Grenze International Journal of Engineering and Technology, January Issue

Enhancing Cystic Fibrosis Diagnosis through Motif Driven ML Models

Abhija R Nair[1] and Prathibha Mol C. P[2]
[1-2]Department of Computer Science and Engineering, Amrita School of Computing, Amritapuri, 690525, Kerala, India
Email: abhijarnair@gmail.com, prathibhamolcp@am.amrita.edu

***Abstract*— The prediction of genetic diseases is a critical area of research in genomics. With the increasing accessibility of DNA sequencing, the identification of genetic variations associated with diseases has become feasible. Cystic fibrosis (CF) is a genetic disorder characterized by the buildup of thick, sticky mucus that can damage various organs in the body. In this study, we proposed a novel approach for classifying CF-affected and not-affected genetic sequences using motif discovery and machine learning techniques. We utilize Gibbs sampling for motif discovery and k-mer representation for sequence transformation. Subsequently, we train and evaluate four machine learning models - Random Forest, Decision Tree, Logistic Regression and Support Vector Machine (SVM) - to classify genetic sequences based on the identified motifs. Our results demonstrate the effectiveness of motif- guided feature selection in improving classification accuracy and highlight the potential of machine learning in CF diagnosis. In our approach, we got more accuracy in the Random Forest model and future researchers can improve the accuracy by collecting more DNA sequences.**

Keywords—Cystic fibrosis, DNA motif discovery, Gibbs sampling, machine learning, Random Forest.

I. INTRODUCTION

DNA (Deoxyribonucleic acid) is the genetic material of humans and almost all other organisms. DNA, [1] available in most of the body's cells, contains the instruction that needs to be utilized by plenty of viruses and for organism's growth, development, and reproduction. Adenine, Cytosine, Guanine, and Thymine are the four chemical bases, that is A, C, G, T involved in the formation of helical structure, resembling a twisted ladder.
Cystic fibrosis (CF) [2] is a severe hereditary disorder that mainly affects the respiratory and digestive systems of the body. The genetic disease cystic fibrosis, caused by mutations in the Cystic Fibrosis Transmembrane Conductance Regulator gene (CFTR), has disastrous consequences for the digestive and respiratory systems. Sweat chloride testing is considered as the golden standard for CF diagnosis, however all of this process requires specialized equipment, sample collection, skilled technicians, etc. The thick, sticky mucus causes bacterial infection, and it affects all in the lungs. With the emergence of a great number of studies in bioinformatics and machine learning, it has made searching genetic data possible in new ways for hidden patterns. One of the techniques in finding common patterns of DNA sequences, called motif discovery [3], has been the cornerstone in elucidating regulatory elements and their relations to diseases. The project's main objective is to use machine learning techniques together with DNA motif discovery to see if there is any way to differentiate among these different DNA sequences and identify the most effective classifier based on some accuracy score.

Grenze ID: 01.GIJET.11.1.505_2

Here, we have used DNA sequences of two groups, not affected and affected in FASTA format, from the NCBI database. After merging, it made a single data frame, and the CF status for each sequence was labelled. We utilized the Gibbs sampling motif discovery technique to identify significant motifs within the DNA sequences mentioned above. This technique is an effective one for finding recurrence patterns within biological sequences [4]. The found motifs were then attached to the data frame; afterwards, we converted the DNA sequences to their 3-mer representations using the Count Vectorizer technique. This will convert the DNA sequences into feature vectors that can be applied as input in machine learning algorithms [5]. The data was then split into the training and testing set. The algorithms considered in this step were Logistic Regression, Decision Tree, SVM and Random Forest. The performance of the models was checked by their classification accuracy.

From the results, the Random Forest model yielded the highest accuracy, while the SVM and Logistic Regression models were close in scores. However, the Decision Tree model has got less accuracy compared to other models, this implied that the model could over-fit and had low capability in capturing complex patterns in the data. On the other hand, the Random Forest model performed the best due to ensemble learning; it pools a large number of decision trees for classification, which increases the model's accuracy and robustness.

