Promoter Prediction in DNA Sequences of Escherichia Coli Using Machine Learning Algorithms

Anveshrithaa S, Balamurugan Aathavan, Jaisankar N

Abstract: The advent of Artificial Intelligence and Machine learning has brought many advancements in the field of computational biology. The significant improvement in the field of Machine Learning has made way for opportunities in demanding fields by enabling machines to automatically learn from data without any explicit programming and improving their ability to solve complex problems through learning and experience. Bioinformatics is one among the many applications of Machine Learning where it is widely utilized especially for classification and identification of patterns in DNA in genomics. The purpose of this research is to implement and improve various Machine learning models including ensemble learning namely boosting and bootstrap aggregation, neural network-based methods, Support Vector Machine, Naïve Bayes, k-nearest neighbors and decision tree for predicting transcription start sites (promoters) in the DNA sequences of a common bacteria, Escherichia coli. The performance of the models is optimized through hyper-parameter tuning for improved prediction. This paper also focuses on the comparison of these machine learning classification models to determine the model that best predicts the promoters in the DNA sequences.

Keywords: Classification, DNA, Ensemble learning, Machine learning, Naïve Bayes, Neural Networks, Promoters.

1. INTRODUCTION

The Deoxyribonucleic Acid (DNA) is a complex organic molecule containing genetic information for development, functioning and growth of all living beings. DNA is also the principal unit of heredity in organisms. It contains the protein-encoding information. DNA is made up of simpler monomeric units called nucleotides. The structure of DNA is a double-helix polymer, a spiral consisting of two DNA strands wound around each other [1]. Gene is a segment of DNA that contains information for the synthesis of protein. Information from a gene is used in the synthesis of a functional gene product, often proteins, by a process called gene expression. The initial step in this process is the copying of a gene's DNA sequence to form an RNA molecule, which is performed by the enzyme RNA polymerase. The DNA sequences contain promoter sequences located near the beginning of a gene, which define where the transcription begins. When RNA polymerase binds to a promoter sequence, transcription is initiated. The stretch of DNA sequence following the promoter region is first transcribed into mRNA which carries codes from DNA to protein synthesis sites, as information contained in DNA cannot be decoded directly into proteins.[2] Gene prediction, which involves identifying the regions of genomic DNA that encode genes, is a serious challenge in the field of computational genomics. Promoter prediction is an important and common facet of many gene prediction methods. Many algorithms have been developed as an aid to detection of promoters in genomic sequences. Many researchers are working towards the aim of improving the current knowledge of this domain of bioinformatics. Machine learning is a branch of computer science that focuses on the study and development of algorithms and statistical models that facilitates computers to learn for themselves from data, without explicit instructions. Machine learning models are widely used for detecting promoter regions in DNA. The objective of this work is to implement and assess various machine learning models to detect promoter regions in the DNA of Escherichia coli bacteria.

- ***Anveshrithaa S,*** *School of Computer Science & Engineering, VIT, Vellore, India. E-mail: anveshrithaas@gmail.com*
- ***Balamurugan Aathavan,*** *School of Information Technology & Engineering, VIT, Vellore, India. E-mail: aathitheva@gmail.com*
- ***Jaisankar N,*** *School of Computer Science & Engineering, VIT, Vellore, India. E-mail: njaisankar@vit.ac.in*

2 LITERATURE REVIEW

Various machine learning techniques and algorithms were proposed by researchers for the prediction of promoters in bacterial DNA sequences. Towel et al. presented the KBANN (Knowledge-Based Artificial Neural Networks) hybrid learning system that combines empirical and explanation-based learning to overcome the problems of each approach, for promoter prediction. In this paper, he demonstrated how hybrid systems like KBANN are superior to empirical and explanation-based learning systems in terms of classification accuracy. It demonstrates how combining domain knowledge can yield better performance of Neural Networks. This approach was compared with other algorithms like simple ANN, decision tree learning with ID3 algorithm and k-nearest neighbor classification [7]. Tavares et al. compared the performances of various machine learning methods for predicting promoter sequences in the DNA of E. coli bacteria. In this comparative study, probabilistic methods, including Hidden Markov Model (HMM) and Bayesian methods gave better results in terms of accuracy. Also results obtained by Neural networks were comparable to that of the probabilistic methods [3]. Gordon et al. discussed an approach to predict promoters and Transcription Start Sites in E. coli, based on an ensemble of Support Vector Machines and this classifier is then combined with Position Weight Matrices [4]. Maleki et al. used Expectation Maximization and Support Vector Machine classifier (EMSVM) to perform promoter detection. The proposed algorithm is of two stages where the data is clustered using the Expectation Maximization (EM) in the first stage and then they are classified using Support Vector Machine (SVM). This technique obtained very high accuracy in detecting promoters in DNA [5]. Umarov et al. in their work, made use of Convolutional Neural Networks (CNN) using Keras with Theano library as backend, utilizing GPU for fast neural network training, to analyse complex prokaryotic and eukaryotic promoters in DNA of four organisms: humans,

