


An Approach to DNA Sequence Classification Through Machine Learning: DNA Sequencing, K Mer Counting, Thresholding, Sequence Analysis


Sapna Juneja, KIET Group of Institutions, India*

Annu Dhankhar, KIET Group of Institutions, India

Abhinav Juneja, KIET Group of Institutions, India

 <https://orcid.org/0000-0003-1984-0125>

Shivani Bali, Jaipuria Institute of Management, Noida, India

 <https://orcid.org/0000-0001-5618-1857>

ABSTRACT

Machine learning (ML) has been instrumental in optimal decision making through relevant historical data, including the domain of bioinformatics. In bioinformatics classification of natural genes and the genes that are infected by disease called invalid gene is a very complex task. In order to find the applicability of a fresh protein through genomic research, DNA sequences need to be classified. The current work identifies classes of DNA sequence using machine learning algorithm. These classes are basically dependent on the sequence of nucleotides. With a fractional mutation in sequence, there is a corresponding change in the class. Each numeric instance representing a class is linked to a gene family including G protein coupled receptors, tyrosine kinase, synthase, etc. In this paper, the authors applied the classification algorithm on three types of datasets to identify which gene class they belong to. They converted sequences into substrings with a defined length. That 'k value' defines the length of substring which is one of the ways to analyze the sequence.

KEYWORDS

DNA Sequencing, K Mer Counting, Sequence Analysis, Thresholding

INTRODUCTION

1.1 Background

DNA comprises of two chains of nucleotides spiraled around each other, joined together through hydrogen bonds while moving in diverse directions. It has a double-helix structure, a spiral consisting of two DNA chains coiled around each other (Chou & Shen, 2006). Each of the chains possess four complementary nucleotides – adenine (A), cytosine (C), guanine (G) and thymine (T) (Akhtar et al., 2008),(Akhtar et al., 2008),(Akhtar et al., 2007),(Ramachandran et al., 2012) with an A on one chain always matched with T on the other, and C always matched with G (Kinsner, 2010). The structure of DNA was discovered by Francis Crick. This methodology of expressing gene in the field of biomedical sciences is employed to determine human disease structure (Kirk et al., 2018),(Phongwattana et al.,

DOI: 10.4018/IJRQEH.299963

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

2015). DNA sequencing is an operation of identifying the state of nucleotides in DNA i.e. nucleic acid sequence. It is the process of identifying the physical order of these bases. There a number of techniques for the identification of the order of four bases. Traditionally most of the biologists use Machine Learning (ML) models to resolve their problems like functional genomics, gene-phenotype associations, gene signatures and gene interactions. Previous research recognizes the genes through experimentation on realistic cells, a veracious but costly job. In contrast the present-day work uses machine based approaches to identify the genes due to inherent accuracy driven advantage of these methods. Machine approaches for gene prediction can be categorized as Content-based approaches and Similarity-based approaches (Wang et al., 2004). Similarity-based formulation search for monotony between candidate and public sequence database of existing genes. Similarity-based formulation are computationally costly and miss original genes. Content-based formulation is advanced technique of gene-prediction that overcomes limitations faced by similarity based technique. These approaches use several attribute of sequences like codon utilization, sequence length and GC content. They then employ supervised learning or applied mathematics approaches to predict whether a read comprises of any genes. ML has been evidentially effective to resolve various problem types like classification, regression and clustering.

1.2 DNA Sequencing and Machine Learning

We have employed the ML algorithm for the classification of DNA sequence. The DNA sequencing methods have an unprecedented impact on the progression of medical research and discovery in different domains including medical diagnosis, biotechnology and virology. Comparing healthy and mutated gene sequences can diagnose various diseases. Genome demonstrate several types depending on the species like human, animal, plant and microbial species. In the early 1970s, DNA sequence was traced for first time by academic researchers. Further, fluorescence-based sequencing methods came into practice which also incorporate a DNA sequencer(Olsvik et al., 1993). DNA sequencing can be very effective in finding out the sequence of individual genes. DNA sequencing technology has played a significantly front runner role in several areas of sciences including medicine, biology and forensics. A Fast Adaptive Shrinkage Thresholding Algorithm (FASTA) file represents a DNA sequence with its details and this data is useful for prediction. FASTA format is a one line affix denoted by the greater than symbol which includes annotations and next line comprises of the sequence. A typical FASTA file may include one or even many gene sequences.

In the current work, the authors have critically analyzed the effect of K value on different types of dataset. First we will check how multinomial algorithm behaves on three type of DNA sequences from our sample data sets of chimpanzee, dog and human. Further, the value of substring is bumped and it is explored that how the variation in values effect on decision accuracy. To show this accuracy we use some performance matrices. The paper is organized as follows. In Section 2, we discuss about kmer counting and related brief description of machine learning. In Section 3, we discuss about the techniques and data set. In section 4, detailed description on the implementation has been presented. Section 5 presents the result of the research which focus on the performance of machine learning models with deviation in the offset of k. Finally, Section 6 presents the conclusion of the research carried out.

