International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-11 Issue-10, September 2022

DNA Sequencing using Machine Learning and Deep Learning Algorithms

Varada Venkata Sai Dileep, Navuduru Rishitha, Rakesh Gummadi, Natarajan.P

Abstract: ***DNA Sequencing plays a vital role in the modern research. It allows a large number of multiple areas to progress, as well as genetics, meta-genetics, and phylogenetics. DNA Sequencing involves extracting and reading the strands of DNA. This research paper aims at comparing DNA Sequencing using "Machine Learning algorithms (Decision Trees, Random Forest, and Naive Bayes) and Deep Learning algorithms (Transform Learning and CNN)". The aim of our proposed system is to implement a better prediction model for DNA research and get the most accurate results out of it. The "machine learning and deep learning models" which are being considered are the most used and reputed. A prediction accuracy of the higher range in deep learning is also being used which is also the better performer in different medical domains. The proposed models include "Decision Tree, Random Forest, Naive Bayes, CNN, and Transform Learning". The Naive Bayes method gave greater accuracy of 98.00 percent in machine learning and the transform learning algorithm produced better accuracy of 94.57 percent in deep learning, respectively.***

Keywords: DNA Sequencing, Machine Learning, Random Forest, Decision Tree, Naive Bayes, Deep Learning, Transform Learning, CNN.

I. INTRODUCTION

This The primary approach of molecular biology and genome research is to implement machine learning and deep learning as the major course to find out the insights of DNA sequencing and perform different levels of research within it. To perform the DNA sequencing and some insights calculation within it, it is the costliest thing and we are implementing "deep learning and machine learning" models to predict the DNA sequencing with a real-time dataset and there is a need to combine and compare the results of the "machine learning and deep learning models". There are nearly 3 billion counts of nucleotides in the human genome and it is highly impossible to perform sequencing manually with the clinical applications it is economically also not suggestable. ML and DL approaches will be used and is useful for deep analysis of the data and plucking some information which is much useful for future implementation. DNA sequencing is a huge domain and for this domain, we need to apply the latest technologies of industry 4.0 to make it more reliable for prediction of the future. The medical domain is very huge and there are a lot of opportunities to perform ML predictions. The future prediction of the person is one of the most important factors we need to focus on. The patient may feel some medical issues that cannot be triggered in the initial stage of diagnosis and there is a situation where we need to implement the prediction models like decision trees, random forest, and some kind of DL algorithms/models to predict the insights of the patient medical condition. The medical sequencing prediction and the future of the patient prediction is the risky factor and they cannot be done perfectly through the general medical diagnosis. We need to Fig out the way to implement the technology in this factor. The targeted DNA sequence will be predicated on the category of the training assigned to the model. DL and the standard ML models can analyze the DNA data deep into the cores and measure the required things. DNA sequencing is a typical task and the deep search of the components and the features related to the patient is a critical task to achieve. The main focus will be on identifying the dataset and making it more acceptable for the machine learning and deep learning methodologies implementation. The medical stream is having a wide range of chances to implement the research and the DNA sequencing is one of the complex tasks here the dataset is consisting of the components or the features which will affect the human health condition and need to be taken care of with the system that can help to analyze the insights and also to process the sequence. In the general procedure, there are a lot of challenges in modeling and getting outcomes related to DNA sequencing. One of the major issues is time-consuming. The DNA components will be huge in size and it will take a lot of time to extract the data and perform translation. ETL (Extract, Transform, Learn) is the basic level of data mining approach, which all the applications will be running in the medical domain. The major motivation of this research implementation is to identify the accuracy of five different "machine learning and deep learning models" performance on the single dataset related to the DNA sequencing and the accuracy will be differed based on the components and the logic used to define the model. The implementation will be the same for each and every model which means the purpose of implementation will be the same to find the best accuracy but our work will be on identifying the best algorithm out of those 5 for the specific dataset. The accuracy and the best performance of an algorithm will be there based on the type of data we are using and the type of logic we are applying to the dataset. The dataset which we are using in this implementation is different from all the approaches.

Manuscript received on 28 August 2022 | Revised Manuscript received on 04 September 2022 | Manuscript Accepted on 15 September 2022 | Manuscript published on 30 September 2022.
***** Correspondence Author(s)
Varada Venkata Sai Dileep*, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: saidileep52@gmail.com
Navuduru Rishitha, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: rishitha2811@gmail.com
Rakesh Gummadi, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: rakesh.gummadi.us@gmail.com
Prof. Natarajan.P, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore (Tamil Nadu), India. Email: pnatarajan@vit.ac.in

