

Implication of ML For Disease Gene Prediction In Lung Cancer



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT, SOLAN 173234, HIMACHAL PRADESH

Name of Supervisor- Dr.Tiratha Raj Singh
DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

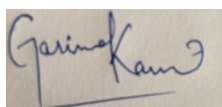
Enrollment No. : 171510,171514

Name of the Students : Garima Singh, Apurva Garg

CERTIFICATE

Candidate's Declaration

I hereby declare that the work presented in the report entitled “**Implication of ML for disease gene prediction in lung cancer**” is in fulfilment of the requirement for the final year project that is submitted in the department of Biotechnology and Bioinformatics, Waknaghat and is an authentic record of our own work carried out over a period from **24 August, 2020 to 12 May, 2021** under the supervision of our assigned guide **Dr. Tiratha Raj Singh**.



Garima Singh

171510



Apurva Garg

171514

This is to certify that the above statement made by the candidate is true to the best of my knowledge.



Dr. Tiratha Raj Singh (Supervisor)

Department of Biotechnology and Bioinformatics

Jaypee University of Information and Technology, Waknaghat

Dated on : 15.5.2021

ACKNOWLEDGEMENT

I take the opportunity to express profound sense of gratitude to my supervisor **Dr. Tiratha Raj Singh, Associate Professor , Jaypee University of Information Technology, Wagnaghat, Solan** without whom this report would not have been possible. He has been a great inspiration throughout the project period.

I extend my sincere thanks to **Dr. Tiratha Raj Singh** for giving me the opportunity to conduct research work in my area of interest. Also his continuous guidance, efforts, and invertible suggestion throughout the duration of the project was blessing in disguise.

Also, I would want to extend thanks to the Phd scholar Mr. Arvind Kumar Yadav without whose guidance this report and project wouldn't have completed. His constant help was a blessing in disguise.

Lastly I would like to thank all the researchers and scholars partially involved in the process of successful completion of my project.

Table of Content

Chapters	Content	Page Number
Chapter 1	Abstract	7
Chapter 2	Introduction	8-9
Chapter 3	Literature Review	10-15
Chapter 4	Material and Methods	16-32
Chapter 5	Result and Discussion	33-44
Chapter 6	Conclusion	45
Chapter 7	Reference	47-49

LIST OF FIGURES:

Fig 4.1 The flowchart showcases all steps followed through the project.....	16
Fig 4.1.1 The dataset of all expressed miRNA.....	17
Fig 4.1.2 The training data file with special header for iLearnPlus.....	18
Fig 4.1.3 The testing data file.....	18
Fig 4.1.4 Code for creating reads of length ranging from 18-22.....	19
Fig 4.1.5 Code for creating special header for training and testing files.....	19
Fig 4.3.1.1 Naïve bayes classifier pertaining to independent probabilities of elements of dataset.....	24
Fig 4.3.2.1 The axis between the two categories in the initial part and the steep axis which is at some definite distance reducing the covariance.....	24
Fig 4.3.4.1 All the three layers of a multi-layer perceptron.....	25
Fig 4.3.5.1 The stochastic gradient descent and the weight steps showing the maximum cost and derivative cost.....	26
Fig 4.3.6.1 The basic modelling process for XG Boost.....	27
Fig 4.3.7.1 Both the diagrams represents linear and logistic regression respectively.....	28
Fig 4.3.8.1 The KNN approach where the nearest category is assigned the respective data point.....	28
Fig 4.3.9.1 The positive and negative hyperplane resulted due to SVM algorithm.....	29
Fig 4.3.10.1 The diagram that clearly explain the decision tree and its elements....	30
Fig 4.3.11.1 The image determine the basic outlook of the random process method.....	31
Fig 5.2 Kmer ROC curve	43
Fig 5.3 Mismatch ROC curve	43
Fig 5.3 NAC ROC curve	43

Fig 5.4 NMBroto ROC curve	43
Fig 5.5 RCKmer ROC curve	43
Fig 5.6 Subsequence ROC curve	43
Fig 5.7 Z_Curve 12 bit ROC curve	44
Fig 5.8 Z_Curve 36 bit ROC curve	44
Fig 5.9 Z_Curve 48 bit ROC curve	44
Fig 5.10 Z_Curve 144 bit ROC curve	44

Lung cancer is known to be quite common in both genders (men and women) because of the unrestrained and aggressive growth of the cells within the lungs. This triggers various respiratory problems and paralysis of the chest. Cigarette smoking, exposure to radiation therapy, and encounter with carcinogens like asbestos are major contributors to the disease that includes both ADC (Adenocarcinoma) and NSCLC (Non-small cell lung cancer) leading to a high number of deaths every year. It is highly imperative to imply certain safety measures in the initial stage of the disease and in this report we have used ML techniques to predict the early start of NSCLC through some important steps. Firstly, the collection of the right dataset (i.e both positive and negative) to assess it further. Secondly, extract the feature descriptors relevant to the given data set like Kmer, Mismatch, NAC, NMBroto, etc. Thirdly, by applying various ML algorithms to check for a range of factors corresponding to each dataset. Later, the performance evaluation is done and the result of interest is discussed. This report will discuss all the descriptors and ML techniques with elaborative description and put keen emphasis on each step.

Lung cancer is the foremost cause of death in the world, with an ever-increasing number of 2 million plus cases every year and around 11.4% of the global cancer burden, according to recent data. In 2020, it is projected that there will be 228,820 new cases of lung and bronchus cancer and an estimated 135,720 people will die of this disease. Lung cancer typically affects older people with age group of 65 to 84 years old. It is rarely diagnosed before age 55. 70.4% of new lung cancer was in people of 65 and older. Although the outcomes of patients in all stages of lung cancer have enhanced in recent duration. In the case of non-small cell lung cancer (NSCLC), undergoing a surgical operation remains the only viable option due to the severity of the disease.