2 RELATED WORKS

Mahmoud & Guo (2021) classified DNA sequences using multilayer perceptron. Using PILAE algorithm it achieves maximum 98% accuracy when applied on five different type of dataset. Kopp et al. (2020) developed a python library entitled Jangu for evaluating the performance of model and data acquisition and visualization in genomics application. This library is compatible with other deep learning libraries also. They also applied this library on the model and it was observed that the performance of model gets significantly better and considerably overcomes the iterative programming overhead. Bartoszewicz et al. (2020) developed Deep Learning Approach to

Pathogenicity Classification (DeePaC). It consider a flexible structure allowing easy evaluation of neural network architectures with reverse-complement parameter sharing Bartoszewicz et al. (2019). Guo et al.(2020) used one of MLP learning method i.e Pseudo Inverse Learning algorithm (PIL) that gives concept for neural network (NN) decay with prescient precision. Its structure has an equal numbers of hidden neurons to quantities that are to be learnt in order to overcome the learning errors. It is a feed-forward network. PIL technique is better than back propagation (BP). Anveshrithaa et al.(2019) presented approaches for the improvement of many machine learning models like support vector machine, Naive Bayes, decision tree and many more for identify promoters in DNA sequences. Dakhli & Amar(2020)proposed a new DNA sequence classifier to classify DNA chain and also used to pullout features from DNA strands. They performed classification by using dynamic time warping (DTW) method .Their model accuracy rate of 97.3% is quite appreciable when applied on three different DNA database. Liu(2018) used machine learning techniques to analyze genome structure. He proposed a BioSeq-Analysis webserver that automatically performs the attribute extraction, predictor building and performance evaluation, wherein user is only needed to upload the dataset. It is a helpful tool for natural sequence analysis. Rysak et al.(2018) came up with a formulation which chooses the least correlated physical properties for DNA classification. This classifier not only uses the sequence and static properties of DNA sequence but can also consider the dynamic properties of DNA. Yang et al.(2018)presented the machine learning models for a set of 1839 UK bacterial discriminate to classify Mycobacterium tuberculosis against eight antiTB drugs and to classify multi-drug resistance. Compared to past rules-based approach, the sensitivities of model increased by 2-4%. Rysak et al.(2018)developed the first classifier to predict bacterial activity of DNA sequence that uses not only static properties of DNA fragment but also uses dynamic properties of DNA. They got accuracy values up to 90% for all types of DNA sequences. B. Yang et al.(2017) developed BiRen that is a hybrid architecture and it is based on deep learning. It is used to target the enhanced elements on a genome-wide level using the DNA sequence alone(Upadhyay et al., 2021). BiRen shows good performance in identification accuracy, generalization to other species, robustness in subdue noise data and based on motifs or k -mers. Nguyen et al.(2016) developed their model using CNN for classifying the DNA sequences and they considered sequences as text data. This model is similar to text classification model and both are based on deep learning. They used one-hot encoding to convert the sequence into a vector and these vectors act as input to the model. By doing this they preserved the necessary position information of each nucleotide in sequences. They used C++ software package to implement their model. They evaluated their proposed model by employing 12 DNA sequences and achieved far better performance in comparison to the conventional approach. Dixit & Prajapati (2015) explained numerous techniques that are used by other researchers for DNA Sequencing problems, and explored the advantage and disadvantage of each of these(A. Juneja et al., 2020). A detailed analysis of various methodologies and their applicability in context it was deliberated that choosing the machine learning techniques is a very crucial component of any classification process. It must be appropriate in the context of the classification problem. Öz& Kaya (2013) in their work communicated that Support Vector Machines technique is a two class classifier as it provides outcomes partitions into two groups linear support vector and Non-Linear support vector. Here they developed new evaluation technique where the quality of DNA is classified into two classes as low or high(S. Juneja, Juneja, & Anand, 2019). They used SVM learning and created a confusion matrix.(Melsted & Pritchard, 2011) presented a method that determines all the k -mers in a DNA sequence data set. They used a Bloom filter that stores all the determined k -mers. With this approach they did a 50% savings in memory usage with modest costs. Table 1 shows a summary of the relevant work in recent research which provided a roadmap for the current work.

Table 1. Summary of inspiration from the earlier research in the domain of DNA sequence classifier

S.No.	Author and Year	Technique	Objective and outcome
1.	C. Suresh Gnana Dhas(2021)	CNN and hybrid model	Classify DNA sequence using probabilistic model and also compare different type of model.
2.	Mahmoud & Guo (2021)	Pseudoinverse learning autoencoder (PILAE) algorithm	Classification of DNA sequences using multilayer perceptron. Using PILAE algo, it achieves 98% accuracy when applied on five different type of dataset.
3.	Bartoszewicz et al. (2020)	DeePaC, a Deep Learning Approach	Developed a DeePaC, a Deep Learning Approach to Pathogenicity Classification.
4.	Kopp et al. (2020)	Deep learning	They developed a python library i.e Janggu for evaluating performance of model and data acquisition and visualization in genomics application.
5.	Guo et al. (2020)	MLP learning method i.e. Pseudoinverse learning algorithm (PIL)	Its structure has equal numbers of hidden neurons to quantities that are to be learned to overcome the learning errors
6.	Anveshritaa et al. (2019)	Support vector machine, Naive Bayes, decision tree	They presented the comparison of their proposed model with existing models and they conclude that their proposed model provides far better performance to identify promoter in DNA sequence.
7.	Liu(2018)	Machine learning technique to analyze genome structure	Proposed a BioSeq-Analysis webserver that automatically performs the attribute extraction, predictor building and performance evaluation for natural sequence analysis.
8.	Dakhli & Amar (2020)	Dynamic time warping (DTW) method	They proposed a new DNA sequence classifier to classify DNA chain and also used to pullout features from DNA strands and their model accuracy rate is 97.3 applied on three different DNA database.
9.	Ryasik et al. (2018)	Naive Bayes and Random Forest Machine learning	This classifier apart from using the sequence and static properties of DNA sequence, also considers the dynamic properties of DNA. Accuracy values is up to 90% for all types of sequences.
10.	Y.Yang et al. (2018)	Principal component analysis and a sparse logistic version (clustering)	Developed one machine learning model, and compared to past rules-based approaches, the sensitivities of model increased by 2-4%.
11.	Ryasik A. et al. (2018)	Variation of parameter correlation and cophenetic coefficient, Naive bayes and Random forest	They developed the first classifier to predict bacterial activity of DNA sequence. They got accuracy values up to 90% for all types of DNA sequences.