II. LITERATURE SURVEY

Sequencing deoxyribonucleic acid (DNA) could be a large part of trendy research. It permits a mess of various areas to progress, we tend toll as together with genetics, meta-genetics, and phylogenetics. This paper studies fearless learning in a very domain wherever the sample house is of polynomial size. Since these ideas are trivially polynomial learnable, not abundant attention has been paid to them. In the past, researchers are focused on the learnability of thought categories whose sample areas are, after all (otherwise the matter would be trivial), super polynomial [1]. This article describes an algorithm program that is supposed to decode spheres in order to efficiently solve the shortcomings of maximum likelihood sequence detection (MLSD) in synthetic sequence systems. It is an analysis of the expected complexity of the algorithm and cannot be denied through simulation. This far exceeds the heuristic method. Now, the important bases of a polymer sample from a given signal are often disguised as a problem with maximum likelihood sequence detection (MLSD). A thorough search solution is prophylactic even with relatively short sequences. This article describes DNA sequencing, maximum likelihood sequence detection, and bead decoding [2]. In this paper, the version of the sequencing-via way of means of-synthesis approach and its nonidealities as a noisy switched linear gadget parameterized via way of means of the unknown deoxyribonucleic acid series, anyplace the extrude is executed via way of means of the enter take a look at series The base calling impediment is then exceptional as a boundary recognition issue: Given a really look at series and its relating loud result series, affirm the device boundaries, i.e., the DNA series that limits the danger of cryptography mistake [3]. In this paper, a model is energized that gets the fundamental close by choices of DNA cataphoretic time series. Then, the cubic centimeter sequencing algorithmic rule jumps on this model. Properties of the assessment are inspected by misdirecting each imitated and confirmed information. Current DNA sequencing assessments are heuristic in nature and unassuming in their use of applied mathematical data. During this paper, a real undeniable model of the DNA estimation is given so acclimated build the best most likelihood (ML) processor. Notwithstanding, blemished assessment of bound highlights of the significant information conceivably added to an exhibition that was beneath that on paper reachable. Recreations recommend a triple decrease in mistake rate is plausible utilized [4]. In this paper, the definition of an optimized approach for DNA collection assessment on a heterogeneous platform is prolonged with the Intel Xeon Phi. Such systems generally encompass one or modern purpose host vital processing units (CPUs) and one or more Xeon Phi devices. This paper gives background facts about the popularity of heterogeneous computing systems and regular expression matching and describes our tool for studying strategies for accelerating the DNA collection assessment. It moreover gives the experimental environment and discusses the experimental evaluation results. The artwork described in this paper is in assessment and contrasted to the fashionable related artwork [5]. The consequences of this paper propose that AI approaches give a promising way to deal with foreseeing drug obstruction in irresistible sicknesses. Explore different avenues regarding deceived ML procedures to foresee deoxyribonucleic corrosive data in light of MTB drug opposition. Thus, it was obvious that the ml strategy precisely anticipated drug obstruction with an exactness of up to close to 100% and a sub bend distance (AUC) of 1 (near). This outcome recommends that the AI approach is a promising hint in foreseeing tuberculosis. Drug obstruction. The outcomes additionally demonstrate the way that model presentation is information explicit, and boundary normalization can essentially (marginally) further develop execution [6]. The utilization of ML to streamline the synthesis of specific sequencing boards represents a promising new way to address improved detection of ctDNA mutations with this infection among patients. In AN in-silico screening, Panel 2 beat the option in policing growth determined ctDNA changes. One was created from a combination of many existing boards and the other upheld the recurrence of neoplastic changes. Techniques, for example, in this region unit are progressively to distinguish changes in cfDNA separated from heterologous disease patients at sequencing depth required to achieve the level of sensitivity required for early detection, especially at a reasonable cost is needed [7]. Allele-explicit articulation (ASE) is estimated by RNA sequencing, taking into account different articulation numbers of different alleles. Several studies have shown that ASE assumes a part in genetic sicknesses by managing penetrance or wound seriousness. Notwithstanding, in light of the fact that the appointment's clinical specialty depends on DNA groupings, it overlooks guidelines on natural peculiarities like ASE. To guarantee the advantages of ASE without any RNA sequencing, only the sacrifice of DNA mutations needs to be predicted [8]. In this review, an organic succession examination stage called BioSeq Analysis was laid out. At the expense of this platform, it is easy to establish machine learning techniques that utilize process prediction. BioSeq-Analysis can stick with its change by incorporating recently projected options and advanced machine learning algorithms therefore on win the goal of intelligent systems in bioinformatics [9]. In any case, the BioSeq-Analysis are frequently exclusively applied to the succession level investigation assignments, forestalling its applications to the build-up level examination errands, related to an astute device that is prepared to precisely produce various indicators for organic arrangement examination at every build-up level and grouping level is very wanted during this regard. To the easiest of our data, the BioSeq-Analysis2.0 is the first instrument for producing indicators for organic succession examination assignments at the build-up level. In particular, experimental results have shown that the predictors developed by BioSeq Analysis 2.0 can perform as well as, or perhaps better than, the general progressive predictors [10]. In this paper, they proposed DNA act-Ran, A Digital DNA Sequencing Engine for Ransomware Detection Using ML. DNA act-Ran utilizes the prerequisites of a computerized DNA sequencing plan and k-mer repeat vector. The proposed DNA act-Ran for ransomware tracking technology. ML calculations are actually applied to ransomware localization. A constant dataset was used to confirm the feasibility and productivity of the proposed DNA act-Ran strategy. A few benchmarks were utilized to assess the proposed DNA act-Ran.

