A Project Report on

Genome Classification

Submitted by **Group 25**

1. Venkata Sai Varun Meka (50419539)

2.Jagadeesh Raghu (50419635)

3.Ravi Raj Chenna (50414161)

4. Chaithanya Koorapati (50412902)

5. Tejaswini Oruganti (50412510)

Subject: **EAS 508 Statistical Learning 1**

10th December 2021

Table of contents

1. Abstract

Genome study plays important role in identifying genetical similarities in the off-springs and curing several diseases. Our project aims at identifying the similarities and differences between genome types of the different organisms. The genome data is made of four key proteins Adenine, Thymine, Cytosine and Guanine (ATCG) which are repeated in sequential manner. We used machine learning and Neural networking algorithms to identify the patterns in the genome which are in turn used to identify the similarities and differences. A comparative analysis on the performance measure of four main machine learning algorithms namely Multilayer preceptor classifier (MLP Classifier), Naive Bayes classifier (NB), Recurrent Neural Network (RNN) and Long Short-term Memory Model (LSTM) are done. The key attributes of these algorithm are calibrated aiming higher performance metrics.

Key Words: Genome, Neural Network, K-mer, Label Encoding, Vectorized, Multilayer preceptor classifier (MLP Classifier), Naive Bayes classifier (NB), Recurrent Neural Network (RNN) and Long Short-term Memory Model (LSTM), Bidirectional LSTM (Bi-LSTM).

2.Introduction

In current world, diseases spread through microbes, life threatening cancers, hereditary genetic disorders are higher than ever. Genome study plays crucial role in identifying the causing agent and find suitable cure for the problem. Due to innovation in genome identifying technology, the genome data availability of variety of species and microbes are explosive, but the tools to identify the inference from it are taking in part to the data. The National Center of Biotechnology Information (NCBI) has been collecting and organizing the genome data of various species for public research and reference. Added to it, the availability of vast data, surging need for cure to the diseases has added to the importance of analyzing the genome data and find the right inference. The genome study has been opened to new gateways with advancements and enhancements in machine learning, and its ability to identify the hidden pattern in large data which was previously impossible with traditional methods of computing. Genome analysis has its own difficulties as even genome of single microbes would cross the combination of billions of proteins. So, to identify the hidden patterns in combination of proteins, cutting edge machine learning techniques are used. There are several methods to classify, analyze, identify the similarities and differences among the genomes. Among the most sequential analysis-based algorithms such as Recurrent Neural Networks (RNN) and Long Short-term Neural Networks (LSTM) are used by major main players to understand the genome structure. All machine models are first splitting the data into training and testing then with the training genome data finds the general pattern and then subjected to classify the genome. It is also used to identify the similarities and differences among data sequences. Artificial Neural Networks, Hidden Markov Chains, Random Forest, Clustering techniques are also used for interpretation of Genome data.

Any Virus/Disease spreads, it will be very difficult to find the cure for that disease and there is no way to test our medical results on human body. In order to test these trail attempts, can consider

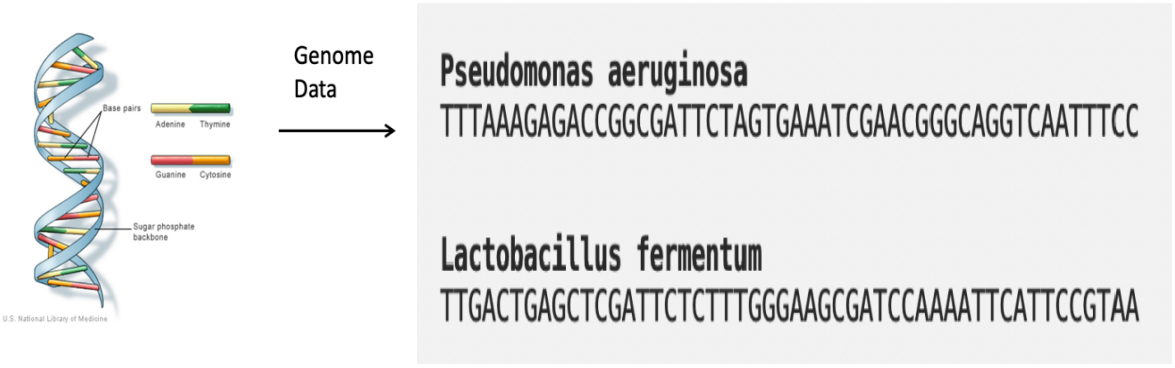
the genomes that are similar to the genomes of the human which would help to find whether the predicted medicine is a cure or not to that particular virus. In order to achieve this, we consider 3 genome sequenced data sets which are Human, Dog and Chimpanzee. The main objective is to find the which organism genome patterns [Dog, Chimpanzee] are more similar with the Human DNA.