plants and two bacteria (E. coli and Mycoplasma pneumonia) [11]. Zhang et al. implemented a feed forward neural network for promoter prediction in bacteria, that can extract statistical characteristics of promoters effectively. The result of this work also suggests that the number of hidden layers seems to have no significant effect in the promoter prediction precision [12]. Askary et al. presented a modified artificial neural network fed by nearest neighbors to predict the promoters in E. coli with high sensitivity and precision [13]. Hong-Hee et al. proposed an ensemble approach called EnsemPro (Ensemble Promoter) which combines the prediction results of already existing predictors, and uses ensemble schemas like majority voting, weighted voting, and Bayesian approach to predict promoters in human gene sequences [14]. Huang et al. developed a prediction algorithm that can increase the promoter detection accuracy. They presented two methods to calculate all possible patterns of features of promoters: FTSS (Fixed Transcriptional Start Site) uses the known transcription start sites positions of promoter sequences and NTSS (Nonfixed TSS) uses the TSS positions of promoters that are assumed to be unknown, hence not taking the absolute positions into consideration [15]. Ryasik et al. presented a systematic approach which selects less correlated physical properties of DNA for classification. The proposed classifier, along with using sequence and static physical properties of DNA sequence also takes the dynamic properties of DNA into consideration [16]. Various researches are carried out on the application of machine learning techniques in the field of molecular biology. Weinert et al. in their work, discuss a fast and efficient biomolecular classification methodology based on multilayer perceptron neural networks which is used to classify protein and infer their functions by analysing the structural similarity [6].

3 PROPOSED ARCHITECTURE

The objective of the proposed work is to develop a model that predicts the promoters (transcription start sites) in the DNA sequences of E. coli bacteria using machine learning approach. Fig. 1 represents the block diagram of the architecture of the proposed system.

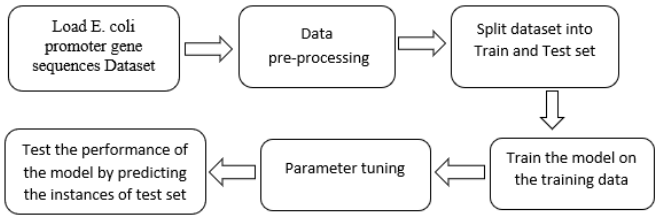

***Fig. 1**. Proposed Architecture*

3.1 Dataset

The E. coli genome dataset used for this work contains 106 instances in total, with the two classes distributed equally, 53 being promoter sequences (positive cases) and the other half being non-promoter sequences (negative cases). This dataset was obtained from the UCI machine learning repository collated by Towell et al. [7] to help evaluate a hybrid learning algorithm, KBANN that uses examples to inductively refine pre-existing knowledge. Every instance contains 57 base-pair positions starting at position -50 and ending at position +7.

3.2 Pre-processing

For any machine learning analysis, the first step is data pre-processing. Pre-processing of raw data to obtain useful information that is feasible for analysis is crucial as it determines the efficiency of the analysis as well as the quality of the output. In the proposed work, the dataset being a biological sequence dataset, has uncertainties. Thus, the initial approach is to pre-process the dataset to make it suitable for classification using machine learning models.