The proposed technique diverged from a couple of existing ML computations [11]. One of the serious problems is to characterize the typical and ineffective properties contaminated by certain diseases. In genomic research, you can get used to different protein elements by classifying groups of DNA into existing classifications. In this way, distinguish those qualities and characterize them. To distinguish the contaminated qualities and the ordinary qualities with the utilization of characterization strategies here we utilize the AI procedures. This paper gives an audit on the instruments of quality grouping arrangement utilizing Machine Learning methods, which remembers a short detail for bioinformatics, writing overview and central points of contention in DNA Sequencing utilizing Machine Learning [12]. DNA (deoxyribonucleic corrosive) is quite possibly the most significant and fundamental macromolecule for all living creatures. Accordingly, DNA sequencing, which is a method for sequencing nucleic acid bases, has become an indispensable technique. In this paper, they exhibited that two-fold abandoned DNA atoms can be ordered very precisely utilizing AI techniques working on testing quantum transport information. Regular arrangement exactness for atoms that are primarily unique surpasses 99.9%. In fact, our strategy allows even a single base pair of E. coli DNA to cross and separate between them with an accuracy of 96% or higher [13].

In this paper, they investigated four normal uses of Machine Learning in DNA grouping information: DNA succession arrangement, order, bunching, and design mining, broke down and examined their relating natural application foundation and importance, and methodically summed up late years research in the field of DNA packing DM by nearby and dark examiners. Summarized in. It then presents a couple of huge issues in the field of DNA batching DM and some future investigation titles and models. Future exploration has acknowledged that more organic spaces and AI will be integrated to provide easier-to-use mining results [14]. This work utilized CNN, CNNLSTM, and CNN Bi-direction LSTM structures, and utilized Label and K-mer coding for DNA gathering requests. The model is assessed utilizing different gathering strategies. From the exploratory outcomes, CNN and CNN bidirectional LSTMs utilizing K-mer 20encoding give 93.16% and 93.13% exactness independently in the test data. In this article, we looked at three profound learning strategies: CNN, CNN, LSTM, and bidirectional CNN-LSTM utilizing name coding and K-mer coding. They found that CNN, which encodes marks, is superior to different models, however shockingly. The precision of the test is low. K-mer encoding is the most ideal way to accomplish superb test and approval precision. This dataset can't be assessed with accurate measurements. Different pointers like fit, review, awareness, and particularity ought to likewise be thought of [15].

III. MOTIVATION AND OBJECTIVE

The medical sector is having a large composition of information using which we can predict the future of a person from it. We are using “machine learning and deep learning models” to make the DNA sequencing and prediction through that sequence in less time and there is a need to help the medical practitioner with some advanced technologies using which they can apply the data on the model and try to predict some future from that data. The implementation will depend on the type of data we are using and the information of the columns or the feature we are using to perform the machine learning model. The initial state of the mechanism is to make the practitioner approach and it is time taking so we can’t depend on the accuracy. The time is another important component of the “machine learning and deep learning models” for implementing in the DNA research. The objective is to obtain a better prediction accuracy out of the prediction model and even identify insights and get the most accurate values for decision trees, random forest, Naïve Bayes algorithms, CNN, and Transform Learning. Comparison of the results of the ML algorithms with the result of DL algorithms and then identify the better performer out of the 5 distinct algorithms.

IV. CLASSIFICATION OF DNA SEQUENCING

Classification is used to predict the category of items with unknown labels based on the data obtained from the training set. These work well with discrete variable or categorical data and are the perfect solution for analyzing DNA data. Sequencing can be used to analyze the sequence similarity (structure or function) of a DNA sequence and predict other sequence categories or classes (function and relationship). Classification helps identify genes within a DNA molecule. Convolutional neural network classifiers are also a great way to classify DNA sequences. They are behind textual data issues such as Gmail spam detection and emotion classification in Grammarly. It is also ideal for extracting features from raw datasets. However, it is not possible to provide raw text as input to the CNN for feature extraction and class prediction. The input must be converted to a numerical representation before it can be input to the neural network. To do this, you can encode regular text (phrases) into numbers by creating a dictionary with words as values and specific numbers as keys. You can then encode the text based on the dictionary of words created (suitable for CNN). However, unlike regular textual data, there are no words in the DNA sequence, only one 100-character word with no spaces. To solve this problem, you can use k-mer to encode the sequence into words and use the one-hot encoding based on the k-mer (of size 6) dictionary to convert it to a number. In this way, you can use any text classification algorithm to classify your DNA. Once the DNA sequence has been numerically preprocessed, it can be inserted into the CNN model for training. The CNN architecture has a convolutional layer that extracts specific features from the input to build a feature map and a max-pooling layer that shrinks dimensions and outputs only the most important information.

V. OVERVIEW OF THE PROPSED SYSTEM

A. Architecture

The proposed architecture seems to be common in implementation but there is complexing in understanding the dataset. The dataset consists of the DNA sequence and which we cannot understand. There are a few unwanted symbols where we need to perform the pre-processing methodology. Fig 5.1 will give the architectural explanation of the proposed methodology.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

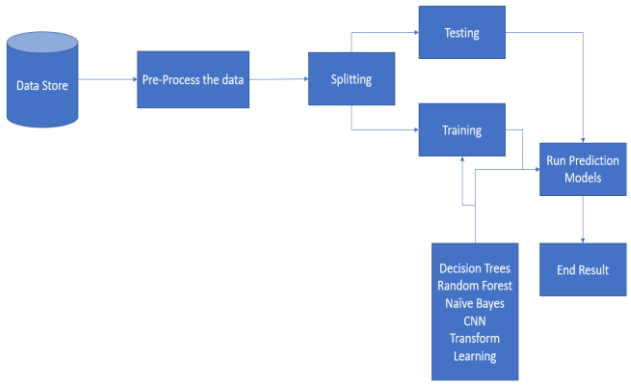

Fig. 1. Architectural view of proposed methodology

1. Data Store: Datastore is the warehouse where we have all the information in the form of raw data. The raw data is the information which is needed to be processed. The data will be collected from different resources and stored in a repository we call a warehouse. The warehouse will be having the capacity of storing any type of data and also size is also not having any restrictions.