3.Data Section

Genome DNA sequence data for **Human**, **dog,** and **Chimpanzee** as the project data set. The dataset is obtained from the **National Centre for Biotechnology Information** (NCBI) (https://www.ncbi.nlm.nih.gov)

The length of the sequence ranges from 8 to 37,971 nucleoids.

Sample Genome data:



How the data set looks like?

sequence	class
ATGCCACAGCTAGATACATCCACCTGATTTATTATAATCTTTTCAATATTTCTCACCCTCTTCATCCTATTTCAACTAAAAATTTCAAATCACTACTACCCAGAAAACCCGA	
ATGAACGAAAATCTATTCGCTTCTTTCGCTGCCCCCTCAATAATAGGTCTCCCTATTGTGGTACTGATCGTCATATTCCCTTCCATTTTATTCCCAACACCCAGTCGCCTAA	
ATGGAAACACCCTTCTACGGCGATGAGGCGCTGAGCGGCCTGGGCGGCGGCGTCAGTAGCAGTGGCGGCGGTGGTAGCTTCGCGTCCCCGGGTCGCCTGTTTCCCGGGGCGC	
ATGTGCACTAAAATGGAACAGCCCTTCTACCACGACGACTCATACGCAGCGGCGGGATACGGCCGGGCTCCGGGCGGCCTTTCTCTACACGACTACAAACTCCTGAAACCCA	
ATGAGCCGGCAGCTAAACAGAAGCCAGAACTGCTCCTTCAGTGACGTCGATGAGCTGATGAAGACGGTGCAGTTGGCTGTCCACATCCCCACCTTCCTCCTGGGCCTCCTCC	
ATGGAGGAGGGCTCCAGCTCGCCCGTGTCCCCCGTGGACAGCCTGGGCACCAGCGAGGAGGAGCTCGAGAGGCAGCCCAAGCGCTTCGGCCGGAAGCGGCGCTACAGCAAGA	
ATGACGTCCACCTGCCCCAATAACACCAGGGAGAGCAACAGCAGCCACACGTGCATGCCCCTTTCCAAAATGCCCATCAGCCTGGCTCACGGCATCATCCGCTCGAGCGTGC	
ATGGCCAACTCCACAGGGCTGACCACCTCGGAAGTCGTGGGCTCGGTGGGCTTGGTCCTGGCGGCCGTCGTGGAGGCAGCGGCCCTGCTGGGCAACGGCGCGCTGCTGGTGG	
ATGGCGAACTATAGCCATGCAGCTGACAACATTTTACAAAATCTTTCTCCTCTAACAGCCTTTCTGAAACTGACTTCCTTGGGTTTCATAATAGGAGTCAGCGTGGTGGGCA	

Class labels used in the data set :0,1,2,3,4,5,6

Below are the number of DNA Genome sequences used:

- Human – 4381 Samples
- Chimpanzee - 1683 Samples
- Dog - 821 samples

4. Methodology Section

4.1 Data Preprocessing:

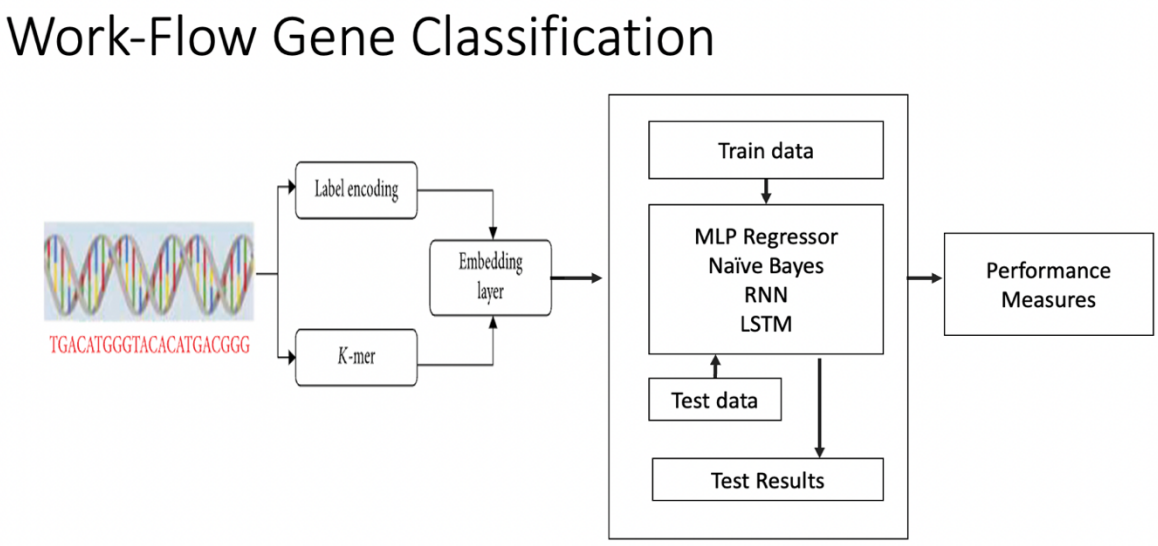
Data Preprocessing is the most critical step in machine learning and deep learning algorithms to train a model.

Steps followed in the Data Preprocessing:

1. Divided the Genome sequence with the length of n using **K-mer** encoding technique which converts the DNA sequence into an English-like statement. Each DNA sequence is splitted and transformed into a given size of substrings (Used k-mer size is 6).
2. In the next step, **vectorized** the data which converts the data into numerical which would helpful to train the model to identify the genome sequence.
3. Data set splitted into 80% train and 20% test.
4. Apply Neural Network algorithms and Machine Learning algorithms.
5. Evaluate the model using Performance Metrics (Accuracy, F1 score, Precision, Recall).

4.2 Classification Models:

Four classification models **LSTM**, **RNN (Bidirectional LSTM), Naïve Bayes** and **MLP Classifier** are used for DNA sequence classification.



4.3 LSTM:

Long short-term memory (LSTM) is a Recurrent neural network (RNN) that can learn long-term dependencies in a sequence and also introduces a memory cell to hold the past data. **Optimization algorithm** is used for training the sequences like gradient descent, combined with **backpropagation** through time to compute the gradients during the optimization process

Embedding layer, Hidden layer and dense layer are used in LSTM**.**

The current state is calculated using the below Equation

$$h_t = f\left(h_{t-1}, X_t\right),$$

where Xt is the input state, is the ht is the current state, and h(t-1) is the previous state.

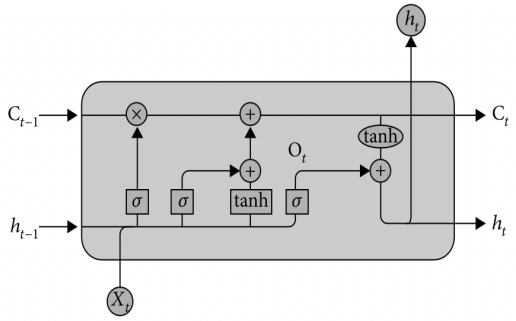


Fig: Architecture of LSTM

```
# Build neural network architecture

adam = Adam(learning_rate=0.005)

model = Sequential()
model.add(Embedding(vocab_size, 10))
model.add((LSTM(20)))
model.add(Dense(7, activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer=adam, metrics=['accuracy'])
model.summary()
```

Code snippet to build a LSTM model

Activation used in dense layer: **SoftMax**

SoftMax is an activation function that scales into probabilities. The output of a SoftMax is a vector (v) with probabilities of each possible outcome. The probabilities in vector, sums to one for all possible outcomes or classes.

$$S(y)_i = \frac{\exp(y_i)}{\sum_{j=1}^{n} \exp(y_j)}$$

Mathematical equation for soft Max

4.4 Bidirectional LSTM:

The bidirectional LSTM contains two RNN, one is used to learn the dependencies in the forward sequence and the another is to learn the dependencies in the backward sequences. Hence the output layers contain backward and forward sequences simultaneously.