3 CLASSIFICATION BUILDING BLOCKS

3.1 Sequence Analysis

DNA, protein and RNA sequence typically represent the primary elements in the sequencing problem. In this paper, the work is related to DNA sequencing (genome sequencing).

DNA sequencing is the activity of identify the nucleotides sequence of nucleotides in a part of DNA. Some techniques are there.

1. **Sanger sequencing**. The final DNA is traced numerous times, creating piece of various lengths.
2. **Next-generation method** is new approach that enhance the speed and bring down the value of DNA sequencing. It is like executing a numerous number of small Sanger sequencing methods in parallel. Due to this parallelization, ample amount of DNA can be sequenced more rapidly and inexpensively with next-generation methods than with Sanger sequencing. Fragments of DNA up to baseborn in length are mainly sequenced using **Sanger sequencing technique** or the **chain termination method**.

Here first we describe what are the different problems occur in sequence analysis.

3.1.1 Genome Sequencing

The Genome Sequencing involves the use of some sequencing technologies for processing, management and study of the sequences. The sequencing is a complex task and has several crucial challenges including data-based design, data representation and examination of data. It is the process where the order of DNA bases is evaluated (S.H. Guo et al., 2014). The human genome is made up of over 3 billion of these DNA bases (Pareek et al., 2011). Genome sequencing has been proven in helping the scientists in identification of genes speedily and easily with high accuracy.

3.1.2 Gene Determination and Genome Annotation

Gene determination is the phenomena of tracing the introns and exons in a particular section of genome sequence. Numerous algorithms and computer programs are available for identification of protein-coding genes. Crucial feature of genome annotation is the investigation of iterative DNA (Abhinav Juneja, 2021).

3.1.3 Sequence Comparison

Comparing the sequence can be accomplished by several tools. This is a well-established process step (S. Juneja, Juneja, Anand, et al., 2019).

3.2 K-MER Counting

With Kmer counting, one can divide the string into substrings according to its 'k' value. When we have huge list of DNA sequences then its analysis using fixed size 'k vector' is easy and prompt (Sapna Juneja, Gahlan, Dhiman, et al., 2021). In DNA sequence, divide the nucleotide sequence into part of nucleotide with 'k value' ($k > 0$).

Dividing the k -mers into tiny sizes also assist to remove the difficulty of variable read lengths. All our machine learning model can work on fixed length input but by using kmer we can take input of variable length. Because in kmer we can divide our input data on the basis of k value. k -mers can also be utilized to discover genome mis-assembly by determining k -mers. In addition, k -mers are also in use to identify microorganism contamination.

3.2.1 K-mer Size

The k -mer size has various effects on the series assembly. These effects differ between small sized and higher sized k -mers. The aim of various k -mer sizes is to attained a suitable size . A lower k -mer size will store the DNA sequence with the less amount of space and if we choose smaller kmer that it also lose our Informat. Larger sized k -mers will take more memory to store the DNA sequence. Larger k -mer sizes assist to solve the problem of small recurrent regions. So we choose that type of k value that give optimum accuracy.

Analyzing DNA Sequence ATTTTCGATCG when value of k is 4:

Table 2. Analysis of DNA Sequence with k=4

Offset	0	1	2	3	4	5	6
4-mer	ATTT	TTTC	TTTCG	TCGA	CGAT	GATC	ATCG

Further, we need not to keep this value fixed on every run, we have the flexibility to change it. Jellyfish is an already developed program that is also used for the purpose of counting the k-mer (for k-mers of upto 32 bp)(Bartoszewicz et al., 2020).

3.3 Machine Learning

Machine learning (Sapna Junej et al.,2021)is the technique where we get desired output from the system using some experience and historical data. Machine learning (Abhinav Juneja et al. 2021) methodology relies on three primary process elements 1) Attribute extraction 2) Predictor building 3) Performance evaluation. Machine learning (Aggarwal s. et al. 2021)is a subfield of artificial intelligence. Machine learning algorithms basically depends on some statistical models (Hüllermeier, 2005),(Juneja et al., 2021). The data and classification algorithms are basically joined together such that the algorithm can learn from attributes and pattern of the data to make significant predictions (Cherkassky & Ma, 2009), Juneja et al. (2021). Machine learning algorithms(Khan S et al 2021) provide many benefits to public and private research areas which makes this technique so much popular and acceptable. Abundant projects exist and are devoted to developing generic libraries and toolkits for machine learning that under development for a number of languages, platforms, and use cases (Varoquaux et al., 2015), (Marks Hall, 1994), (Kohavi et al., 2002), (Dignam et al., 1983). The machine learning methods are performing fundamental roles in sequencing problems Liuet al. (2015). In BioSequence (RNA, DNA) Analysis, four normally used machine learning algorithms like support vector machine SVM and Random forest RF (Liuet al., 2016), optimized evidence-theoretic K-nearest neighbor (OET-KNN) (Chou& Shen, 2006) and covariance discriminant algorithm (Jia et al., 2016)

Deoxyribonucleic acid (DNA) is a natural molecule and used to store some information. The data of DNA sequence is growing at an explosive rate, So the work of DNA sequences is necessary in the wave of big data. Machine learning is a vigorous technique for examination of large amount of data and learns impulsive to increase knowledge. It has been extensively utilised in DNA sequence analysis and acquire a heap of research accomplishment. Machine learning models we can use for classification of data on the bases of label like email is spam or not spam. Like in human genes also we can classify it which DNA will make protein or not. So we can label these classes and classify it using multinomial naive classifier.