2. Pre-Process: It is the concept of cleaning the data. The dataset will be consisting of some kind of unwanted information and what needs to be focused on for cleaning. The machine learning model will not have the capacity of handling unwanted symbols or information.

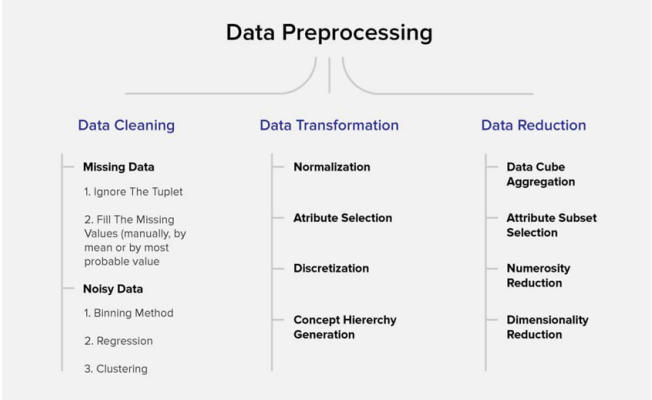

Fig. 2. Data Preprocessing

3. Splitting Data: The given dataset is then split into training and test set. The composition may be three different types. 70%-30%, 75%-25%, and 80%-20% is the composition, in general, all the researchers will use. The greater value is for the training set and the remaining small portion is for testing.

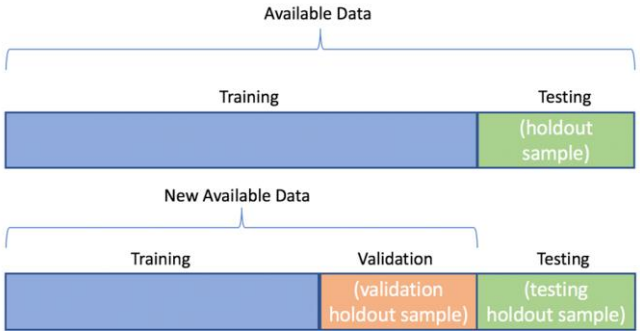

Fig. 3. Splitting Data

4. Prediction model implementation: We are implementing 5 different algorithms and technically they are called models: Decision Trees, Random Forest, Naïve Bayes, Convolutional Neural Networks and Transform Learning.

B. Implementation of the Algorithms Used in the Proposed System

Let us discuss each and every machine learning model which the proposed system stated individually to understand the operation of the models on the dataset and the objective of the problem statement solving approach.

1. Decision Tree Algorithm Implementation

In this section, we are going to look at the implementation of DNA sequencing using a decision tree algorithm. In the decision tree algorithm, we will be using the Gini method to calculate the probability of pairing the features with respect to the methodology.

In this methodology, the purity and impurity will be calculated by the Gini method. We call it Gini impurity. The Impurity will be taking care of identifying the loss happening in the dataset while the model is being created. The implementation will be identifying the accuracy of the pair of the independent variables and then we will try to reduce the Gini methodology.

The decision tree classification problem will be working on a top-down approach. There are different nodes. The first node is the root node and the middle nides which holds the data are called internal nodes and then the last nodes are called the leaf node. Each of them will be used for testing and if there is any new data it will be navigated to all the nodes based on the Yes/No condition.

The Yes/No condition will be continued until the new data is passed to any leaf node. When the decision is made, eventually predicated on the number of nodes it traveled and the probability of Yes/No traveled, precision and recall will be calculated. Using the confusion matrix, we will get the 4 variable results of the confusion matrix. As mentioned in fig 1.1 the matrix and the tree traversal will be formed.

2. Random Forest Implementation

Same as the decision tree classification algorithm, the random forest will be implemented. When we consider the random forest, it will be the group of decision tree classifiers. That means we will be plotting different random subsets of the independent variables and making them classify independently. For each decision tree, internal model accuracy will be calculated again using the Gini method.

The following are the steps for random forest using which DNA sequencing will be working on:

Step1: Identify the K number of data points to form the training set

Step2: With the selected subsets form the decision tree

Step3: Choose the decision trees to count N to build the tree

Step4: repeat 1 and 2

Step 5: For the new data point, find the prediction value of those N decision trees and assign the data point to the node or category to which it belongs to.

Fig 1.2 will explain the procedure of random forest algorithm implementation.

3. Naive Bayes Classifier

This algorithm will work on the probability of occurrence. It is based on the Bayes theorem. It is based on the reasoning. We need to understand the reason behind an action occurring.