3.3 Training and Testing

Firstly, the dataset is divided into training set and test set and each model is separately trained on the training set. After training the model, the parameters are tuned using techniques like Grid search and cross validation to get better accuracy in prediction. Parameter tuning is the process of selecting the optimal set of hyperparameters for the model for improved accuracy. Then its performance is tested on the test set in order to give an unbiased estimate of the performance of the model.

3.4 Implementation

This work is implemented in python, using various libraries for python. Keras, an open-source neural networks library running on TensorFlow was used for building the artificial neural networks. Scikit-learn, another machine learning library in python featuring many classification, regression and clustering algorithms is also used for the classification of DNA sequences. The machine learning techniques that are implemented in this work, for the prediction of promoter sequences are as follows.

A. Ensemble Learning

Ensemble learning, as the name suggests, is an efficient technique which is an ensemble of various machine learning algorithms, aimed at performing high accuracy predictions. This model possesses an advantage of alleviating the problem of overfitting the training data in small sized datasets by averaging and combining many different models [8]. Two important approaches in ensemble learning are Bootstrap Aggregation, also known as Bagging and Boosting. In Bagging method, various classifiers are trained on the training data and the prediction is done based on majority voting by all the trained classifiers. This method reduces variance error in the prediction and avoids overfitting. Boosting emphasizes on sequential learning of the classifiers. Initially the dataset is classified by an algorithm and equal weights are given to each observation. If there is error in prediction, then higher weight is assigned to the data points which were predicted incorrectly, with more focus given to them by the classification algorithm. In this way, every classifier dictates what features the subsequent classifier will focus on. This method reduces bias. We use ADABoost and Gradient Boosting in this work, which are boosting approaches in ensemble learning. For an ensemble model of N classifiers, and all classifier models having same error rate ϵ, the error rate of majority voting, ϵ_{mv} is given by

$$\epsilon_{mv} = \sum_{i=\lceil \frac{N}{2} \rceil + 1}^{N} \binom{N}{i} \epsilon^{i} (1-\epsilon)^{N-i}$$

Given this, the majority voting error rate ϵ_{mv} monotonically decreases and tends to 0 as N→ ∞ [9].

B. Artificial Neural Network

Artificial neural network is a computational model that emulates the functioning of the human brain. It replicates the way in which a human brain learns and processes information. A neural network is an interconnection of huge number of processing units called neurons with a weight associated with

each connection. Neural networks are used for complex computational analysis, pattern recognition, prediction, classification and is the foundation for most of the advances in what is called as Artificial Intelligence. A Deep learning Neural Network has multiple layers including the input, output and the hidden layers. Neural Networks learn through a technique called backpropagation which allows the network to adjust the weights of the hidden layer in order to minimize the value of the error obtained by the cost function. This optimization is done using a method called Gradient descent which calculates the derivative of the loss function with respect to the weights. The cost function used in building the neural network for the purpose of this work is Binary cross-entropy, which is used for binary classification. An Activation function is responsible for the non-linearity of the model. It determines if a neuron should be activated or not, and the output signal is then fed as input to another neuron. The activation functions used in this work are ReLU (Rectified Linear Unit) function for the input layer and the three hidden layers and Sigmoid function for the output layer of the ANN. ReLU function is given by f(x)=max(0, x) indicating that if the input is negative, the ReLU function converts it to zero and the neuron is not activated. But for any positive value x, it returns the same. It is computationally economical compared to other activation functions. The sigmoid function is given by

$$f(x)=\frac{1}{(1+e^{-x})}$$

It takes an input and maps it to a value between 0 and 1. Thus, this function is suitable for the output layer of the ANN that classifies as promoter or not.

C. Naïve Bayes

The Naïve Bayes is a probabilistic model for classification, which works on the principle of Bayes theorem of conditional probability, mathematically given by

$$P(A|B) = (P(B|A)\ P(A))/(P(B))$$

The probabilistic approach used in this work is Gaussian Naïve Bayes classifier which assumes Gaussian distribution of continuous data associated with each feature and hence the mathematical formulation of the conditional probability is given by,

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(\frac{-(x-\mu_y)^2}{2\sigma_y^2}\right)$$

where σ is the standard deviation and μ is the mean.