3.4 K-Mer's Algorithm

Here we are using Multinomial Naive Bayes Classifier to classify a DNA sequence. In this classifier we divide long string into substring that have fixed length. DNA sequences are long strings of the alphabet AGCT. Machine learning algorithms trust on fixed-length sequence for processing. First we have to fragment DNA sequence in to number of substrings on the basis of k-mer counting(Shao et al., 2022). A k-mer is a nucleotide sequence of k character in a given DNA sequence. To identify all K-mers from a given DNA sequence we need to first get k nucleotides and after that this sliding window will move with a single character. When sliding window will move to the next character that is the starting of next k-mer and so on. The default size of K is 6 but we can change it according to

our requirement. Examination of alteration in gene sequence have considerably precocious knowledge of disease cause and classification (Learn et al., 2004),(Zhang et al., 2016).

3.4.1 Importance of K-mers

Dividing a DNA sequence into fixed size substrings according to its 'k value' will help to classify the sequence. Screening of fixed size chunk is more efficient. K-mer counting is useful in other applications also like set dealing, string matching, sequence matching etc. K-mer technique is also helpful to predict that this random sequence S1 is belong to an organism O1 and O2, to solve this problem first take the genome sequence of O1 and O2 that are known so after that if S1 has more k-mers available in O1 or O2. Kmer counting directly analyzes the counts of sequence bases of length k between instances. These k -mer based formulations have been center to the field of genomics, where they are utilized to detect unique k -mers to classify sequences (Ounit et al., 2015),(F.P. Breitwieser, 2002).

3.4.2 Algorithm: K-mer Counting

For DNA sequence classification we used Multinomial algorithm. From any DNA sequence string, we can extract all possible overlapping string with some length given by K. In python we can do this thing using getKmer method. After that we need to transfer our training data into some short sequence with k-mers of length 6. At this value of k we got best accuracy. We used scikit-learn processing tools to do the counting, Now we need to change the lists of k-mers into series of words. After that count vectorizer will use that word. We also need one more variable to grasp the class labels. So at this point we know how to change our nucleotides sequences into fixed length, After we can apply any machine learning model that can classify our DNA sequence. For classification purpose we used multinomial naive bayes classifier (Sapna Juneja, Jain, Suneja, et al., 2021).

We implemented this algorithm using python. The following sequence of steps was executed for the classification.

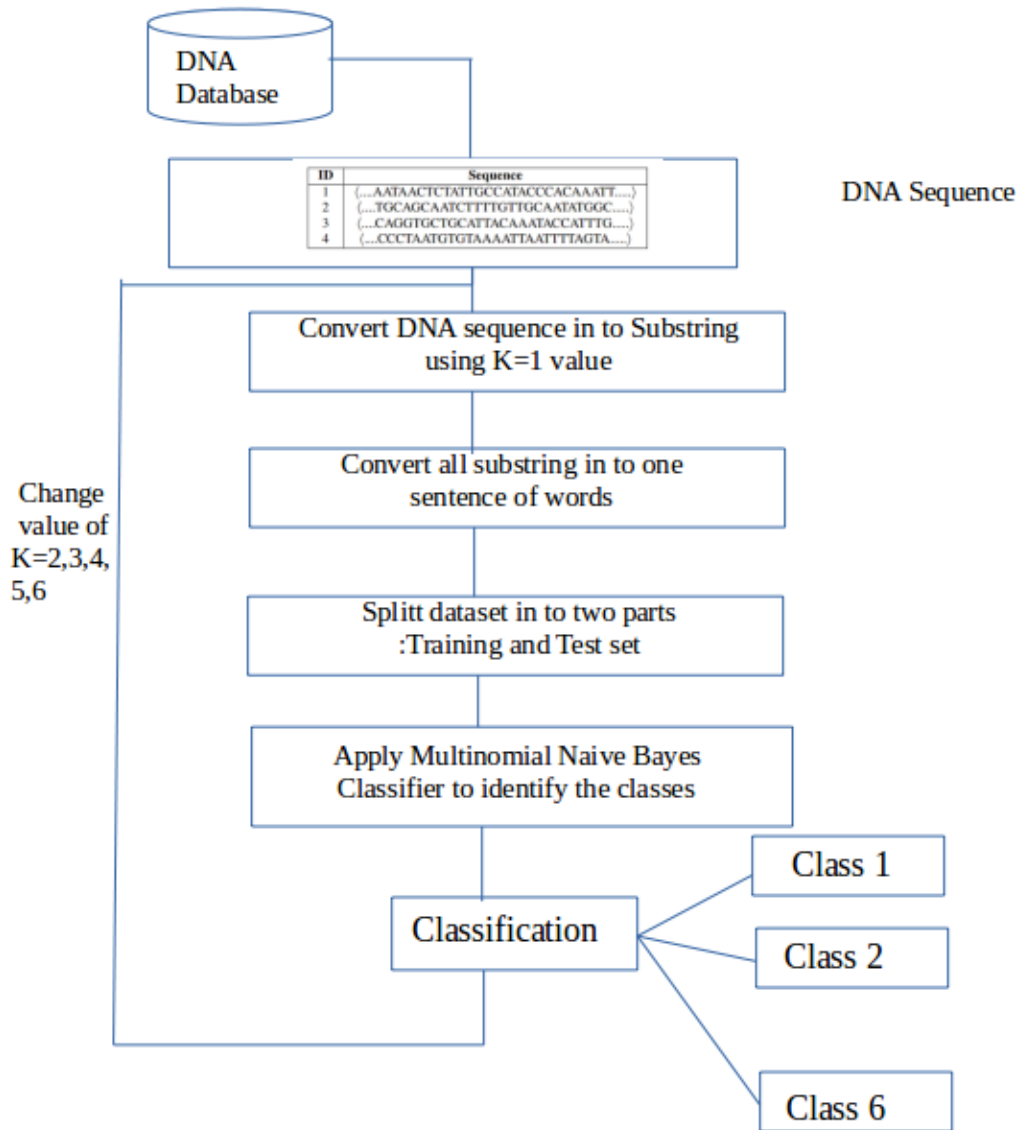
The processing is done in different phases as described below:

- Phase 1: First convert DNA sequence strings into k-mer words, default size = 6. If we choose size=4 then it breaks the whole sequence of nucleotides into 4 nucleotides. First four characters are taken, next time it skips the first character and takes next 4 nucleotides.
- Phase 2: After that we need to convert all substrings in list into one sentence of words so that we can handle it easily with one variable.
- Phase 3: Splitting the dataset into two parts i.e. training set and test set. Here we divide into 80:20 ratio means 80% data we use as a training data and 20% as a testing data.
- Phase 4: Now apply Multinomial Naive Bayes Classifier to identify the classes. Using grid search we can fix the alpha value.
- Phase 5: Now check whether kmer counting is working on gene sequence or not, checked through confusion matrix.

3.4.3 Multinomial Naive Bayes Classifier

There are many softwares or tools for the investigation of numerical data but there are less softwares are available for texts. For the analysis of the categorical text data, the most popular supervised learning classifications technique is available that is Multinomial Naive Bayes. Classification of text data is gaining popularity because there is an tremendous amount of data present in websites, documents etc. that needs to be analyzed. Multinomial Naive Bayes formula is based on the Bayes theorem that used probabilistic method. Naive Bayes classifier is a grouping of numerous formula where all the algorithmic program share one joint principle, that is each feature is independent means its not related to other features. The existence of one attribute does not impact the attribute of other attribute. This classifier used in many domain area like news categorization where they categorize news into various

Figure 1. Proposed Approach



sections such as fake,non fake, political etc. Multinomial Naive Bayes classifier is so popular because of its swift learning rate and simple design. In text or DNA categorization this classifier giving good accuracy rate because of their powerful naive hypothesis.

3.5 Datasets

We took dataset of gene sequence from the open DNA sequence dataset available at Kaggle. We used three text files of DNA sequence i.e chimpanzee text file, Dog text file and human text file. In these text files we have DNA sequences with their respective classes. The description of the dataset used has been given in Table 3.

Table 3. Description of each database

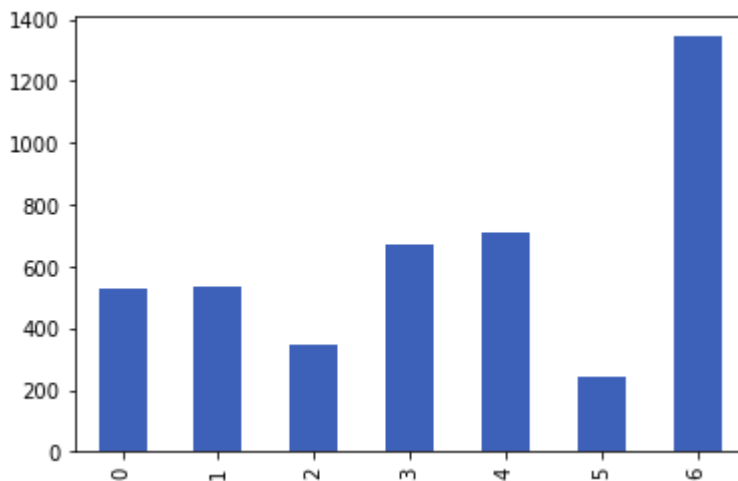
Item	Chimpanzee	Dog	Human
Number of sequences	1682	820	4380
Number of Attribute	2	2	2
Training	1346	656	3504
Test	336	164	876

DNA sequence data is very crucial for scientists research to identify the role of genes. We have to divergent species data set to check our model. The machine learning model that we applied on human data, will that work on other species also. We have human DNA sequence data coding regions with class label. We also have dog data and a more divergent species, the chimpanzee.

4. IMPLEMENTATION

This is the main part of paper, here we apply the Multinomial algorithm on a variety of datasets i.e. Chimpanzee dataset and Dog dataset and human Dataset, and evaluate the effect of k value on these type of datasets. We used python to implement that. First we apply multinomial algorithm on chimpanzee dataset to classify the gene sequence into its classes as a classifier. Figure 2 illustrates different types of class available in this data set.

Figure 2. Types of classes in chimpanzee dataset



With this chimpanzee dataset, we now evaluated the effect on the performance after changing the value of K. Now it is observed that the accuracy is bumped up when the value of k is increased. Further if we keep on increasing the value of k, the performance will decrease.

This concludes that we need to evaluate all the performance measurement parameters like accuracy, precision, recall and F1 to get the optimum results. Table 4 shows the performance of Chimpanzee dataset for various values of k.

Table 4. Performance of Chimpanzee Dataset

Chimpanzee Dataset	Accuracy	Precision	Recall	F1
K=3	62.9	71.6	62.9	64.6
K=4	86.9	88.0	86.9	86.9
K=5	91.1	92.5	91.1	90.9
K=6	91.4	92.0	91.4	91.1

Next is to apply this algorithm on Dog sequence dataset. This algorithm is also used to classify the gene sequence in to different classes. Further, we will monitor what are the effects of ‘k value’ on its performance to evaluate the classes of gene sequence (Table 5).

Lastly, we applied the algorithm on human data set; here we are showing different types of classes in dataset in front of their DNA sequence.