However, there are a lot of cases that stay unsure even after surgery. In fact, 30-55% of patients with NSCLC can still die from the disease even after proper treatment. Therefore, there is a pressing need for new biomarkers for lung cancer that can be used in clinical practice and more widened research is required to recognize and confirm these new biomarkers for predicting and also detecting lung cancer [2].

Treatment for lung cancer includes surgery, chemotherapy, radiation therapy, immunotherapy, etc. Without these treatment options, the diagnosis of lung cancer would be rather hard because the doctor will only be able to diagnose cancer in a much advanced or deadly stage. Predictability ahead of the last phase is therefore very important so that the mortality rate can be suppressed with effective and efficient control procedures. The rate of surviving this disease also varies in the patients depending on age, race, and health status. Nowadays, machine learning (ML) plays a very important role in diagnosing medical conditions in the early stages of the disease generation. ML simplifies the diagnostic process and determines factors that could lead to assess the development of cancer much earlier. Modern ML has already dominated the medical field with many districts that now use electronic learning methods in their healthcare sector. ML helps in extracting features for miRNA (microRNA) sequences consisting of nucleotides and protein sequences. ML facilitates simple analysis or examination of datasets and also inspects the valid attributes or details and aids in the identification of the underlying cause of the disease [16].

ML helps in better disease prognosis to predict the severity of disease and its effect. The need for further progress in ML algorithms will therefore assist physicians in making correct

medical decisions with effectiveness and accuracy. The correct calculation for the outcome of the disease is one of the most appealing and tough jobs for doctors. As a consequence, ML methods have become a prominent tool for healthcare researchers. These methods can discover and spot some patterns and associations from vast and complex data sets while being able to work successfully in predicting the potential effects of this type of cancer. Also, we consider the types of ML approaches utilized the kind of data they merge, the overall output of the proposed method while discussing their advantages and disadvantages [1].

The dataset that was extracted from GEO Database is transcriptomic data of miRNA. These miRNA are a family of 22-nucleotide very small RNA sequence that determines and modulates the expression of any respective gene at the post-transcriptional level. They function by combining to partially complementary sites on the gene on interest or the target gene to encourage breaking or repression of the translation by not letting the gene produce functional peptides and proteins. Despite many developments made in the understanding of miRNA and its interaction, the primary norms that dictate their interaction with the target gene is not completely understood by researchers. This miRNA dataset is the positive dataset and the negative dataset is fetched from NCBI for about 26 proteins. Now as there are both positive and negative dataset, the prediction process can begin by extracting the descriptors and diving the data into 80:20 for performing five-fold cross-validation by segregating data into training and testing files.

ML enables the system to find an explanation for a problem with some learning methodologies. The work mentioned in this report is done on CD hits and feature descriptors. After this, ML algorithms are used such as SVM, random forest classification, Multilayer perceptron, XG boost, Logistic Regression, and were final results were obtained for the analysis of data.

Lung cancer occurs when a malignant (cancerous) tumor grows inside the lungs, in structures such as the bronchi (small tubes that connect the windpipe to the inner surfaces of the lungs where gas transfer takes place). Like many other types of cancer, lung cancer is capable of multiplying and widely spreading (metastasizing) to other parts of the body. Here cancer begins in the lungs most generally spreads to the brain, bones, adrenal glands, and liver, via whichever of three mechanisms: direct extension, via the blood vessels, or the lymph system. Direct extension occurs when a tumor develops rapidly in size in such a way that it begins to contact an adjacent organ or structure and then starts to pierce itself into that adjoining organ or structure. tumor cells are also capable of getting into the blood and lymph circulatory systems and pass through, one by one, to distant structures.

Lung carcinoma is considered a deadly ailment and a major reason for death in today's world. Lung cancer affects a person to a large extent and is predicting it now ranks 7th in the mortality rate which accounts for 1.5% of the global mortality rate [4].

Some of the symptoms linked with patients such as rigorous chest pain, dry cough, shortness of breath, losing weight, etc. In terms of the development of lung cancer and the causes behind it, the doctors lay specific emphasis on smoking and second-hand smoke as the prime factors contributing to the development of lung cancer. Cancer is considered to be a complex disease made up of many different subtypes. Lung carcinoma is a painful tumor that is categorized by escalated and uncontrollable multiplication of lung tissues. The two key categories are:

1. Small-cell carcinoma (SCLC)

2. Non-small cell lung carcinoma (NSCLC)

- 3.1. NSCLC: There are three types of NSCLC tumors:

- Adenocarcinoma: It starts right in the cells in the airways that secrete mucus and other elements, usually on the exterior of the lungs. The largely widespread form of lung carcinoma in people who smoke and non-smokers and in people under the age of 45 years is that the tumor generally grows slower in comparison to other lung diseases.

- Squamous cell (epidermoid) carcinoma: This tumor begins in the cells that are on the internal layer of the lungs. There is about 1/4th of cancer that is of such a type.
- Large (undifferentiated) carcinoma: It is known to develop and expand very fast which in result makes it quite challenging to treat. It accounts for about 10% of cases.

3.2. SCLC: When lung cells begin to grow speedily in an uninhibited manner and spread in distinct ways, the condition is called small cell lung cancer.

Types of SCLC: 2 main types are small cell carcinoma also called oat cell cancer and the other is combined small cell carcinoma.

- Both above-mentioned cancers involve any kind of cells that triggered to grow and multiply in a myriad of ways and are therefore named based on the shape of the cell.