D. Support Vector Machine

Support Vector Machine is a machine learning algorithm that performs classification by constructing a hyper plane that best segregates the data classes in an N-dimensional space by maximizing the distances between the nearest support vectors and the hyperplane. If the data is not separable by a linear hyperplane, then the kernel maps the data points from a low dimensional space to a higher dimensional feature space, thus making it separable by a hyperplane. Different kernel functions are used in SVM algorithm, one of which is the Gaussian kernel or the Radial Basis Function (RBF) kernel which is given by the formula

$$K(x^{(i)}, x^{(j)}) = exp\left(-\frac{||x^{(i)}-x^{(j)}||^2}{2\sigma^2}\right)$$

E. K- Nearest Neighbors

K Nearest Neighbors is a classification algorithm which uses similarity measure to classify data. Classification of a data point is done based on the majority vote of its neighbors, assigning it to the class that has the highest frequency amongst its K nearest neighbors, measured by a distance function. In this work, we have taken the value of K to be 3, by considering the three nearest neighbors of any data point and the distance metric used is minkowski.

F. Decision Tree

Decision tree is a supervised learning algorithm majorly used for classification. It uses a tree-like graph structure that derives the possible outcome of each decision. Decision trees classifies the data points into classes containing homogeneous data points by identifying the most significant variable whose value can best segregate the data points into homogeneous sets that are heterogeneous to each other. Decision tree divides the nodes on all the variables and choses the split that gives the highest degree of homogeneity within a node.

3.5 Accuracy Measures

The primary objective of the work is to build machine learning algorithms that have better accuracies in predicting promoters than the existing systems. The accuracy of the model is calculated using the formula,

$$Accuracy= (TP+TN)/(TP+TN+FP+FN)$$

Where TP (True Positive) denotes the number of promoter instances (positive) that are correctly classified as promoters, TN (True Negative) denotes the number of non-promoter instances (negative) that were correctly classified as negative, FP (False Positive) gives the number of negative instances wrongly classified as positive and FN (False Negative) denotes the number of promoters that were incorrectly classified as negative. These parameters are used to calculate various performance metrics like recall score, also known as sensitivity, which is the fraction of positive instances predicted correctly, precision, which is the fraction of predicted positives instances that are actually positive, f1 score which is the harmonic mean of recall and precision and the specificity of the model. Higher the f1-score, better the performance of the model. It is important to consider these performance metrics as accuracy is not always reliable and it can be deceiving in case of an unbalanced dataset. These metrics are given as follows.

$$Recall = TP/(TP+FN)$$

$$Precision = TP/(TP+FP)$$

$$F1= (2 \times (precision \times sensitivity))/ (precision + sensitivity)$$

4 RESULTS AND DISCUSSION

The results obtained for each model are summarized in Table 1. It contains the accuracy and other performance metrics of the model such as recall, precision and f1-score. Fig. 2 visualizes the performance metrics of each model in the form of a bar graph and Fig. 3 graphically represents the number of correctly and incorrectly classified instances for each model.

TABLE 1. *PERFORMANCE METRICS OF EACH MODEL*

Method	Accuracy (%)	Precision	Recall	F1-score
ANN with back-propagation	96.87	0.97	0.97	0.97
Ensemble (Bootstrap aggregation)	96.29	0.95	0.97	0.96
SVM (Linear kernel)	96.29	0.95	0.97	0.96
Naïve Bayes	92.59	0.92	0.94	0.92

Gradient Boosting	88.88	0.88	0.91	0.89
ADABoost (Boosting)	85.18	0.86	0.88	0.85
K-Nearest Neighbors	85.18	0.84	0.84	0.84
Decision Tree	77.77	0.78	0.80	0.78

