area of development. Training deep learning models requires significant computing power. Explore cloud-based resources or optimizing model architectures for efficiency. The model's performance on *E. coli* data is promising, but how well does it generalize to other species or data types? Test the model on a wider range of data to assess its robustness.

10. IMPLICATIONS FOR FUTURE RESEARCH

The future exploration of deep learning in genomic research which is the base for the emerging research area are as follows, Train models to predict disease risk based on DNA sequences. This could revolutionize preventative medicine. Use deep learning to identify DNA targets for new drugs, accelerating the drug development process. Develop models that tailor treatments to individual patients based on their unique genetic makeup. Combine DNA data with other biological information (RNA, protein) to create more comprehensive models. Explore techniques to make deep learning models more interpretable, allowing researchers to understand their decision-making processes. Investigate new deep learning architectures specifically designed for genomic data, potentially improving accuracy and efficiency.

11. CONCLUSION

This research effectively demonstrates the potential of deep learning, particularly ANNs, for tasks in omics. Compared to other machine learning algorithms, our findings are highlighting ANNs' ability to achieve superior accuracy in DNA sequence prediction, especially for large datasets. The low loss function (0.1533) further signifies the model's effectiveness. This approach boasts broad applications, encompassing tasks like predicting biological properties, discovering novel biological knowledge, and aiding in disease-related studies and data-driven pharmaceuticals. While this research lays a strong foundation, further exploration can unlock even greater potential. Future work could involve extending the model to classify different DNA sequences and identify their functionalities, developing methods to compare predicted sequences with existing databases, and refining the model to predict specific families within antimicrobial classes. By pursuing these areas and potentially incorporating specific details about your research methodology (e.g., network architecture, training data characteristics), we can further strengthen your conclusions and contribute significantly to the advancement of deep learning in omics. This research delves into the exciting potential of deep learning for biological prediction tasks. By analyzing DNA sequences, the model can predict various properties, like specific functionalities of certain regions, or even uncover novel biological knowledge by grouping genes based on interactions. Additionally, it holds promise for disease prediction by identifying genes impacted by pathogens or assessing patient risk. Beyond biology, applications extend to data-driven pharmaceuticals, where the model can predict the effect of chemical agents on cells or even identify agents administered to a sample. Despite the promising results with a simple architectural model, further exploration is crucial to unlock its full potential. Future work should focus on expanding the model's capabilities through tasks like DNA sequence classification to identify their functionalities, comparing predicted sequences with existing databases to assess their novelty, and refining the model to predict specific families within antimicrobial classes. These advancements will solidify the contributions of this research to the burgeoning field of deep learning in omics, paving the way for groundbreaking discoveries across various biological and medical fields.




REFERENCES

- [1] S. C. Rastogi, P. Rastogi, and N. Mendiratta, *Bioinformatics: Methods and Applications-Genomics, Proteomics and Drug Discovery*. PHI Learning Private Limited, Delhi, India, 2022.
- [2] G. Goussarov, M. Mysara, P. Vandamme, and R. V. Houdt, "Introduction to the principles and methods underlying the recovery of metagenome-assembled genomes from metagenomic data," *Microbiology Open*, vol. 11, no. 3, 2022, doi: 10.1002/mbo3.1298.