Small-lung cell cancer differs from non-small cell lung cancer in the following ways:

- Small cell lung cancer establishes itself in various parts of the body much rapidly than NSCLC.
- Small cell lung responds fine to chemotherapy (using drugs for affected cells) and radiation therapy (utilizing high-dose X-rays or other high-energy rays to curb affected cells).

3.3 Cell of Origin of NSCLC:

As we have discussed, cancer cells that begin to invade cancer may reflect structures found in normal stem cells. Increased research recently has revealed tumor-genic cells with stem cell features in lung tumors. Additional studies in mutants of but-K-Ras-induced mouse lung adenocarcinomas disclosed the existence of a rare amount of double-positive cells (DPCs) shown to signify Clara Cell Antigen 10 (CC10) cell marking; is known as Clara cell secretory protein, uteroglobin, and Secretoglobin 1a1 (Scgb1a1) and the alveolar II type, Surfactant Protein C (SFTPC), displayed that these DPCs found in BADJ were out of the usual homeostasis of the lungs, but regenerates itself and raises bronchiolar and alveolar cells post naphthalene injury. These DPCs have been showcasing continuously to produce stem cell surface markers for hematopoietic and skin cells, stem cell antigen-1 (Sca-1) and antigenation group (CD34) antigen, respectively. Ultimately, tumorigenic lesions in mutant K-Ras mutant mice revealed elevated records of DPCs, and further, continued progression of these cell groups associated with tumor progression in these mice. In addition, combined treatment of

naphthalene with K-RasG12D activation has led to a surge in the size and number of tumors [5].

3.4 Cell of Origin of Small Cell Lung Cancer:

Although significant advancement has been done in finding the definite number of cells that cause NSCLC mutates resulting in genetic mutations, it is unclear whether that A similar origin cell is responsible for tumorigenesis in SCLC. This is essentially due to the later stage of the disease in many patients at the moment of diagnosis. Nevertheless, it has been noted that at least half of SCLCs show signs of NSCLC traits, which may be contraindicated in the "normal" cell lung cancer cell in those exhibiting combined phenotypes, although it is still unclear if the same traditional cells are the determinants for initiating both types of cancer [3]. In addition, SCLCs are routinely shown to generate neuroendocrine markers and markers that have an imperative role in neuroendocrine differentiation, suggestive of that, an abnormal quantity of neuroendocrine cells could be the progenitors of SCLC. On the other hand, although small areas in mouse lungs found near neuroepithelial carcasses (NEBs) displays that it retains stem cells, the pulmonary neuroendocrine cells related with these NEBs which show weaker cell structures rather than inhibiting cells. However, seeing that SCLC can show adenocarcinoma or epidermoid carcinoma or features such as cell carcinoma such as these may contradict the existence of a “normal” cell source of this lung cancer [5].

3.5 Causes of Small Cell Lung Cancer:

- The cigarette smoking is a contributing risk factor for developing lung cancer. Those who passively intake some amount of smoke around a smoker has about a 30% increase in the risk of developing non-small cell lung cancer whereas there is about more than 55% increase in the risk of small cell cancer compared to people who are not directly around the person who smokes.
- Almost every type of lung cancer occur with rising frequency in uranium miners, but small cell lung cancer is more widespread. The pervasiveness is escalating for individuals working in Uranium mines.
- If there is a direct divulgence in any space consisting of radon gas or asbestos etc can also, harm the respiratory tract causing lung cancer.

3.6 Causes of Non-Small Cell Lung Cancer:

Doctors are not sure of the exact cause of the disease. Smoking is the most talked-about cause and especially for patients who are constant smokers or chain smokers Rest of the causes of lung cancer may be:

- Radon which is a radioactive gas existing in nature like soil and rocks
- Asbestos
- Mineral dust and iron
- Polluted air, harmful rays of radiation therapy on your chest [5] [6]

3.7 Diagnosis of Lung Cancer:

- Symptoms corresponding to lung cancer often emerge only with complex diseases. If the doctor discovers something apprehensive in the test or has some symptoms of lung cancer, additional tests will be needed to identify the condition.

3.8 Experimental Testing:

X-ray: Commonly this is the test doctors recommend initially to patients to determine any residual weight in the lungs and on encountering anything that raises concern is further dealt with additional required tests for better assistance [6].

Computed tomography (CT) scanning: This scanning test helps to evaluate the mass of the lungs and is known to be better than X-ray, as it gives the entire information related to the size or the shape and also the posture of lung tissue or to detect what organs are affected [21].

• Laboratory tests:

Sputum→ Cytology: A sputum sample (mucus that comes out of the lungs uses a microscope to spot the cells involved in cancer generation. This takes a sample of a deep cough right in the morning for 3 regular days.

Needle→ biopsy: In this process, an empty needle is inserted to obtain a little sample for testing and can also be done with aspiration biopsy where a syringe is used to remove or implant cells and fragments[5].

As lung cancer is diagnosed, the doctor will try to find out the stage (stage) of cancer. The revelation of the stage of cancer aids the patient and doctor to look for better treatment options considering the severity and complexity of the disease.