This study has far-reaching implications. Determining the most accurate machine learning model [6] for classification of DNA sequences related to cystic fibrosis means that diagnosis can become more accurate and, in the future, personal treatment can be developed on this basis. Secondly, such research enhances the view that machine learning in genetic research can cope with the complications of genetic data.

II. Related Works

Since genetic disease prediction is an important area, there has been much research done based on this. Each study came up with different approaches, which helped in achieving various aspects of our study. Different keywords like 'Motif discovery', 'Cystic fibrosis', 'genetic disease prediction' and many more have been used to find papers to conduct a thorough review of literature on this area. This section covers different research articles in this area.

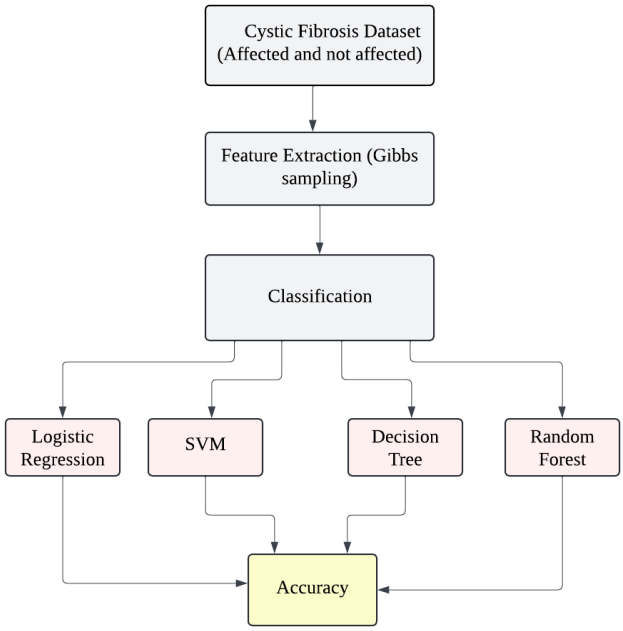

Figure 1. Block diagram of the proposed approach

Traditional studies have looked at certain mutations, but recent advances have made more comprehensive genomic studies identifying patterns associated with the disease. Genetic studies on the etiology of cystic fibrosis have managed to identify many variations in the CFTR gene that cause the disease. Turcios et al. [7] provided an extensive study of the CFTR gene, detailing the spectrum of mutations and their relevance to CF pathology. Extended the understanding of CF by investigating the genotype in CF patients. Their study showed the complexity of CF genetic factors and the use of state-of-the-art methods to establish the underlying genetic

structures. This research lays the framework for using advanced bioinformatics research. Mayara S., et al. [8], collected data of CF patients who are using wearable sensors to record their physical activities. Then different ML algorithms are used such as random forest, decision tree, logistic regression and all. Random Forest outperforms the other models having highest accuracy. From all the related work papers, it's evident that random forest works better for most of the disease prediction using classifiers which is related to current genetic disease research using ML Models.

The identification of DNA Motifs [9] and the application of Machine learning techniques [10] have been studied in the field of bioinformatics. The major findings and approaches in previous research are described in the following section.

Researches in the field of motif discovery [11] have significantly helped to understand gene regulation and expression. Gibbs Sampling and MEME (Multiple Em for Motif Elicitation) have been used for this purpose. MEME, Timothy L., et al. [12], which employs a statistical method based on the expectation maximization approach for motif finding. Gibbs sampling is an iterative process with high efficiency, identifying motifs using a sample from the conditional probability distribution of motif sites. The ability of Gibbs sampling to identify protein binding regions in DNA sequences was demonstrated by Changjun, et al. [13]. The approach has been used successfully in identifying motifs in complicated genomic data due to its superior accuracy and ability to process large volumes of data. The ability of the approach to trace weak patterns in noisy data has been exploited successfully in the interpretation of regulatory processes in several organisms.