Retrieval Number: 100.1/ijitee.J927309111022
DOI: 10.35940/ijitee.J9273.09111022
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

In this theorem, we can generate the pairs of probability and estimate the occurrence of an event. Prediction using an NB classifier is easy than other prediction models but the biggest downvote is time-consuming.
Step 1: Take the features which are used for prediction
Step 2: Categorize the variables which are considered to be predicted in numerical format
Step 3: Estimate the probability event happening for each row.
Step 4: The given dataset is then split into training and testing. After training then test the results with test set variables
Step 5: Record the probability of occurrence. Either positive or negative.
Step 6: Repeat 1 to 5
In DNA sequencing the dataset is consisting of human, dog, and chimpanzee data, and the mapping is based on the properties which are highly affecting to map a specific category. The categorical variables will function according to the requirement and the features which are mapped to the specific category are the most important factor.

4. Convolutional Neural Network

There are different steps to be followed in this procedure. First, we need to understand the important components in this dataset that are most required to prepare the neural network.
Step 1: Perform the label encoding and on-hot encoding. There are three different categories that form the encoding for those categories.
Step 2: Take the K value, which means the number of samples to train.
Step 3: we have an embedding layer, where the K value and the labels which are encoded are included for training.
Step 4: Generate the neural network with any number of hidden layers and implement pooling internally
Step 5: repeat 1 to 4

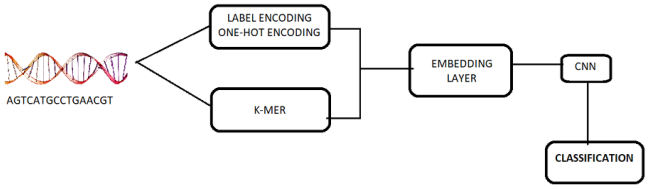

Fig. 4. DNA sequencing using CNN implementation

5. Transform Learning

In this deep transform learning methodology, we are generating deep results with a neural network that is already generated. It is a kind of grouping the results of different neural network scenarios and making them create a new unsupervised learning methodology.
Step 1: Generate the Embedded layer with K different variables and subsets
Step 2: Implement training using CNN on different subsets. Create different CNN for different subsets.
Step 3: Generate the individual results category-based.
Step 4: Get the new data input
Step 5: Repeat 1 to 3
It will be the combination of different neural networks and implement the loss function on each and every result of CNN and make a new deep neural network. The matrix created at each level will diagnose the data perfectly and map it to the right category after prediction. For suppose, we have been training a dataset or an architecture already, but we won't be having the number of labels, then what we do is, the number of labels we have will be adapted to the conditions where the final layer can be changed i.e., we will be changing the classification layer and the data we have is extended, this is called as transform learning. Actually, every dataset is different, like the human or dog, or chimpanzee dataset. Keeping aside its content size we won't be having the same type of genes. Training a dataset to our dataset, we would be changing the last layer, doing this process is known as transform learning. In Transform learning the dataset is trained easily and the classification would be effective and we can adapt to many.

VI. IMPLEMENTATION

Our project is basically about understanding how we will do DNA sequencing using ML, DL techniques and algorithms. we know that DNA in our human beings consists of a sequence type ATGC or a different kind of sequence. Here we just came across a data set in Kaggle and from that, we've basically taken up this particular project what we did is that by using DNA sequencing we will apply a classification algorithm that will be able to classify these particular sequences in human beings, for example, to what kind of gene class the particular sequence belongs to. A few libraries like NumPy, pandas, and matplotlib have been imported. The data is called human_data.txt, this data set was downloaded from Kaggle. After running this data set, we got the human DNA sequence and their class. Basically, now we have sequence and class. When we read this particular dataset, based on the sequence it should be able to predict which class the sequence belongs to. This sequence may be the gene sequence or a DNA sequence of a particular human being and they are basically classified into various classes. Apart from this particular data set we also have some data set like chimpanzee data and dog data which we got it from Kaggle and these data sets will also be having the same thing which is called a sequence and a class.

Gene family	Number	Class label
G protein coupled receptors	531	0
Tyrosine kinase	534	1
Tyrosine phosphatase	349	2
Synthetase	672	3
Synthase	711	4
Ion channel	240	5
Transcription factor	1343	6

Fig. 5. Gene Family

This is basically a gene family, if a particular sequence belongs to class 0 then it means that it belongs to the g protein-coupled receptors gene family, this number basically shows that class label 0 is present on 531 and so on, we just retrieved it and saw which class this belonged to, and this is about gene family. k-mer counting has been used. While working with DNA sequencing, we basically convert this DNA sequences as languages, and in order to convert those into languages we basically use this particular technique k-mer counting. Here we have used words of length six which is also called as hexamers. Finally, this sequence is broken down into 4 hexamer words.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In the same way our sequence will be converted into a vector by using NLP. Next, this particular data set will be converted into 6 sets of words, and each word of length 6. We have done this so that we can apply count of words or bag of words for this particular data. We have done chimpanzee data and dog data. Next, we need to convert the list of k-mers for each gene into string sentences. So now we need to combine all the sequences together because it makes it easy to convert it into a bag of words. Till here the same steps will be repeated for chimpanzee and dog data. Those steps will be done later. Next, we will try to convert the strings by applying bag of words using count vectorizer. We did this to have our independent feature in the form of strings, and in NLP we cannot use data key strings directly to our model, so for this we convert the strings into bag of words using count vectorizer. Now we check whether the data set is balanced or not, so for that I'm just trying to see the human data value counts.