These tests include CT scan, MRI, positron emission tomography (PET). Not all tests are suitable for everyone, so talk to your doctor about what procedures are best for you. A lot of these tests are discussed above determining the process.[6]

3.8 Lung cancer can create problems, such as:

1. Shortness of breath: If the tumor develops to block the main path of the air, it can cause shortness of breath and troubled breathing. And at times due to the collection of fluid, the lungs face problems expanding completely as we breathe.
2. Coughing up blood: The growth of tumor cells and multiplication might also cause bleeding in the airways of the lungs leading to hemoptysis. As it becomes more complex this bleeding issue has to be resolved by taking some medications.
3. Pain: As cancer progresses, many organs of the body get affected and can lead to pain but there are a lot of available options and medications that the doctor can prescribe to reduce the ache and relieve the stressful area of the organ.
4. Pleural effusion: The lung carcinoma may lead to the collection of fluid in the chest and around the lung that is affected and causes breathing problems. There is some treatment to get rid of this fluid and deteriorate the risk associated with the pleural recurrence in the chest.
5. Metastasis: In this stage, cancer has erupted and escalated to all major organs of the body causing extreme inconvenience. Widespread cancer can be painful, induce nausea, headaches, or any other related signs and symptoms dictated by the cause. At this final stage, cancer cannot be cured. [5][6]

3.9 Treatment Modalities:

- Surgery: It is done mostly in the case of non-small cell lung cancer and hardly ever with small cell lung cancer when the carcinoma is just occurring in the initial phase. The surgeon ends up doing wedge resection wherein they cut a small portion of the lungs or lobectomy

where they remove a large part of the lungs or one lobe of the lung or pneumonectomy where they eliminate the whole lung.

- **Chemotherapy:** There is a mixture of certain drugs that are normally given at intervals i.e weekly, monthly, with few disruptive breaks to not overwhelm the body. It is okay to take that before a certain time of surgery or afterward with proper doctor's consultations.
- **Radiofrequency Ablation:** At times, patients are too weak to have surgery or have many other complications where surgical operations are rather difficult. So, to counter-effect the tumor, a thin needle-like structure is inserted in the lungs and through electrical energy, the cells are heated to regress the multiplication and development of cells [6]

3.10 MicroRNAs: The miRNAs are a class of small highly-conserved, non-coding RNAs that were discovered in the early 1990s, and are around 18-25 nucleotides in length. These molecules are essential post-transcriptional gene expression regulators linked to fundamental processes such as cellular proliferation, differentiation, development, and apoptosis [10] [7]. Altered miRNA levels have been described in several pathologies, like cancer, and some different studies have shown that miRNAs could be valuable as diagnostic and prognostic biomarkers in lung cancer. Furthermore, microRNAs are also involved in resistance to chemotherapy and novel targeted agents in non-small cell lung cancer. Also, rising technologies such as next generation sequencing (NGS) have shown great potential as a platform for small RNA analysis and its use is now being extended to find novel cancer biomarkers. MI classification methods have also been really helpful to predict various attributes associated with transcriptomic data [8] [9].

3.11 miRNA transcriptomic dataset: As we have used transcriptomic data of miRNA for the prediction, it is intriguing to know that even though the miRNA targets are computationally assessed, there is a very limited number that has been verified through experimentation. As many target assessing algorithms are applied, the results are found to be inconsistent, and appropriately finding functional miRNA targets is still a posing challenge [17]. Canonical sites are containing higher miRNA interactions and non-canonical sites that are supposedly less in showcasing much relevance. But on the other hand in other contrary studies, it has been noted that the entire miRNA should be considered for better verification or validation [19]. As is also conducive to the performance of target prediction tools which generally identify almost 80% of recognized miRNA targets and 20% compromising non-

canonical targets. It also poses a method of novel target approaches to be used to deduce relations and patterns among them.

Chapter 4

MATERIAL AND METHODS

The steps that are followed to validate the transcriptomic dataset through some ML techniques.