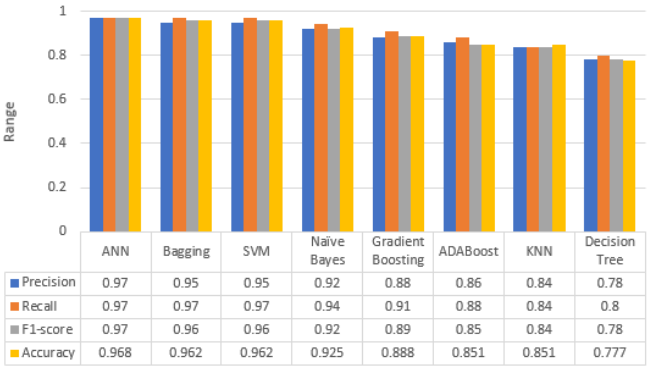

***Fig. 2**. Performance Metrics*

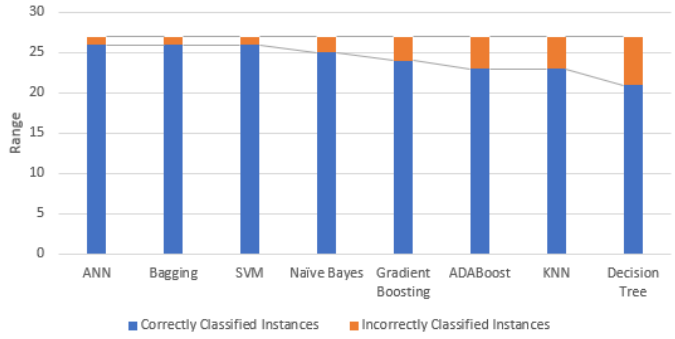

***Fig. 3**. Correctly and incorrectly classified instances*

From Table 1, we conclude that Neural network trained with back-propagation best classifies the promoters in the DNA sequences. This high accuracy is attributed to the optimization through back-propagation and hyperparameter tuning using grid search. We find that other models including the Bootstrap Aggregation which is an ensemble learning method and Support Vector Machine with linear kernel are equally effective in classifying the DNA sequences. This observation is not surprising considering the fact that neural networks and SVM have the ability to form complex hyperplanes in n-dimensional space to classify data with high accuracy. Thus, this dataset having large number of features, is handled well by this property. Bagging reduces overfitting the training data in small sized datasets by averaging and combining many different models. This is one of the reasons for obtaining high accuracy in promoter prediction by this ensemble model as the dataset used is small. Also, Naïve Bayes, being a probabilistic method, is suitable to handle the specific nature of biological sequences that have high uncertainties and thus achieved comparable results to those of neural networks, ensemble method and SVM. Boosting algorithms like Gradient Boosting and ADABoost are not on par with neural-networks, bagging, SVM and Bayesian classification method. Though K- nearest neighbor is considered to be a powerful machine learning technique, it did not yield satisfactory results compared to others, for this classification problem and for the dataset used in this work because of the fact that KNN suffers from the curse of dimensionality. As far as decision tree algorithm is concerned, a small change in the dataset can lead to a large change in the structure of the optimal decision tree. With the high uncertainties in DNA sequences, this attribute of decision tree has resulted in poor performance in predicting the promoters in the DNA sequences. Comparing our results to those of already published work of Towell et al.[7], in which the proposed hybrid approach named Knowledge Based Neural Network (KBNN) yielded an accuracy rate of 96.22%, our model of neural networks, SVM and bagging outperform it. Also, comparing with the results obtained in the comparative study by Tavares et al [3], we have obtained higher accuracy than the Hidden Markov Model (HMM) which has an accuracy of 92.45% and the probabilistic methods like Naïve Bayes that has an accuracy of 92.45%. The neural network trained with back-propagation implemented in this work has performed better than the feed-forward neural network that was reported with an accuracy of 93.40% by Taveras et al. [3]

5 CONCLUSION

The accuracy of the proposed system achieved from this work is better than the previously proposed systems. Hyper-parameter tuning for the models have helped in achieving high accuracies. From this work we conclude that ensemble methods, neural networks, SVM, and Bayesian method exhibit comparable results, outperforming other models and are suitable for tasks like pattern recognition, sequence aligning and other biological sequence analysis, keeping in mind that biological sequences being specific in nature, exhibit uncertainty in their analysis. The result of this work justifies why these machine learning models are traditionally being used in bioinformatics. Future work focuses on implementing new algorithms for improved promoter prediction as well as reducing the complexities associated with the analysis of biological sequences.