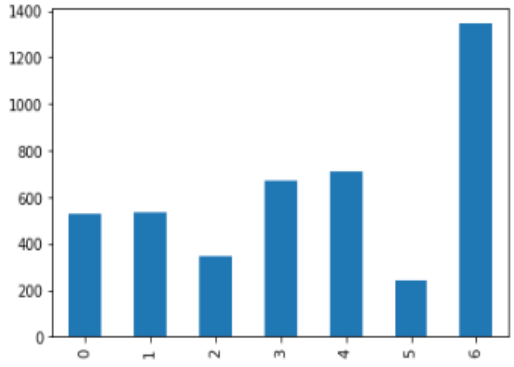

Fig. 6. Class Balance of Human

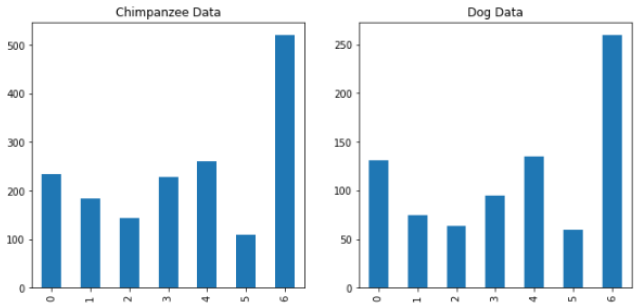

Fig. 7. Class Balance of Chimpanzee and Dog

Now we can see here it is very precisely all the classes are approximately balanced. Some of the data sets are low but other classes are approximately balanced so we can basically use this directly and with this we can also handle imbalanced dataset. If there is an imbalanced dataset problem then you can do oversampling. The Train test split and will take 20% as per the test size. Next, we applied all our classification algorithms to our three datasets to check the accuracy. These classification algorithms are estimated using different classification metrics. All these mentioned metrics are estimated from the confusion matrix. The model performs well on human data. It also does on Chimpanzee. This is not surprising as we already know that man and chimpanzees are genetically related. But the model when used for dog data did not show good results as both man and dog are not much related to each other.

VII. MODEL COMPARISON AND ANALYSIS

In this paper the proposed system is compared and analyzed with the different "machine learning and deep learning techniques" to prove the accuracy of the model. The proposed models are Decision tree, Random Forest, Naïve Bayes, CNN and Transform Learning. The Naive Bayes method gave greater accuracy of 98.00 percent in machine learning and the transform learning algorithm produced better accuracy of 94.57 percent in deep learning, respectively. The accuracy acquired in categorizing the seven distinct classes is obtained. We obtained an accuracy of 82.6 for the Decision Tree, 91.8 for the Random Forest, and 91.45 for CNN. The experimental findings have shown above clearly demonstrate that the suggested model works effectively for DNA sequence categorization. However, naive byes fared the best, with a 98.00 percent accuracy. Figure 8.1 and 8.2 depicts the comparison analysis.

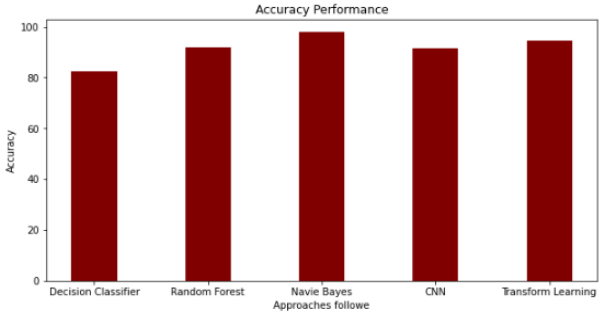

Fig. 8. Accuracy performance comparison of all algorithms

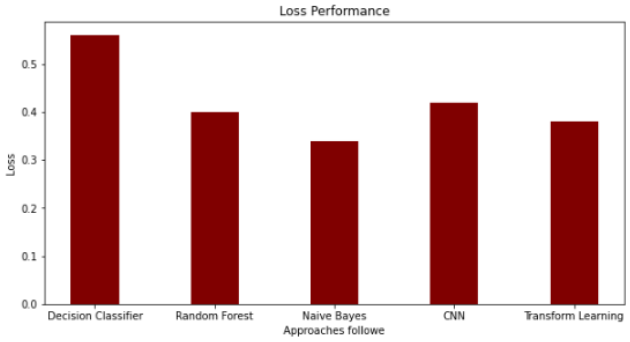

Fig. 9. Loss Performance comparison of all algorithms

Table: Accuracy Performance of all Algorithms

Algorithm Used	Accuracy Score
Decision Classifier	82.60
Random Forest	91.80
Naive Bayes	98.00
CNN	91.45
Transform Learning	94.57

VIII. RESULT AND DISCUSSION

This paper is about basically understanding how we can do DNA sequencing using Deep Learning, Machine Learning techniques and algorithms. We have taken three datasets namely human, chimpanzee and dog data set form Kaggle.