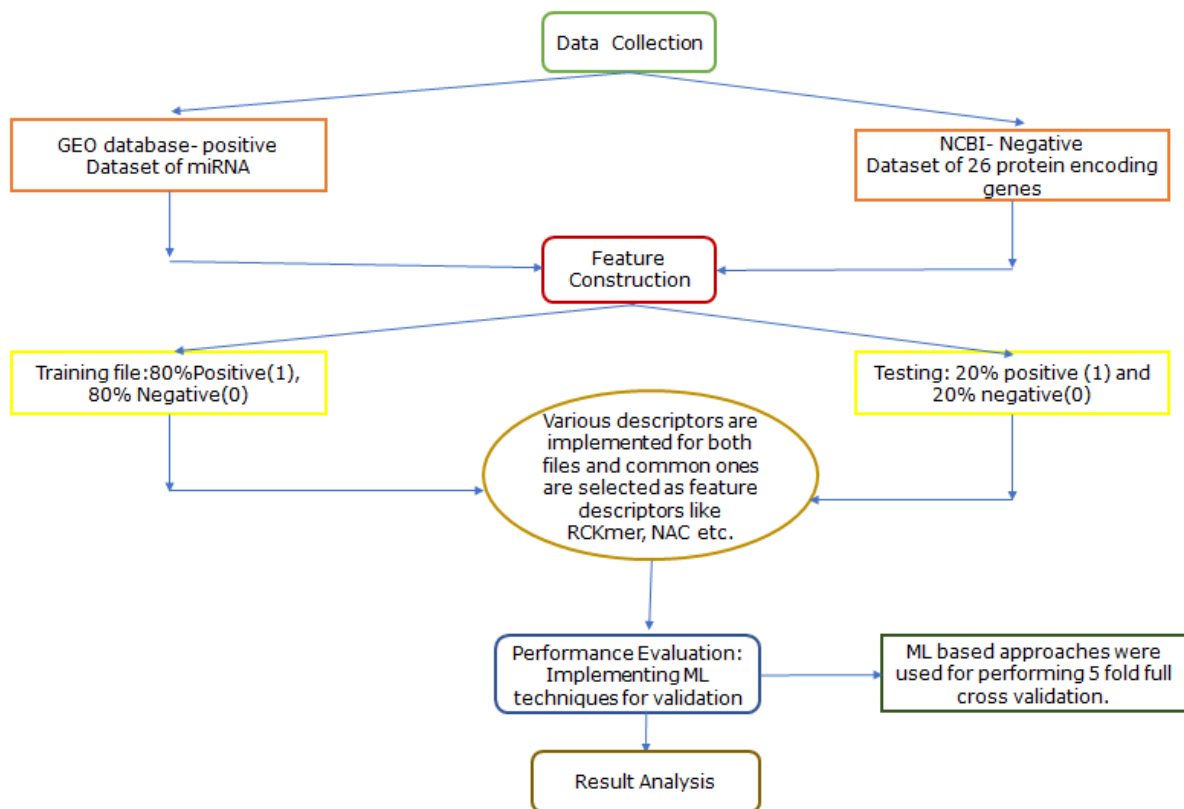


Fig 4.1 This flowchart showcases all steps followed through the project.

All these steps are enveloped in the overall methodologies that are followed during this project. We will discuss each of the steps mentioned above further in this report.

4.1 Data Collection and preprocessing: The dataset was obtained from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122452>) with miRNA reads of about 18-22 nucleotides in length. There were about 4863 such reads and this was considered a positive data set.

miR_name	miR_seq	len	genome	strand	start	end	pre-miF#mir/MimirIDs, rep_mirID	rep_mi	rep_miF	rep_miFtype	Sequenc
hsa-miR-1-3p	TGGAATGTAAGAA	22	chr18	-	2E+07	2E+07	agctaac 11 hsa-mirhsa-mir-1-2	acctaact	hsa-miF	TGGAATG 3'	Yes
hsa-miR-1-3p	TGGAATGTAAGAA	22	chr20	+	6E+07	6E+07	cctgctt 11 hsa-mirhsa-mir-1-1	tgggaa	hsa-miF	TGGAATG 3'	Yes
hsa-let-7a-5p	TGAGGTAGTAGGT	22	chr22	+	5E+07	5E+07	agaccge 12 hsa-lethsa-let-7a-3	gggTGAC	hsa-let	TGAGGTA 5'	Yes
hsa-let-7a-3p_R+1_1	CTATACAATCTACI	22	chr22	+	5E+07	5E+07	agaccge 12 hsa-lethsa-let-7a-3	gggTGAC	hsa-let	CTATACA 3'	Diff
hsa-miR-7-5p_R+1	TGGAAGACTAGTGA	24	chr19	+	5E+06	5E+06	agattac 10 hsa-mirhsa-mir-7-3	agattac	hsa-miF	TGGAAGT 5'	Diff
hsa-let-7b-5p	TGAGGTAGTAGGT	22	chr22	+	5E+07	5E+07	caaggoc 14 hsa-lethsa-let-7b	cggggTC	hsa-let	TGAGGTA 5'	Yes
hsa-let-7b-3p_1ss22CT	CTATACAACCTACI	22	chr22	+	5E+07	5E+07	caaggoc 14 hsa-lethsa-let-7b	cggggTC	hsa-let	CTATACA 3'	Diff
hsa-let-7i-5p	TGAGGTAGTAGTT	22	chr12	+	6E+07	6E+07	cccgcac 15 hsa-lethsa-let-7i	ctggcTC	hsa-let	TGAGGTA 5'	Yes
hsa-let-7i-3p	CTGCGCAAGCTACI	22	chr12	+	6E+07	6E+07	cccgcac 15 hsa-lethsa-let-7i	ctggcTC	hsa-let	CTGCGCA 3'	Yes
hsa-let-7g-5p	TGAGGTAGTAGTT	22	chr3	-	5E+07	5E+07	ccttttc 17 hsa-lethsa-let-7g	aggcTGA	hsa-let	TGAGGTA 5'	Yes
hsa-let-7g-3p_R+1	CTGTACAGGCCACI	22	chr3	-	5E+07	5E+07	ccttttc 17 hsa-lethsa-let-7g	aggcTGA	hsa-let	CTGTACA 3'	Diff
hsa-miR-7-5p_R+1	TGGAAGACTAGTGA	24	chr15	+	9E+07	9E+07	ctggate 11 hsa-mirhsa-mir-7-2	ctggate	hsa-miF	TGGAAGT 5'	Diff
hsa-mir-7-2-p3_4ss10C	CAACAAATCACAGI	22	chr15	+	9E+07	9E+07	ctggate 11 hsa-mirhsa-mir-7-2	ctggate	hsa-miF	CAACAA 3'	New
hsa-let-7e-5p	TGAGGTAGGAGGT	22	chr19	+	5E+07	5E+07	gtctgtc 14 hsa-lethsa-let-7e	ccgggTC	hsa-let	TGAGGTA 5'	Yes
hsa-let-7e-3p	CTATACGGCCTCCI	22	chr19	+	5E+07	5E+07	gtctgtc 14 hsa-lethsa-let-7e	ccgggTC	hsa-let	CTATACC 3'	Yes
hsa-let-7c-5p	TGAGGTAGTAGGT	22	chr21	+	2E+07	2E+07	taaggac 17 hsa-lethsa-let-7c	gcacccq	hsa-let	TGAGGTA 5'	Yes
hsa-let-7c-3p	CTGTACAACCTTCI	22	chr21	+	2E+07	2E+07	taaggac 17 hsa-lethsa-let-7c	gcacccq	hsa-let	CTGTACA 3'	Yes