Machine learning has made it possible to predict and classify genetic traits based on sequence information. Machine learning is used to develop models. Decision Tree, SVM, Logistic Regression and Random Forests are some of the models extensively used in the context of this paper. All of these models have a special advantage when dealing with genomic data; for instance, the resistance of Random Forest to over-fitting due to its ensemble approach, while SVM excels in high-dimensional spaces. In the review of the use of machine learning techniques in genomics, Shahadat et al.

[14] noted the effectiveness of these models in challenges like sequence classification, gene expression prediction, and motif finding. The significance of feature extraction techniques, such as k-mer counting, which convert raw sequence data into a format suitable for machine learning algorithms.

Sujatha et al. [15] have done research in prediction of heart disease via performance of supervised machine learning models. In this paper they have taken a heart disease dataset with different attributes such as age, cholesterol level, sex, and all. Various classifiers are used such as SVM, Random Forest, Naive Bayes, decision tree and all are evaluated using the performance metrics such as precision, recall, accuracy. Here random forest emerged as the best performing algorithm.

Deepa, R., et al. [16], in this paper different supervised ML algorithms are used to classify breast cancer. They have taken the Wisconsin Breast Cancer Dataset (WBCD) that contains images of breast mass. The models evaluated are decision tree, random forest, SVM, Naive Bayes and all. Random forest and SVM are found to be the better models that help in breast cancer diagnosis on the basis of evaluating accuracy.

II. METHODOLOGY

This section details the methodology followed for prediction of cystic fibrosis through Motif discovery via ML Model building. The proposed approach block diagram is shown in Figure 1 . The overall flow is expounded upon in subsections.

A. Dataset Extraction and Preprocessing

The study had utilized two datasets obtainable from NCBI, one that contains DNA sequences with CF-affected cases, and the other holds normal, not-affected sequences. Then these two datasets are randomly integrated into a single data frame, as shown in Figure 2, classifying every sequence either affected or non-affected. The sequences would then be cleaned and pre-processed to make it consistent and rid of any irrelevant and noisy data. This integrated dataset will be utilized for motif discovery and machine learning.

B. Motif Discovery

There are many motif discovery algorithms [17] to find the motifs and here we have taken Gibbs sampling motif discovery algorithm to identify significant motifs within the DNA sequences. We have utilized this algorithm to find motifs of both affected and not-affected sequences to extract the motifs. Then store the result in a data frame as shown in Figure 3.

Gibbs Sampling is an iterative method, which is used to find common motifs, that means the recurring patterns in a DNA sequence. The following are the main steps done in Gibbs sampling: First, Initialize a randomly selected motif from each DNA sequence. Create a background model from the set of all other sequences, after excluding one sequence from the set. Calculate the probability of each potential motif leading position for the excluded sequence using the background model. Then, based on these probabilities, select a new motif position for the excluded sequence. Using the new motifs, update the background model. Repeat sampling and update for a fixed number of iterations or until convergence is achieved. Finally, we have some motifs that are often resampled in the iterations.

III. Similarity Analysis

Similarity of DNA motifs was calculated by applying the Hamming distance metric. The analysis was done at three different levels: the Hamming distance among the motifs in the Affected group, the Hamming distance among the motifs in the not affected group, and the Hamming distance between the motifs of the Affected and not affected groups. It gives an average Hamming distance of 3.63 within the Affected group and 3.67 within the not-affected group but significantly larger i.e., 7.3, between the affected and not-affected. Results suggest a clear differentiation in motif patterns for affected and not-affected sequences, while pointing toward cystic fibrosis-specific biomarkers. This major difference in motif similarity underlines the importance of motif discovery in understanding the variations of genes associated with the disease.

IV. Converting Sequences To K-Mers

In this research, numerical features were made from DNA sequences, which were further used for analyses. The current study applied the k-mer analysis with k = 3, where DNA sequences are broken down into overlapping sub sequences of three nucleotides. Countvectorizer was used to turn these k-mers into a numerical feature matrix by counting the frequency of each k-mer in these sequences. It returned a vectorized form of the sequences, where each point raised all the possible occurrences of 3- mer combi- nations, such as AAA, AAC, AAG, and so on. This numerical feature matrix shall be used as input to the machine learning models to learn from DNA sequence data.