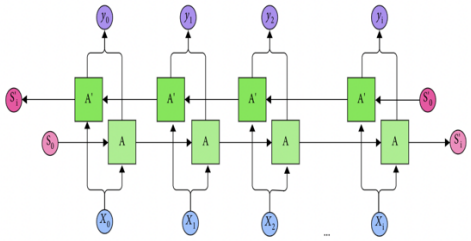


Fig: Architecture of Bidirectional LSTM

4.5 MLP and Naïve Bayes:

4.5.1 MLP Classifier (Multi-layer perceptron):

It is a feed forward Artificial Neural Network (ANN) and fully connected consists of multiple layers and each layer is fully connected to the following layer nodes. The nodes of the layers are neurons with nonlinear activation functions, except for the nodes of the input layer.

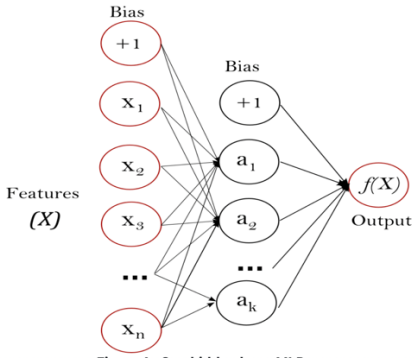


Fig: Basic architecture for MLP

```
from sklearn.neural_network import MLPClassifier
classifier = MLPClassifier(max_iter=1000,random_state=1)
classifier.fit(X_train, y_train)
```

Code Snippet for MLP Classifier

4.5.2Naïve Bayes:

It is a probabilistic classifier based on **Bayes theorem** with the "**naive**" assumption of conditional independence between every pair of features corresponding to the value of the class variable.

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \left(\prod_{j=1}^{j=n} P(x_j|\, C_i)\right) \cdot \frac{P(C_i)}{P(x_1, x_2 \ldots, x_n)}$$

Where Ci is the class variable x1,x2,…xn are the feature vectors

```
from sklearn.naive_bayes import MultinomialNB
classifier = MultinomialNB(alpha=0.1)
classifier.fit(X_train, y_train)
```

Code Snippet for NB

5.Analysis and Results:

The performance metrics used are **Accuracy, Precision, Recall** and **F1 score**.

Accuracy: The number of correctly predicted data points out of all the data points.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Precision: Ratio of number of true positives divided by the total number of positive predictions.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

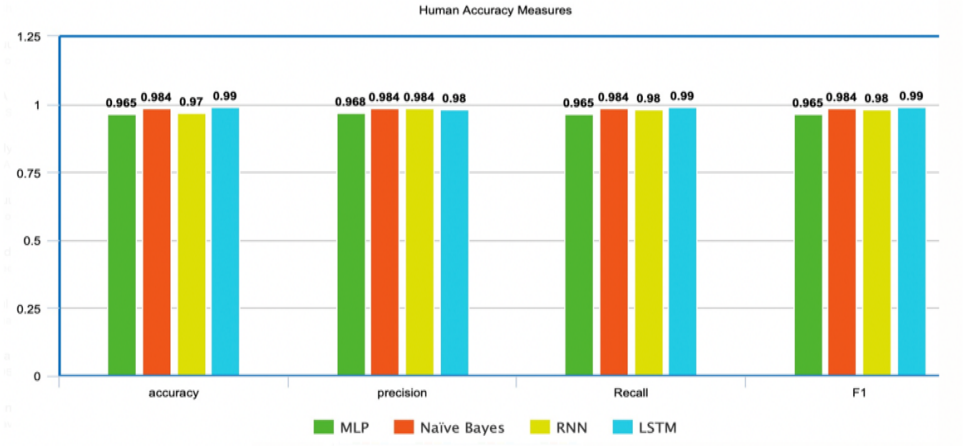
Recall: How many of the true positives were recalled (found) in the model.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