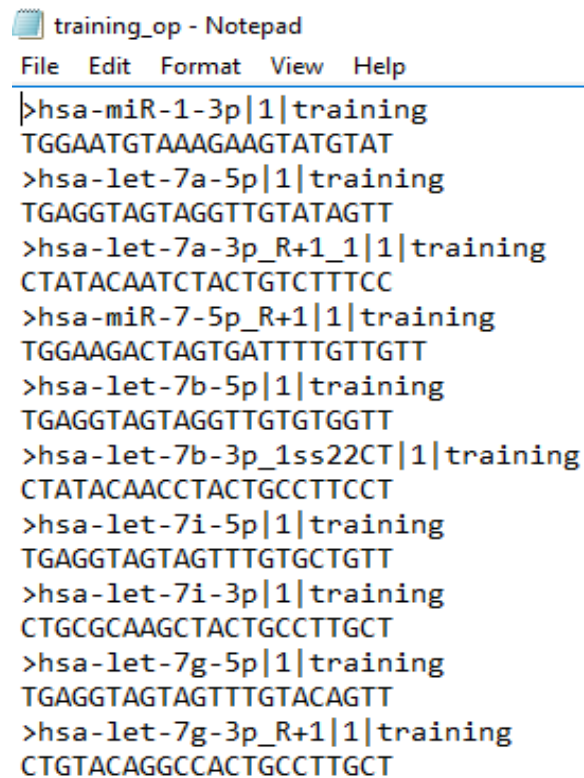
Fig 4.1.1 The dataset of all expressed miRNA

Now, for the known or negative dataset, we downloaded nucleotide sequences of 26 protein-coding genes through NCBI. Those protein-coding genes were: A1BG, A1CF, A2M, A2ML1, A3GALT2, A3GALT, A4GNT, AAAS, AACS, AADAC, AADACL2, AADACL3, AADACL4, AADAT, AAGAB, AAK1, AAMDC, AAMP, AANAT, AAR2, AARD, AARS1, AARS2, AARSD1, AASDH, AASHDPT. This sequence data was divided into equal reads that were up to 4891 with length ranging from 18-22 nucleotides.

Pre-processing of data: The most imperative step of any type of data analysis is pre-processing and normalization of raw data which is ultimately subjected to further analysis. This process reduces the noise resulting from technical variations and consequently permits data to be compared for predicting the actual biological changes. The implementation of data normalization aids in stabilizing imbalanced quantities of starting RNA, differences in labelling or detection efficiencies between the used fluorescent dyes and systematic biases in expression levels [25].

To remove any redundancy the reads with high sequence similarity were eliminated and both positive and negative sequence data was reduced to 1:1 using Cd Hit with an equal number of reads that remained to be 4187 in both cases. Further, the sequence data was divided into 80:20 for training and testing implementation. And finally, the training dataset is made up of 3349 (80%) positive sequence and 3349 (80%) of negative sequence [18]. Similarly, the testing file was composed of 837 (20%) of positive and 837 (20%) negative

sequence. Now for the positive dataset, it was binary classified as 1 and the negative dataset was classified as 0.



```
training_op - Notepad
File Edit Format View Help
>hsa-miR-1-3p|1|training
TGGAATGTAAAGAAGTATGTAT
>hsa-let-7a-5p|1|training
TGAGGTAGTAGGTTGTATAGTT
>hsa-let-7a-3p_R+1_1|1|training
CTATACAATCTACTGTCTTTCC
>hsa-miR-7-5p_R+1|1|training
TGGAAGACTAGTGATTTTGTGGTT
>hsa-let-7b-5p|1|training
TGAGGTAGTAGGTTGTGTGGTT
>hsa-let-7b-3p_1ss22CT|1|training
CTATACAACCTACTGCCTTCCT
>hsa-let-7i-5p|1|training
TGAGGTAGTAGTTTGTGCTGTT
>hsa-let-7i-3p|1|training
CTGCGCAAGCTACTGCCTTGCT
>hsa-let-7g-5p|1|training
TGAGGTAGTAGTTTGTACAGTT
>hsa-let-7g-3p_R+1|1|training
CTGTACAGGCCACTGCCTTGCT
```

Fig 4.1.2 This is the training data file with special header for iLearnPlus

testing_ot (1) - Notepad

File Edit Format View Help

```

>cgr-miR-7b_1ss7AC|1|testing
TGGAAGCCTTGTGATTTTGTGTT
>mml-miR-9-3-3p_1ss1CT|1|testing
TTCAAGCTAGATAACCGAAAGT
>oan-miR-15b-5p_R+1|1|testing
TAGCAGCACATCATGGTTTGCA
>oan-miR-15b-3p_R-1|1|testing
CGAATCATTATTTGCTGCTT
>eca-miR-17_R+1_1ss5GT|1|testing
CAAATTGCTTACAGTGCAGGTAGC
>oan-miR-19a-3p_R+1|1|testing
TGTGCAAATCTATGCAAACTGAC
>sha-miR-21_L+2R-2|1|testing
AATAGCTTATCAGACTGATGTTGAC
>mmu-miR-21a-3p_L-1_1ss5AC|1|testing
AACCGCAGTCGATGGGCTGTC
>mmu-miR-21b_1ss4TC|1|testing
TAGCTTATCAGACTGATATTTCC
>mdo-miR-22-5p_L+1_2ss16GA23AT|1|testing
CAGTTCTTCAGTGGCAAGCTTTT
>mdo-miR-22-3p_R+1|1|testing
AAGCTGCCAGTTGAAGAACTGCC
>eca-miR-22b-3p_R-1|1|testing

```

Fig 4.1.3 This is the testing data file

```

import str = ""
stack = []
prev = 0
# str = str.st
for item in str:

    n = random.randint(18,24)
    stack.append(str[prev:prev+n].replace('\n',''))
    prev = prev+n
    if prev>=len(str)-1 or len(stack)>4920:
        break
with open('genomes.txt','w') as f:
    for item in stack:
        f.write("%s\n"%item)