	Sequence	Label
0	AGACAGAGTCTAGATCTGTCATCCCAGGCTGGAGTGCAGTGGCACG...	Not Affected
1	CAAGCAATTCTCCTGCCTCA	Not Affected
2	GTGTTGCATTTTACAAGTTATTTTTTAGGAAGCATCAAACTAATTG...	Affected
3	AGAAATACTAGTGCCATAAAAAGTCAATATTTCAAATATAAAATTT...	Not Affected
4	GATCCTATATCTAAAATAAAATAAAATAAAATGAATAAATTGTGAG...	Affected

Figure 2. Concatenating of CF datasets into a single data frame.

	Sequence	Label	Motifs
0	GCCTCTTGGGAAGAACTGGATCAGGGAAGAGTACTTTGTTATCAGC...	Affected	GATCTGTGAT
1	GATTGATTGATTGATTGATT	Affected	ATTGATTGAT
2	GAGTCCCTGTCTTCACAGTGCTGACATTCCTGTAGAATATTACTCC...	Affected	CTATACCAAC
3	ATAACTTGAAAGTTTTAAAAATTATGTTTTCAAAAGCCCACTTTAG...	Not Affected	AATTATGTTT
4	GACTGGTGGTACCATCTTCTTCAGAACTGCATCTAAGAGGCTGTGC...	Not Affected	CATCTAAGAG

Figure 3. Adding Motifs of data frame

V. ML Model Training and Evaluation

We have taken four supervised machine learning models [18] for training such as Random Forest, Decision tree, SVM, Logistic Regression. Each model was trained under the training subset of the dataset. Performance was measured using the testing subset, with accuracy as the primary evaluation metric. Following are the ML Models we have used in this paper.

A. Random Forest

Random Forest is an ensemble learning method where many decision trees are created for training and then combined for better accuracy in prediction. Trees are based on random subsets of the forest training set. The ultimate prediction is obtained through the average of all the predictions from trees. Hyper-parameters used in this are n-estimators, max-depth, min-sample-split, max-features, bootstrap—these all are a few of the parameters used. n-estimators help know the number of trees in the forest. Increasing the number of trees generally helps increase accuracy.

These are hyperparameters which can be tuned around in order to optimize Random Forest Model performance on a given dataset. It has high predictive accuracy because it includes an ensemble of trees. One of the important advantages of Random Forest is that it does not overfit since it is the average of a lot of different trees. It can treat big datasets of greater dimension. It also gives feature importance, which helps so much in understanding the model. The disadvantages are that it uses more memory, the results are not interpretable as a single decision tree. Can be used in DNA classification, Random Forest is able to take on board the complexity and variability of DNA sequence data, so the algorithm is appropriate for the prediction of genetic conditions like Cystic Fibrosis.

B. Decision Tree

A decision tree is one of the supervised ML models that generates branches of the data based on the input values, which leads to a decision at each node. While constructing the tree, it recursively partitions the data until the leaves contain uniform samples. The tree structure includes nodes that represent decisions based on feature values and also includes leaves. It's a Rule-Based Model.

The major hyper parameters used in the Decision tree are criterion, splitter, max-depth, min-samples-split, max-features, and all. The main advantages are simplicity to understand, explicit visualization of the process, and it can deal with numerical and categorical data. The main disadvantage is overfitting. While decision trees are very clear and interpretable in their decision processes, with complex genetic data, they become less reliable because of their tendency to overfit. Other techniques, such as pruning, which makes a smaller tree by trimming parts that do not contribute more power to classifying instances, or ensembling, like Random Forests, combining multiple decision trees for a better decision, could handle these problems. The latter are particularly useful in DNA classification, which includes highly complex and variable data, and hence their inclusion is really essential in raising model reliability and accuracy.