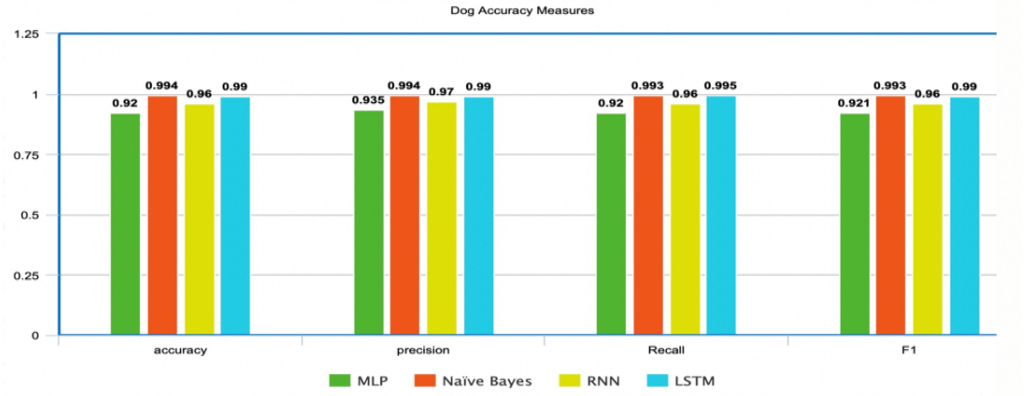
F1 Score: Harmonic mean of Precision and Recall.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

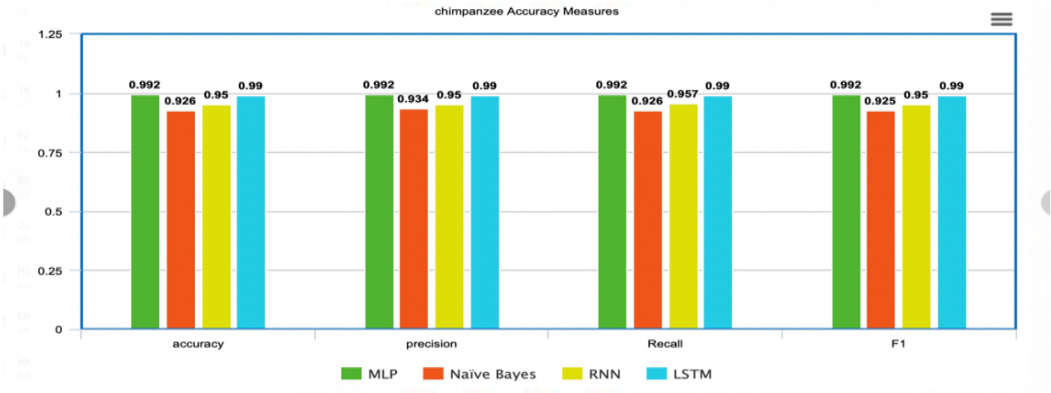
Predictions of Human



Predictions of Dog



Predictions of Chimpanzee



The above bar graph represents the comparison of the 4 metrics against the 4 classifiers. As observed, **LSTM** classifier provides the **highest** metrics. **Dog** gives similar features to that of **human** genomes.

6. Conclusion

The project mainly compared with **LSTM, RNN, MLP Classifier and Naïve bayes** models with **label** encoding and **K-mer** encoding in the terms of **Accuracy, Precision, Recall** and **F1 score**. We observed that LSTM & MLP Classifier are giving the most accurate results for the genomes data set. Both the classifiers are meant to solve a statistical classification whereas **LSTM** provides with large range of parameters such as learning rates, inputs, outputs and hidden layers as there is no need for line adjustments. However due to the memory feature of the layers in **LSTM** it performs better than **MLP Classifier**. **Dog** Genomes gives much similar to that of human genomes with more metrics than Chimpanzee.

Future Work:

Currently, considered the genomes of **chimpanzee** and **dog**. If we consider more genome data sequences for different organisms so that we can get genome's that are much similar to that of human.

As **LSTM** takes more time to compute, have to tune the parameters which would take less time to compute the metrics.

References:

- https://en.wikipedia.org/wiki/Machine_learning_in_bioinformatics
- https://www.ncbi.nlm.nih.gov/
- https://ieeexplore.ieee.org/document/7079049
- https://www.nature.com/articles/s41467-021-24497-8
- https://towardsdatascience.com/machine-learning-for-genomics-c02270a51795
- https://www.frontiersin.org/articles/10.3389/fbioe.2020.01032/full
- https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723