```

Fig 4.1.4 Code for creating reads of length ranging from 18-22

```

with open("training.txt", "r") as inp, open("training_op.txt", "w") as output:
    for line in inp:
        l = line.strip()
        if l.startswith(">"):
            output.write("{}|1|training\n".format(l))
        else:
            output.write("{}\n".format(l))

```

Fig 4.1.5 Code for creating special header for training and testing files

4.2. Feature extraction: The process of extracting feature descriptors were done through iLearnPlus. iLearnPlus is a ML-based software with graph-based and web-based user interface that helps in the construction of automated ML pipelines for computer based analysis and evaluation [14]. Four main modules are iLearnPlus-Basic, iLearnPlus-Estimator, iLearnPlus-AutoML, iLearnPlus-LoadModel for bioinformaticians to conduct customizable sequences based on feature engineering and analysis, machine-learning algorithm construction, performance assessment, statistical analysis and visualization of data.

Descriptors for both training and testing datasets:

4.2.1 Kmer: It helps in calculating the frequency of occurrence of k neighboring nucleotides which was usually used to facilitate the identification and regulatory sequence prediction

The Kmer descriptor is calculated using the formula :

$$f(t) = \frac{N(t)}{N}, \quad t \in \{AAA, AAC, AAG, \dots, TTT\},$$

Where where N(t) is the quantity of kmer type t, while N is the length of a nucleotide sequence. The Kmer descriptor has been effectively applied to lncRNA calculation [12].

4.2.2 Mismatch: The occurrence of kmers, enabling at most m mismatches is the mismatch profile which also aids to estimate the occurrences of kmers, but permits max m imprecise matching (m < k). There are two parameters for this descriptor, k neighboring nucleic acids and m imprecise matching. The mismatch descriptor is defined as:

$$f_{k,m} = \left(\sum_{j=0}^m c_{1,j}, \sum_{j=0}^m c_{2,j}, \dots, \sum_{j=0}^m c_{4^k,j} \right),$$

where $c_{i,j}$ represents the occurrences of i -th kmer type with j mismatches, $i = 1, 2, 3, \dots, 4k$; $j = 0, 1, 2, \dots, m$. The mismatch descriptor has been effectively applied to protein classification prediction, B-cell epitopes identification, and transposon-derived piRNA prediction. [12] [13]

4.2.3 NAC (Nucleic Acid Composition): The Nucleic Acid Composition (NAC) encoding estimates the frequency of every nucleic acid type in a sequence. The frequencies of all four natural nucleic acids (i.e. ACGT or U) can be quantified as:

$$f(t) = \frac{N(t)}{N}, \quad t \in \{A, C, G, T(U)\},$$

where $N(t)$ is the amount of nucleic acid type t , while N is the length of a nucleotide sequence [14].

4.2.4 NMBroto: It is normalized moreau broto autocorrelation. It is utilized to determine the distribution of the properties of amino acid across the sequence that is used. The formula assigned for the same is:

$$AC(d) = \sum_{i=1}^{N-d} P_i \times P_{i+d}, \quad d = 1, 2, \dots, nlag.$$

The normalized autocorrelation :

$$ATS(d) = \frac{AC(d)}{N-d}, \quad d = 1, 2, \dots, nlag.$$

4.2.5 RCKmer : This is known as reverse compliment kmer. It is a type of kmer where the kmers which are present are not obligated to be specific to a particular strand and also aids in estimating the reverse complement of k and the rate of occurrence. For example : there are 16 types of 2-mers (i.e. ‘AA’, ‘AC’, ‘AG’, ‘AT’, ‘CA’, ‘CC’, ‘CT’, ‘CG’, ‘GA’, ‘GC’, ‘GG’, ‘GT’, ‘TA’, ‘TC’, ‘TG’ and ‘TT’) in a DNA sequence. Among them, ‘TT’ is reverse compliment with ‘AA’. Thus, there are about ten kind of 2-mers in the RCKmer approach (i.e. ‘AA’, ‘AC’, ‘AG’, ‘AT’, ‘CA’, ‘CC’, ‘CG’, ‘GA’, ‘GC’ and ‘TA’) by eliminating the reverse complimentary Kmers. [12] [14]

4.2.6 Subsequence Profile: This descriptor allows for non-contiguous matching. For instance: the 3-mer “AAC” in the sequence “AACTACG”. Through accurate non-

contiguous matching, we can attain AAC, AA-C, A-AC, A-AC (“-” implies the gap in non-contiguous matching). AAC is the literal form of “AAC”, and AA-C, A-AC, A-AC are non-contiguous forms of “AAC”. The occurrences of non-contiguous types are penalized with their extent l and the factor δ ($0 \leq \delta \leq 1$), defined as δl . Thus, the occurrence of “AAC” in above example is $1 + 2\delta 6 + \delta 5$. The subsequence descriptor has been effectively implemented to B-cell epitopes identification, transposon-derived piRNA prediction. [12]

4.2.7 Z Curve 12bit: Z Curve 12bit is the criteria of Z Curve for phase independent dinucleotide. The Z_curve_12bit descriptor takes the frequency of dinucleotides, demonstrated by $p(XY)$, where $X, Y = A, C, G$ and T . This descriptor can be estimated as following:

$$\begin{cases} x_X = (p(XA) + p(XG)) - (p(XC) + p(XT)), \\ y_X = (p(XA) + p(XC)) - (p(XG) + p(XT)), \\ z_X = (p(XA) + p(XT)) - (p(XG) + p(XC)), \\ X = A, C, G, T, \end{cases}$$