C. SVM

Support Vector Machine or SVM, one of the classification models that performs the selection of the best hyperplane, which separates data points belonging to different classes while maximizing the margin between them. Therefore, this technique solves both linear and nonlinear classification problems by using different kernel functions for the purpose. The important key features are that it maximizes the margin and is based upon the use of kernel functions. The kernel, degree, gamma, coef0, and all are considered as hyper parameters. The greatest advantage is that it performs very well with high-dimensional data and can use datasets where the number of features is higher than the number of samples. Simultaneously, it helps in reducing the risk of overfitting. One major point of inconvenience is in the case of large datasets; this can be computationally intensive. Thus, SVM has been found to be effective in DNA sequence classification due to its inherent ability to handle the high-dimensional feature space created by k-mer analysis. In effect, its effectiveness in handling complicated and high-dimensional genetic data makes it very fine for distinguishing different genetic conditions like cystic fibrosis.

D. Logistic Regression

Logistic Regression is a statistical model for predicting the event probability of a binary outcome based on one or more predictor variables. In this case, there is a relationship between the input features and the outcome or target variable, which the logistic function models. Some of the major hyper-parameters used are penalty, solver, max-iter, and all. The major advantages are simplicity and its efficiency. Class membership probabilities are given, which gives insight into how confident the predictions are. They include the sensitivity to outliers and linearity assumption, and all that.

It can provide at least a very good baseline model for logistic regression in binary classification problems, such as classifying between the DNA sequences affected and not affected by cystic fibrosis. This paper tries to find the most suitable machine learning classifier model [19] from DNA sequence information for cystic fibrosis classification in terms of accuracy using the proposed method of DNA motif discovery and ML model building.

VI. RESULTS

The proposed work focuses on the importance of CF disease prediction via motif discovery and the data are analyzed using four different classifiers such as random forest, decision tree, SVM and logistic regression, in which Random Forest out performs the other classifiers with better accuracy as shown in Table 1.

TABLE I. COMPARISON OF ML MODELS

ML Models	Accuracy
Random Forest	72%
SVM	70%
Decision Tree	63%
Logistic Regression	70%

Comparing all four models estimates a better model with the highest accuracy. This feature is achieved by Random Forest through averaging the predictions by many trees. In this way, overfitting of trees is managed and minimized, which may be advantageous for a complex dataset like DNA sequence data that is highly variant by nature, hence may cause non-generalized results in overfitted models. Moreover, it allows for more than a thousand input variables that are not deleted, which in the case of genetic data analysis may, for example, be useful due to possible interactions among a large number of genetic markers. The Decision tree classifier had a minimum accuracy among the rest. It is therefore a consequence that the Decision Tree model represents the most overfitted model on the training data because it tends to capture noise rather than real underlying patterns. While a single decision tree model is easily understandable, it often does not perform or generalize well in unseen datasets, the same could be possible in the high-dimensional feature space, which includes DNA sequence analysis.

VII. CONCLUSION

This proposed work focused on evaluating the motif of cystic fibrosis disease and then compare the different ML Models such as SVM, Decision tree, logistic regression and Random Forest using accuracy as key features in classifying DNA sequence. The result shows that the random forest model achieves the highest accuracy. On the journey of this work, several challenges were faced. One main challenge was, finding the dataset having DNA sequence of genetic diseases. The integration of motif discovery using Gibbs sampling and the comparison of classifiers provided a novel approach in this genetic disease area. Then the comparison of different ML Models helps to find the performance of different classifiers in this approach. Overall, this study helps for the robust understanding of this area for the upcoming research also.

VIII. LIMITATIONS

Mainly, the issue faced is limited dataset that are publicly available due to privacy issues, need more access and all to different hospitals and sites. Future Works has to consider a dataset having more patients so that researchers can try this in different models, this proposed work have also tried in deep learning models [20] referring to some papers ,but due to limited dataset, accuracy was very less. In addition, the further researches can take this as a pathway and they can look into exploration of other ensemble methods or hybrid models that can be envisioned to obtain more insights and enhancements with respect to the classification of genetic data.