Next step is evaluation of the performance of DNA sequence classifier. Earlier we applied similar algorithm on animal DNA sequence to identify its gene family using DNS sequence classifier. Now with this Human dataset, we are looking to apply same algorithm to evaluate the effect on the performance after changing the value of K. Under this dataset as well we also observe that some classifier good accuracy scores with the value of k that was defined. Moreover ‘k value’ should be chosen carefully as increasing this value of k does not warranty better accuracy. The said recommendation is used to evaluate the performance as depicted in Table 6

Table 5. Performance of Dog dataset

Dog Dataset	Accuracy	Precision	Recall	F1
K=3	51.2	57.9	51.2	52.8
K=4	65.2	71.2	65.2	64.9
K=5	64.0	73.1	64.0	61.6
K=6	69.5	78.5	69.5	67.8

Figure 3. Different types of classes in human dataset

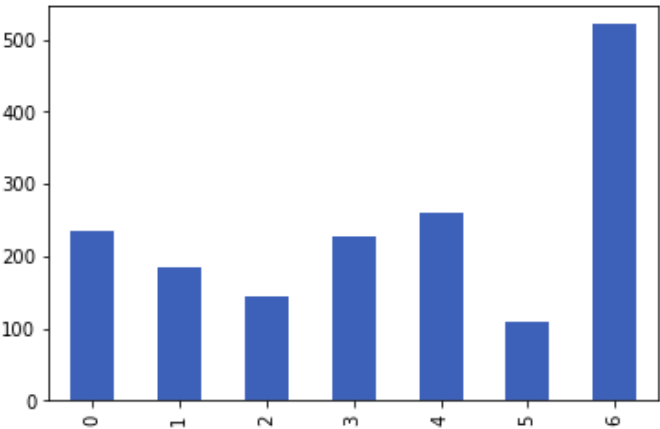


Table 6. Performance of Human Dataset

Human Dataset	Accuracy	Precision	Recall	F1
K=3	62.2	71.3	62.2	64.6
K=4	88.4	88.8	88.4	88.4
K=5	95.5	95.8	95.5	95.6
K=6	98.4	98.4	98.4	98.4

5. RESULT

This paper utilized three varied datasets of Chimpanzee, Dog and Human to determine their gene family. Studying any gene sequence is also fruitful to distinguish whether the gene sequence is original or a mutant. Each of these datasets was taken from open source. In the current work, we have evaluated how the performance of classifier improvises with changes in the values of k. 'k value' is basically used to regulate the length of the substring in a given gene sequence. At certain level when we bump the value of k the value of accuracy, precision, recall and F1 score get increases. But after certain threshold value the performance gets decrease because overhead to maintain so many substring is also get increases. The confusion matrix is a table that gives idea about the performance of our model (Fawcett, 2006), the confusion matrix gives a representation of all True- Positives, True-Negatives, False-Positives and False-Negatives as predicted by the model. Accuracy is the most spontaneous criterion of performance and it is a relation between the precisely predicted observation and all observations.

6. CONCLUSION AND FUTURE SCOPE

We conclude that progression in machine learning technique can be one of the outstanding resolution in solving classification problems in distinguishing irregular genes that distorting gene sequences. The paper shows a machine learning algorithm to classify gene sequence and that algorithm applied on three different types of dataset. Numerous machine learning algorithms like SVM, naive bayes and logistic regression are already used to classifying normal genes from abnormal genes. According to our results, multinomial algorithm performs better when K=6 means number of nucleotides in substring is 6 and one thing we also check if we tweak that value greater than 6 at that time also performance get degrade. It results in increase in the performance adding a significant impact to the result. The Goal of this study was to focus on the importance of K value to elevate the performance of classifier and we also compared the result with previously developed classifier.

We just worked on DNA sequencing problem but there are several other issues also where we can continue our work like RNA sequencing and protein sequence. The main drawback of Machine learning (ML) is that it cannot handle expeditiously natural information in their raw form as compared with Deep learning (DL). The word "deep" in DL represents the number of layers used by the data and these layers also called hidden layers. DL networks can have as many as three hundred layers but Conventional neural networks contain only two to three layers. So, in future we will work on Deep learning architecture.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

Table 7. Comparison of proposed model with previously classifier

Authors	Basic approach and algorithms	Accuracy of Classifier	Deviations from the proposed approach
Mohammed A. B. Mahmoud et. al(2021)	Pseudoinverse learning autoencoder (PILAE) algorithm	98%	When the classes size of the classification task is large,it will not work better.
Abdesselem Dakhli et. al(2019)	Dynamic time warping (DTW) method	97.3%	Size of DNA effect the execution time.
Artem Ryasik et. al(2018)	Naive bayes and Random forest	90%	Work only on three classes
C. Suresh Gnana Dhas et. al (2021)	CNN and Hybrid model	93.3%	Comparison of proposed with existing model
Proposed model	Multinomial naive bayes algorithm	98.4%	Its implemented in python and also works on variable length of gene sequence.