- [3] R. Kumar, M. Gupta, and M. Sarwat, "Bioinformatics in drug design and delivery," *Computer Aided Pharmaceutics and Drug Delivery: An Application Guide for Students and Researchers of Pharmaceutical Sciences*, Springer, 2022, pp. 641–664, doi: 10.1007/978-981-16-5180-9_21.
- [4] D. A. Benson, I. K. Mizrahi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Research*, vol. 33, no. 1, pp. D34–D38, 2005, doi: 10.1093/nar/gki063.
- [5] C. Choudhury, N. A. Murugan, and U. D. Priyakumar, "Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods," *Drug Discovery Today*, vol. 27, no. 7, 2022, doi: 10.1016/j.drudis.2022.03.006.
- [6] C. Phumichai *et al.*, "Genome-wide association mapping and genomic prediction of yield-related traits and starch pasting properties in cassava," *Theoretical and Applied Genetics*, vol. 135, pp. 145–171, 2022, doi: 10.1007/s00122-021-03956-2.
- [7] S. Tsimenidis, E. Vrochidou, and G. A. Papakostas, "Omics data and data representations for deep learning-based predictive modeling," *International Journal of Molecular Sciences*, vol. 23, no. 20, 2022, doi: 10.3390/ijms232012272.
- [8] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012, doi: 10.1093/bioinformatics/bts565.
- [9] Y. Long, M. Wu, C. K. Kwok, J. Luo, and X. Li, "Predicting human microbe–drug associations via graph convolutional network with conditional random field," *Bioinformatics*, vol. 36, no. 19, pp. 4918–4927, 2020, doi: 10.1093/bioinformatics/btaa598.
- [10] J. Zhang, W. Xie, C. Wang, R. Tu, and Z. Tu, "Graph-aware transformer for skeleton-based action recognition," *The Visual Computer*, vol. 39, no. 10, pp. 4501–4512, 2022, doi: 10.1007/s00371-022-02603-1.
- [11] O. A. M.-López *et al.*, "A review of deep learning applications for genomic selection," *BMC Genomics*, vol. 22, pp. 1–23, 2021, doi: 10.1186/s12864-020-07319-x.
- [12] X. Zhang, F. T. Chan, and S. Mahadevan, "Explainable machine learning in image classification models: An uncertainty quantification perspective," *Knowledge-Based Systems*, vol. 243, 2022, doi: 10.1016/j.knosys.2022.108418.
- [13] S. Xu, M. Hao, G. Liu, Y. Meng, J. Han, and Y. Shi, "Concrete crack segmentation based on convolution–deconvolution feature fusion with holistically nested networks," *Structural Control and Health Monitoring*, vol. 29, no. 8, 2022.
- [14] Z. Zhang *et al.*, "Deep learning in omics: a survey and guideline," *Briefings In Functional Genomics*, vol. 18, no. 1, pp. 41–57, 2019, doi: 10.1093/bfpg/ely030.
- [15] Y. Ren *et al.*, "Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning," *Bioinformatics*, vol. 38, no. 2, pp. 325–334, 2022, doi: 10.1093/bioinformatics/btab681.
- [16] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing data preprocessing in credit classification: One-hot encoding or imputation?," *Emerging Markets Finance and Trade*, vol. 58, no. 2, pp. 472–482, 2022.
- [17] P. Mahé and M. Tournoud, "Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–11, 2018, doi: 10.1186/s12859-018-2403-z.
- [18] A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, and A. Sharma, "IoT and machine learning approaches for automation of farm irrigation system," *Procedia Computer Science*, vol. 167, pp. 1250–1257, 2020, doi: 10.1016/j.procs.2020.03.440.
- [19] J. R. Quinlan, "Program for machine learning," *Machine Learning*, vol. 16, pp. 235–240, 1993, doi: 10.1007/BF00993309.
- [20] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [21] M. L. -Williams, "Case studies in the data mining approach to health information analysis," *IEE Colloquium on Knowledge Discovery and Data Mining*, London, UK, 1998, pp. 1–4, doi: 10.1049/ic:19980641.
- [22] J. S. Raikwal and K. Saxena, "Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set," *International Journal of Computer Applications*, vol. 50, no. 14, pp. 35–39, 2012.
- [23] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data," *PLoS computational biology*, vol. 14, no. 4, 2018, doi: 10.1371/journal.pcbi.1006076.
- [24] I. N. D. Silva *et al.*, "Artificial neural network architectures and training processes," *Artificial Neural Networks*, pp. 21–28, 2017, doi: 10.1007/978-3-319-43162-8_2.
- [25] C. N. Vassallo, C. R. Doering, M. L. Littlehale, G. I. Teodoro, and M. T. Laub, "A functional selection reveals previously undetected anti-phage defence systems in the E. coli pangenome," *Nature Microbiology*, vol. 7, no. 10, pp. 1568–1579, 2022, doi: 10.1038/s41564-022-01219-4.
- [26] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004, doi: 10.1002/cem.873.
- [27] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017, doi: 10.17849/insm-47-01-31-39.1.
- [28] Y. Qi, "Random forest for bioinformatics," in *Ensemble machine learning: Methods and applications*, 2012, pp. 307–323, doi: 10.1007/978-1-4419-9326-7_11.
- [29] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-10358-x.
- [30] C. T. Chu *et al.*, "Map-reduce for machine learning on multicore," in *Advances in Neural Information Processing Systems*, MIT Press, pp. 281–288, 2006.
- [31] G. A. Anastassiou, "Generalized symmetrical sigmoid function activated neural network multivariate approximation," *Journal of Applied and Pure Mathematics*, vol. 4, no. 3, pp. 185–209, 2022.

BIOGRAPHIES OF AUTHORS






S. M. Shifana Rayesha    holds a Masters in Engineering (M.E.) Degree from Anna University, India in 2016. He also received his B.Tech. (CSE) from B. S. Abdur Rahman Institute of Science and technology, India in 2014, respectively. She is working as Assistant Professor in B. S. Abdur Rahman Crescent Institute of Science and Technology for past 4 years. She is doing her research in bioinformatics and deep learning algorithms. She can be contacted at email: shifana@crescent.education.






Dr. Aisha Banu    received the Ph.D. in Computer Science and Engineering from the Anna Univeristy. She is a Professor at Department of Computer Science since 1998. Her research interests are in computer architecture, informational retrieval, pattern recognition, and deep learning architecture. She is currently working as Head of the Department in Computer Science and Engineering. Totally she published 45 research papers. She can be contacted at email: aisha@crescent.education.



Dr. Afzalur Rahman    holds a Ph.D. (Commerce) in Central University of Bihar, Patna, Master of Philosophy (Commerce) Master of Business Administration (MBA). He is currently working as Professor in School of Commerce in Presidency University. He is skilled in accounts and finance, data analytics using Python. He is working as professor for 8 years. His research areas of interest include biometrics, accounts and fiancé, blockchain, and data analytics using python. He also published 10 research papers. He can be contacted at email: afzalur@outlook.com.



Dr. Sharon Priya    received the Ph.D. in Computer Science and Engineering from the Anna Univeristy. She is working as assistant professor at Department of Computer Science since 2009. Her research interests are in computer architecture, informational retrieval, pattern recognition, and deep learning architecture. She is currently working as Associate Professor in Computer Science and Engineering. Totally she published 10 research papers. She can be contacted at email: sharonpriya@crescent.education.