REFERENCES

[1] J. D. Watson & F. H. C. Crick, “A Structure for Deoxyribose Nucleic Acid”, Nature 171, pp.737–738, 1953.
[2] Nelson, D.L., Cox, M.M.,” Lehninger Principles of Biochemistry”, 4th ed. W.H. Freeman, Chicago, 2006.
[3] G. Tavares, Heitor S. Lopes, Carlos R., “A Comparative Study of Machine Learning Methods for Detecting Promoters in Bacterial DNA Sequences", 4th International Conference on Intelligent Computing, Proceedings, Shanghai, 2008.
[4] Gordon JJ, Towsey MW, Hogan JM, Mathews SA, Timms P., “Improved prediction of bacterial transcription starts sites”, Bioinformatics 22, pp.142-148, 2006.
[5] Ahmad Maleki, Vahid Vaezinia, Ayda Fekri, “Promoter Prediction in Bacterial DNA Sequences Using Expectation Maximization and Support Vector Machine Learning Approach”, Journal of Data Mining in Genomics and Proteomics, vol. 6(2), 2015.
[6] Weinert, W., Lopes, H.S., “Neural Networks for Protein Classification”, Appl. Bioin-formatics, vol. 3, pp.41–48, 2004.
[7] Towell, G., Shavlik, J., Noordewier, M., “Refinement of Approximate Domain The-ories by Knowledge-based Artificial Neural Networks”, 8th National Conference on Artificial Intelligence, AAAI Press, Menlo Park, pp.861–

866, 1990.
[8] Dietterich TG, “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization”, Machine Learning, vol. 40, pp.139–158, 2000.
[9] Lam L, Suen Y., “Application of majority voting to pattern recognition: an analysis of its behaviour and performance”, IEEE Transactions on Systems, Man, and Cybernetics, vol. 27, pp.553–568, 1997.
[10] Alipanahi B, Delong A, Weirauch MT, Frey BJ., “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning”, Nature Biotechnology, vol. 33(8), pp.831–838, 2015.
[11] R.K. Umarov, V.V. Solovyev, "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks", Plos One, vol. 12, no. 2, pp. e0171410, 2017.
[12] Fan Zhang, D. Kuo, and A. Brunkhors, "E. coli Promoter Prediction Using Feed-Forward Neural Networks," Proceedings of the 28th IEEE EMBS Annual International Conference, vol. 6, pp. 2025-2027, 2006.
[13] Askary A, Masoudi-Nejad A, Sharafi R, Mizbani A, Parizi SN, Purmasjedi M., “N4: A precise and highly sensitive promoter predictor using neural network fed by nearest neighbors”, Genes Genet. Syst., vol. 84, pp. 425-430, 2009.
[14] Won H.-H., Kim M.-J., Kim S., Kim J.-W., “EnsemPro: An ensemble approach to predicting transcription start sites in human genomic DNA sequences”, Genomics, vol. 91(3), pp. 259-266, 2008.
[15] Huang J., Yang, C-B., Tseng, K-T., “Algorithms for promoter prediction in DNA sequences”, Taiwan, 1–5, 2002.
[16] Ryasik A, Orlov M, Zykova E, Ermak T, Sorokin A, “Bacterial promoter prediction: Selection of dynamic and static physical properties of DNA for reliable sequence classification”, Journal of Bioinformatics and Computational Biology Vol. 16, 2018.
[17] R. Fang, S. Wu, W. Zhang, Q. Liu and Y. Song, "A new algorithm of promoter prediction and identification", The Fourth International Workshop on Advanced Computational Intelligence, Wuhan, pp. 236-241, 2011.
[18] M.I. Monteiro, M.C.P. Souto, L.M.G.Gonalves and L.F. Agnez-Lima, “Machine learning techniques for predicting bacillus subtilis promoters”, Proc. of the Brazilian Symposium on Bioinformatics, vol.3594, Springer-Verlag, pp.77- 84, 2005.
[19] Robertas Damaševicius, "Optimization of SVM Parameters for Pattern Recognition In DNA Sequences", World Academy of Science, Engineering and Technology, vol.7, 2005.
[20] V. Palade, "Ensembles of classifiers and their application to bioinformatics problems," 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), Perth, WA, pp. 1-1, 2010.
[21] Robertas Damaševicius,”Optimization of SVM Parameters for Pattern Recognition In DNA Sequences", World Academy of Science, Engineering and Technology, vol.7, 2005.