REFERENCES

- Akhtar, M., Epps, J., & Ambikairajah, E. (2007). On DNA numerical representations for period-3 based exon prediction. *GENSIPS'07 - 5th IEEE International Workshop on Genomic Signal Processing and Statistics*, 2. doi:10.1109/GENSIPS.2007.4365821
- Akhtar, M., Epps, J., & Ambikairajah, E. (2008). Signal processing in sequence analysis: Advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*, 2(3), 310–321. doi:10.1109/JSTSP.2008.923854
- Anveshrihaa, S., Aathavan, B., & Jaisankar, N. (2019). Promoter prediction in DNA sequences of escherichia coli using machine learning algorithms. *International Journal of Scientific and Technology Research*, 8(11), 3000–3004.
- Bartoszewicz J.M., Seidel, A., Rentzsch, R., Renard, B. Y., (2020). DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, 36(1):81-89. doi: 10.1093/bioinformatics/btz541. PMID: 31298694.
- Bartoszewicz, J. M., Seidel, A., Rentzsch, R., & Renard, B. Y. (2020). DeePaC: Predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics (Oxford, England)*, 36(1), 81–89. doi:10.1093/bioinformatics/btz541 PMID:31298694
- Cherkassky, V., & Ma, Y. (2009). Another look at statistical learning theory and regularization. *Neural Networks*, 22(7), 958–969. doi:10.1016/j.neunet.2009.04.005 PMID:19443179
- Chou, K. C., & Shen, H. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *Journal of Proteome Research*, 5(8), 1888–1897. doi:10.1021/pr060167c PMID:16889410
- Dakhli, A., & Ben Amar, C. (2020). Power spectrum and dynamic time warping for DNA sequences classification. *Evolving Systems*, 11(4), 637–646. doi:10.1007/s12530-019-09306-4
- Dignam, J. D., Martin, P. L., Shastry, B. S., & Roeder, R. G. (1983). Eukaryotic gene transcription with purified components. *Methods in Enzymology*, 101(C), 582–598. doi:10.1016/0076-6879(83)01039-3 PMID:6888276
- Dixit, P., & Prajapati, G. I. (2015). Machine learning in bioinformatics: A novel approach for DNA sequencing. *International Conference on Advanced Computing and Communication Technologies, ACCT*, 41–47. doi:10.1109/ACCT.2015.73
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. doi:10.1016/j.patrec.2005.10.010
- Guo, P., Zhao, D., Han, M., & Feng, S. (2020). Pseudoinverse Learners. *New Trend and Applications to Big Data.*, 1(April), 158–168. doi:10.1007/978-3-030-16841-4_17
- Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., Lin, H., Chen, W., & Chou, K. C. (2014). INuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics (Oxford, England)*, 30(11), 1522–1529. doi:10.1093/bioinformatics/btu083 PMID:24504871
- Hüllermeier, E. (2005). Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems*, 156(3), 387–406. doi:10.1016/j.fss.2005.05.036
- Jia, J., Zhang, L., Liu, Z., Xiao, X., & Chou, K. C. (2016). pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics (Oxford, England)*, 32(20), 3133–3141. doi:10.1093/bioinformatics/btw387 PMID:27354696
- Juneja, A., Juneja, S., Soneja, A., & Jain, S. (2021). Real time object detection using CNN based single shot detector model. *Journal of Information Technology Management*, 13(1), 62–80. doi:10.22059/jitm.2021.80025
- Juneja, A., Bajaj, S., Anand, R., & Sindhwani, N. (2020). Improvising green computing using multi-criterion decision making. *Journal of Advanced Research in Dynamical and Control Systems*, 12(3). 10.5373/JARDCS/V12SP3/20201362
- Juneja, A., Juneja, S., Bali, V., & Mahajan, S. (2021). Multi-Criterion Decision Making for Wireless Communication Technologies Adoption in IoT. *Multi-Criterion Decision Making for Wireless Communication Technologies Adoption in IoT*, 10(1), 1–15. doi:10.4018/IJSDA.2021010101
- Juneja, A., Juneja, S., Kaur, S., & Kumar, V. (2021). Predicting Diabetes Mellitus With Machine Learning Techniques Using Multi-Criteria Decision Making. *International Journal of Information Retrieval Research*, 11(2), 38–52. doi:10.4018/IJIRR.2021040103

Juneja, S., Gahlan, M., Dhiman, G., & Kautish, S. (2021). *Futuristic Cyber-Twin Architecture for 6G Technology to Support Internet of Everything*. Academic Press.

Juneja, S., Jain, S., Suneja, A., Kaur, G., Alharbi, Y., Alferaidi, A., Alharbi, A., Viriyasitavat, W., & Dhiman, G. (2021). Gender and Age Classification Enabled Blockchain Security Mechanism for Assisting Mobile Application. *Journal of the Institution of Electronics and Telecommunication Engineers*, 1–13. doi:10.1080/03772063.2021.1982418

Juneja, S., Juneja, A., & Anand, R. (2019). Reliability Modeling for Embedded System Environment compared to available Software Reliability Growth Models. *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*. doi:10.1109/ICACTM.2019.8776814

Juneja, S., Juneja, A., Anand, R., & Chawla, P. (2019). Mining aspects on the social network. *International Journal of Innovative Technology and Exploring Engineering*, 8(9). 10.35940/ijitee.I1045.0789S19

Kinsner, W. (2010). *Towards cognitive analysis of DNA*. 10.1109/COGINF.2010.5599728

Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J., & Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, 50(10), 1474–1482. doi:10.1038/s41588-018-0207-8 PMID:30224646

Kohavi, R., John, G., Long, R., Manley, D., & Pflieger, K. (2002). *MLC++: a machine learning library in C++*. 10.1109/TAI.1994.346412

Kopp, W., Monti, R., Tamburrini, A., Ohler, U., & Akalin, A. (2020). Deep learning for genomics using Janggu. *Nature Communications*, 11(1), 1–7. doi:10.1038/s41467-020-17155-y PMID:32661261

Learn, C. A., Hartzell, T. L., Wikstrand, C. J., Archer, G. E., Rich, J. N., Friedman, A. H., Friedman, H. S., Bigner, D. D., & Sampson, J. H. (2004). Resistance to Tyrosine Kinase Inhibition by Mutant Epidermal Growth Factor Receptor Variant III Contributes to the Neoplastic Phenotype of Glioblastoma Multiforme. *Clinical Cancer Research*, 10(9), 3216–3224. doi:10.1158/1078-0432.CCR-03-0521 PMID:15131063

Liu, B. (2018). BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Briefings in Bioinformatics*, 20(4), 1280–1294. doi:10.1093/bib/bbx165 PMID:29272359

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., & Chou, K. C. (2015). Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research*, 43(W1), W65–W71. doi:10.1093/nar/gkv458 PMID:25958395

Liu, B., Long, R., & Chou, K. C. (2016). IDHS-EL: Identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. *Bioinformatics (Oxford, England)*, 32(16), 2411–2418. doi:10.1093/bioinformatics/btw186 PMID:27153623

Mahmoud, M. A. B., & Guo, P. (2021). DNA sequence classification based on MLP with PILAE algorithm. *Soft Computing*, 25(5), 4003–4014. doi:10.1007/s00500-020-05429-y

Marks Hall, G. H. (1994). *WEKA: Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press.