Our datasets consist of sequences and labels and each Dataset is divided into training and testing ratio of 80%, 20% for ML algorithms and 75%, 25% respectively for DL algorithms. And we have done DNA sequencing for these datasets and have applied five different classification algorithms from both machine learning and deep learning namely Decision tree, Random Forest, Naive Bayes, CNN and Transform Learning. We have used Label encoding, k-mer encoding and one-hot encoding for sequence encoding. For machine learning algorithms we have used k-mer encoding and for deep learning we have used one-hot and label encoding. For each data set in machine learning we converted this DNA sequences as languages, and in order to convert those into languages we used this k-mer counting technique and k-mer size of 6. Then we converted the list of k-mers for each gene into string sentences and also converted the strings by applying bag of words by using count vectorizer. For each data set in deep learning, we have used one-hot and label encoding. These classification algorithms are estimated using "different classification metrics like F1 score, accuracy, recall, and precision". All these mentioned metrics are estimated from the confusion matrix. The Confusion matrix of both label encoding (deep learning algorithms) and k-mer counting (machine learning algorithms) for DNA sequencing is shown in the below Figures 7.1 to 7.4.

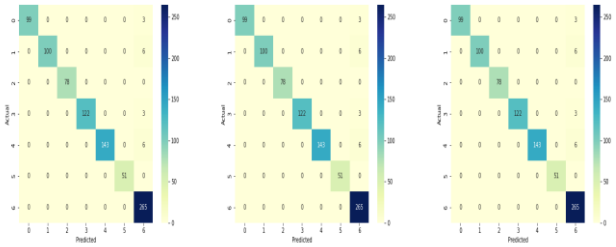

Fig. 10. Confusion matrix of k-mer encoding

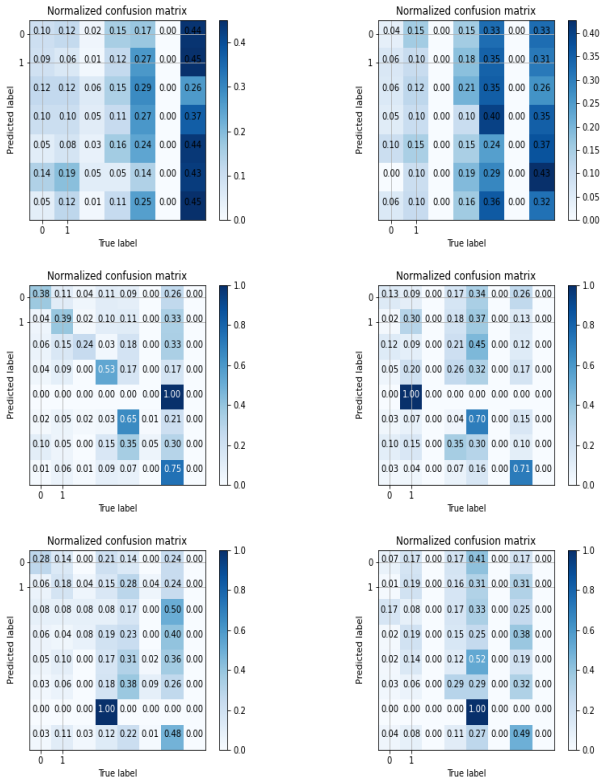

Fig. 11. Confusion matrix of Label encoding

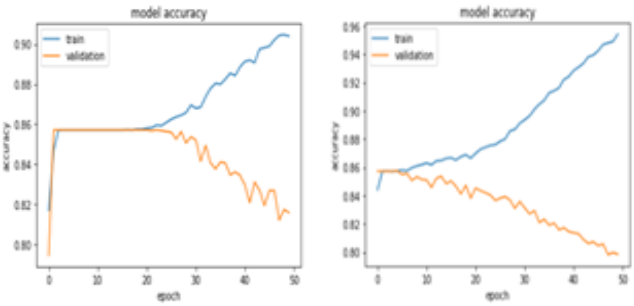

Fig. 12. Training and validation accuracy curve for Label encoding

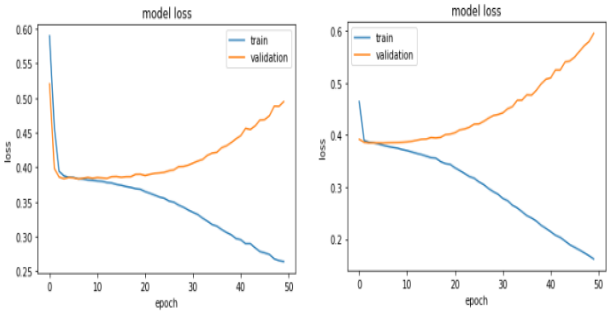

Fig. 13. Training and Validation Loss Curve for Label Encoding

IX. CONCLUSION

This paper compared five different classification algorithms from both ML and DL namely Decision tree, Random Forest, Naïve Bayes, CNN and Transform Learning with k-mer and label encoding. All these DNA sequence string encoding namely k-mer counting, One-hot Encoding and label encoding were presented, compared and analyzed. NLP bag of words algorithm by using count vectorizer was implemented for DNA sequence strings processing. We found that machine learning algorithm Naïve bayes with k-mer encoding performs better accuracy than decision tree and random forest with k-mer encoding. Transform Learning with label encoding performs better accuracy than the CNN with label encoding. However, among all the algorithms naïve byes had performed well with a highest accuracy of 98.00% than the proposed algorithm, Transform learning. Thus, when compared to label encoding, k-mer counting performed well for our datasets. This dataset is evaluated with different metrics like precision, recall and accuracy.