REFERENCES

[1] Watson, James D., and Francis HC Crick. “The structure of DNA.” Cold Spring Harbor symposia on quantitative biology. Vol. 18. Cold Spring Harbor Laboratory Press, 1953.

[2] Naehrig, Susanne, Cho-Ming Chao, and Lutz Naehrlich. “Cystic fibrosis: diagnosis and treatment.” DeutschesA¨ rzteblattInternational 114.33-34 (2017): 564.

[3] Das, Modan K., and Ho-Kwok Dai. “A survey of DNA motif finding algorithms.” BMC bioinformatics 8 (2007): 1-13.

[4] D’haeseleer, Patrik. “What are DNA sequence motifs?” Nature biotechnology 24.4 (2006): 423-425.

[5] Sujatha, P., and K. Mahalakshmi. “Performance evaluation of supervised machine learning algorithms in prediction of heart disease.” 2020 IEEE international conference for innovation in technology (INOCON). IEEE, 2020.

[6] Singh, Amanpreet, Narina Thakur, and Aakanksha Sharma. “A review of supervised machine learning algorithms.” 2016 3rd international conference on computing for sustainable global development (INDIACom). Ieee, 2016.

[7] Turcios, Nelson L. “Cystic fibrosis lung disease: an overview.” Respiratory care 65.2 (2020): 233-251.

[8] Bianchim, Mayara S., et al. “A Machine Learning Approach for Physical Activity Recognition in Cystic Fibrosis.” Measurement in Physical Education and Exercise Science 28.2 (2024): 172-181.

[9] Lee, Nung Kion, Xi Li, and Dianhui Wang. “A comprehensive survey on genetic algorithms for DNA motif prediction.” Information Sciences 466 (2018): 25-43.
[10] Ibrahim, Ibrahim, and Adnan Abdulazeez. “The role of machine learning algorithms for diagnosing diseases.” Journal of Applied Science and Technology Trends 2.01 (2021): 10-19.
[11] Mohanty, Satarupa, et al. “A Review on Planted (l, d) Motif Discovery Algorithms for Medical Diagnose.” Sensors 22.3 (2022): 1204.
[12] Bailey, Timothy L., et al. “MEME: discovering and analyzing DNA and protein sequence motifs.” Nucleic acids research 34. suppl2 (2006): W369-W373.
[13] Zhou, Changjun, et al. “An improved genetic algorithm for DNA motif discovery with Gibbs sampling algorithm.” Journal of Bio nanoscience 8.3 (2014): 219-225.
[14] Uddin, Shahadat, et al. “Comparing different supervised machine learning algorithms for disease prediction.” BMC medical informatics and decision making 19.1 (2019): 1-16.
[15] Sujatha, P., and K. Mahalakshmi. “Performance evaluation of supervised machine learning algorithms in prediction of heart disease.” 2020 IEEE international conference for innovation in technology (INOCON). IEEE, 2020.
[16] Deepa, R., et al. “Breast Cancer Classification using the Supervised Learning Algorithms.” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021.
[17] Hashim, Fatma A., Mai S. Mabrouk, and Walid Al-Atabany. “Review of different sequence motif finding algorithms.” Avicenna journal of medical biotechnology 11.2 (2019): 130
[18] Uddin, Shahadat, et al. “Comparing different supervised machine learning algorithms for disease prediction.” BMC medical informatics and decision making 19.1 (2019): 1-16
[19] Osisanwo, F. Y., et al. “Supervised machine learning algorithms: classification and comparison.” International Journal of Computer Trends and Technology (IJCTT) 48.3 (2017): 128-138.
[20] Wu, Qianfan, et al. “Deep learning methods for predicting disease status using genomic data.” Journal of biometrics and biostatistics 9.5 (2018).