Melsted, P., & Pritchard, J. K. (2011). Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics*, 12(1), 333. Advance online publication. doi:10.1186/1471-2105-12-333 PMID:21831268

Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R., Abapihi, B., Kubo, M., & Satou, K. (2016). DNA Sequence Classification by Convolutional Neural Network. *Journal of Biomedical Science and Engineering*, 9(5), 280–286. doi:10.4236/jbise.2016.95021

Olsvik, O., Wahlberg, J., Petterson, B., Uhlen, M., Popovic, T., Wachsmuth, I. K., & Fields, P. I. (1993). Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in *Vibrio cholerae* O1 strains. *Journal of Clinical Microbiology*, 31(1), 22–25. doi:10.1128/jcm.31.1.22-25.1993 PMID:7678018

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1), 1–13. doi:10.1186/s12864-015-1419-2 PMID:25879410

Öz, E., & Kaya, H. (2013). Support vector machines for quality control of DNA sequencing. *Journal of Inequalities and Applications*, 2013(1), 1–9. doi:10.1186/1029-242X-2013-85

- Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52(4), 413–435. doi:10.1007/s13353-011-0057-x PMID:21698376
- Phongwattana, T., Engchuan, W., & Chan, J. H. (2015). Clustering-based multi-class classification of complex disease. *Proceedings of the 2015-7th International Conference on Knowledge and Smart Technology, KST 2015*, 25–29. doi:10.1109/KST.2015.7051475
- Ramachandran, P., Lu, W. S., & Antoniou, A. (2012). Filter-based methodology for the location of hot spots in proteins and exons in DNA. *IEEE Transactions on Biomedical Engineering*, 59(6), 1598–1609. doi:10.1109/TBME.2012.2190512 PMID:22410955
- Ryasik, A., Orlov, M., Zykova, E., Ermak, T., & Sorokin, A. (2018). Bacterial promoter prediction: Selection of dynamic and static physical properties of DNA for reliable sequence classification. *Journal of Bioinformatics and Computational Biology*, 16(1), 1–16. doi:10.1142/S0219720018400036 PMID:29382253
- Shao, C., Yang, Y., Juneja, S., & GSeetharam, T. (2022). IoT data visualization for business intelligence in corporate finance. *Information Processing & Management*, 59(1), 102736. doi:10.1016/j.ipm.2021.102736
- Upadhyay, H., Juneja, S., Juneja, A., Dhiman, G., & Kautish, S. (2021). Evaluation of Ergonomics-Related Disorders in Online Education Using Fuzzy AHP. *Computational Intelligence and Neuroscience*, 2021, 1–11. doi:10.1155/2021/2214971 PMID:34616442
- Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, 19(1), 29–33. doi:10.1145/2786984.2786995
- Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics / Beijing Genomics Institute*, 2(4), 216–221. doi:10.1016/S1672-0229(04)02028-5
- Yang, B., Liu, F., Ren, C., Ouyang, Z., Xie, Z., Bo, X., & Shu, W. (2017). BiRen: Predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics (Oxford, England)*, 33(13), 1930–1936. doi:10.1093/bioinformatics/btx105 PMID:28334114
- Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., Walker, A. S., Wilson, D. J., Peto, T. E. A., Crook, D. W., Smith, E. G., Zhu, T., & Clifton, D. A. (2018). Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics (Oxford, England)*, 34(10), 1666–1671. doi:10.1093/bioinformatics/btx801 PMID:29240876
- Zhang, Z. M., Wu, J. F., Luo, Q. C., Liu, Q. F., Wu, Q. W., Ye, G. D., She, H. Q., & Li, B. A. (2016). Pygo2 activates MDR1 expression and mediates chemoresistance in breast cancer via the Wnt/ β -catenin pathway. *Oncogene*, 35(36), 4787–4797. doi:10.1038/onc.2016.10 PMID:26876203

Abhinav Juneja is working as Professor, Deptt of IT at KIET Group of Institutions, Ghaziabad. He has 20 years of teaching experience for teaching postgraduate and undergraduate engineering students. He has research interests in the field on Software Reliability, IoT, Machine Learning and soft computing. He has published several papers in reputed national and international journals. He has been involved in the role of editor of several books and also has been active in organizing various conferences.

Dr. Sapna Juneja is working as Professor in Department of Computer Science at KIET Group of Institutions, Ghaziabad, India. She has more than 17 years of teaching experience. She completed her doctorate and Masters in Computer Science and Engineering from M. D. University, Rohtak in 2018 and 2010 respectively. Her topic of research is Software Reliability of Embedded System. Her areas of Interest are Software Engineering, Computer Networks, Operating System, Database Management Systems, and Artificial Intelligence etc. She has guided several research thesis of UG and PG students in Computer Science and Engineering. She is the reviewer of several international journals of repute. She has published 6 patents. She has published various research papers in the renowned National and International Journals.