ACKNOWLEDGMENT

"We extend our heartfelt gratitude to Computer Science and Engineering ingeniously. Natarajan P, Prof. We express our heartfelt gratitude to Computer Science and Engineering Prof. Natarajan P, Sr. Associate Professor, School of Computer Science and Engineering, and all teaching staff and members working as limbs of our university for their non-self-centered enthusiasm coupled with timely encouragement showered on me with zeal, which prompted the acquisition of the necessary knowledge to successfully complete my course study."

REFERENCES

1. M. Li, "Towards a DNA sequencing theory (learning a string)," Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science, 1990, pp. 125-134 vol.1, doi: 10.1109/FSCS.1990.89531. [CrossRef]
2. T. Wu and H. Vikalo, "Maximum likelihood DNA sequence detection via sphere decoding," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 586-589, doi: 10.1109/ICASSP.2010.5495564. [CrossRef]
3. H. Eltoukhy and A. El Gamal, "Modeling and base-calling for Dna Sequencing-By-Synthesis," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006, pp. II-II, doi: 10.1109/ICASSP.2006.1660522. [CrossRef]
4. S. W. Davies, M. Eizenman and S. Pasupathy, "Optimal structure for automatic processing of DNA sequences," in IEEE Transactions on Biomedical Engineering, vol. 46, no. 9, pp. 1044-1056, Sept. 1999, doi: 10.1109/10.784135. [CrossRef]
5. Memeti S, Pllana S. A machine learning approach for accelerating DNA sequence analysis. The International Journal of High Performance Computing Applications. 2018; 32(3):363-379. Doi: 10.1177/1094342016654214 [CrossRef]
6. Hadikurniawati, W., Anwar, M. T., Marlina, D., & Kusumo, H. (2021, April). Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data. In Journal of Physics: Conference Series (Vol. 1869, No. 1, p. 012093). IOP Publishing. [CrossRef]
7. Cario, C. L., Chen, E., Leong, L., Emami, N. C., Lopez, K., Tenggara, I., & Witte, J. S. (2020). A machine learning approach to optimizing cell-free DNA sequencing panels: with an application to prostate cancer. BMC cancer, 20(1), 1-9. [CrossRef]
8. Zhang, Z., van Dijk, F., de Klein, N. et al. Feasibility of predicting allele specific expression from DNA sequencing using machine learning. Sci Rep 11, 10606 (2021). https://doi.org/10.1038/s41598-021-89904-y. [CrossRef]
9. Bin Liu, BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches, Briefings in Bioinformatics, Volume 20, Issue 4, July 2019, Pages 1280–1294, https://doi.org/10.1093/bib/bbx165. [CrossRef]
10. Bin Liu, Xin Gao, Hanyu Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue levelbased on machine learning approaches, Nucleic Acids Research, Volume 47, Issue 20, 18 November 2019, Page e127, https://doi.org/10.1093/nar/gkz740. [CrossRef]
11. F. Khan, C. Ncube, L. K. Ramasamy, S. Kadry and Y. Nam, "A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning," in IEEE Access, vol. 8, pp. 119710-119719, 2020, doi: 10.1109/ACCESS.2020.3003785. [CrossRef]
12. P. Dixit and G. I. Prajapati, "Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015, pp. 41-47, doi: 10.1109/ACCT.2015.73. [CrossRef]
13. Wang, Y., Alangari, M., Hihath, J. et al. A machine learning approach for accurate and real-time DNA sequence identification. BMC Genomics 22, 525 (2021). [CrossRef]
14. Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y. and Zhang, L., 2020. Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Frontiers in Bioengineering and Biotechnology, 8, p.1032. [CrossRef]
15. Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C. and Suresh Gnana Dhas, C., 2021. Analysis of DNA Sequence Classification Using CNN and Hybrid Models. Computational and Mathematical Methods in Medicine, 2021. [CrossRef]

AUTHORS PROFILE

Varada Venkata Sai Dileep, pursued his Bachelor of Technology (B. Tech) in Computer Science and Engineering from Vellore Institute of Technology, Vellore. He graduated in the month of July, 2022 and now currently working as an Advanced Application Engineering Analyst at Accenture. He has wide range of experience in machine learning, web development and completed internships in those fields.

Navuduru Rishitha, pursued her Bachelor of Technology (B. Tech) in Computer Science and Engineering from Vellore Institute of Technology, Vellore. She graduated in the month of July, 2022 and now currently working as a Java Full Stack Project Engineer at Wipro Technologies, Bangalore.

Rakesh Gummadi, pursued his Bachelor of Technology (B. Tech) in Computer Science and Engineering from Vellore Institute of Technology, Vellore. He graduated in the month of July, 2022 and now currently working as Full Stack Developer at Cognizant Technology Solutions Corp.

Natarajan P, He obtained is B. E (CSE), M. Tech (CSE & IT), Ph. D (CSE), working as an Associate Professor in Dept. of CSE at Vellore Institute of Technology, Vellore. He is having more than 18 years of teaching experience and published more than 12 papers in referred Scopus – Elsevier based journals. His total number of publications are 16. He has presented many papers in National and International Journals. He is an academic researcher from VIT University. His area of teaching interest includes Image Processing, Medical Image processing, Computer Graphics, Video Processing, Game Development. His area of research and academic interest includes: Soft Computing, Artificial Intelligence, Machine Learning, Deep Learning, Game Programming.