4.2.8 Z Curve 36bit: This Z Curve criteria corresponds to phase specific dinucleotides. It is demonstrated same as the Z Curve 12 bit and the descriptors is estimated using:

$$\begin{cases} x_X^k = (p^k(XA) + p^k(XG)) - (p^k(XC) + p^k(XT)), \\ y_X^k = (p^k(XA) + p^k(XC)) - (p^k(XG) + p^k(XT)), \\ z_X^k = (p^k(XA) + p^k(XT)) - (p^k(XG) + p^k(XC)), \\ X = A, C, G, T; k = 1, 2, 3, \end{cases}$$

4.2.9 Z Curve 48bit: This parameter of Z curve is used for phase independent trinucleotides. Using similar definitions the descriptor is estimated using the following:

$$\begin{cases} x_{XY} = (p(XYA) + p(XYG)) - (p(XYC) + p(XYT)), \\ y_{XY} = (p(XYA) + p(XYC)) - (p(XYG) + p(XYT)), \\ z_{XY} = (p(XYA) + p(XYT)) - (p(XYG) + p(XYC)), \\ X = A, C, G, T; Y = A, C, G, T \end{cases}$$

4.2.10 Z Curve 144bit: In this criteria of Z Curve the descriptor is used to evaluate the phase specific trinucleotides and is represented as:

$$\begin{cases} x_{XY}^k = (p^k(XYA) + p^k(XYG)) - (p^k(XYC) + p^k(XYT)), \\ y_{XY}^k = (p^k(XYA) + p^k(XYC)) - (p^k(XYG) + p^k(XYT)), \\ z_{XY}^k = (p^k(XYA) + p^k(XYT)) - (p^k(XYG) + p^k(XYC)), \\ X = A, C, G, T; Y = A, C, G, T; k = 1, 2, 3. \end{cases}$$

This is effectively used for short coding sequences and their evaluation.

4.3 Implementation of ML approaches: The techniques of ML are used to evaluate the datasets which are already divided into training and testing in the previous step of data collection. There are about 11 ML methods that we implemented on both files for better results. Every method has some key factor of interest and different ways of approach and creating a model for further assessment and observation. These 11 methods are Naive Bayes, Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Multilayer perceptron (MLP), Stochastic gradient descent (SGD), eXtreme gradient boost (XG Boost), Random forest (RF), Logistic regression (LR), Support vector machine (SVM), K-nearest neighbors (KNN), Decision Tree. We will discuss thoroughly each of the methods ahead in this report. These methods were applied and five-fold cross-validation was executed. [13].

Five-Fold cross-validation: This is a type of k-fold cross-validation used for resampling of data where the dataset is of particular quantity and helps in deep evaluation of ML models that are utilized for this validation process. So the data sample is split into 5 groups in case of fivefold cross validation [20]. Cross-validation is primarily used in the applied ML to elucidate the significance or skill of the model. It is preferred because it a rather simpler approach to validate data and can easily be understood by anyone. Also, there is comparatively less biasness and the result can be trusted for the reliability of the model. It demands easy division of training and testing data and running the files to model. The ML method used is as following:

4.3.1. Naive Bayes: It is a supervised learning method and works on the principle theorem known as Bayes theorem that states that the probability of event A to take place on the given probability that event B has already occurred. This technique is based on probability statistics of independent events. The formula representing the theorem is:

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

Here the occurrence of event A is not affected by the occurrence of event B.

The part P(A|B) denotes the probability of hypothesis A w.r.t to the observed event B

$P(A)$ is the probability before observing and $P(B)$ is marginal probability. This classifier is quite simple and is an efficient classification method that assists in making rapid ML models and also helps in predicting quickly and effectively.

4.3 The ML method used for analysis are as following:

4.3.1. Naive Bayes: It is a supervised learning method and works on the principle theorem known as Bayes theorem that states that the probability of event A to take place on the given probability that event B has already occurred. This technique is based on probability statistics of independent events. The formula representing the theorem is:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Here the occurrence of event A is not affected by the occurrence of event B.

The part $P(A|B)$ denotes the probability of hypothesis A w.r.t to the observed event B

$P(A)$ is the probability before observing and $P(B)$ is marginal probability. This classifier is quite simple and is an efficient classification method that assists in making rapid ML models and also helps in predicting quickly and effectively [11].

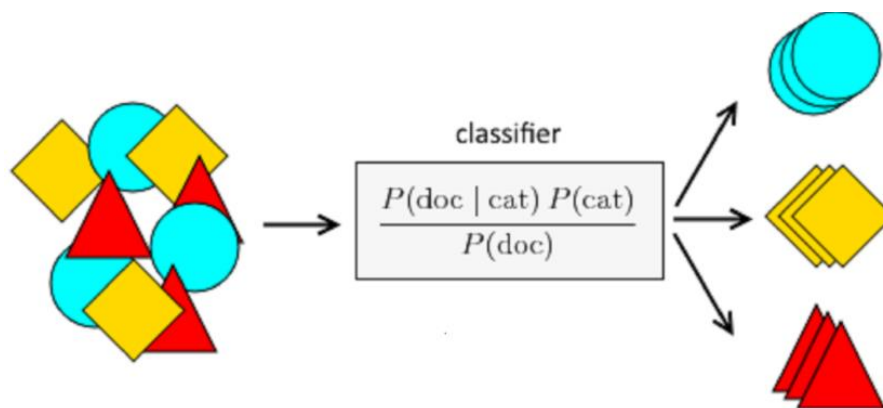


Fig 4.3.1.1 demonstrates Naive bayes classifier pertaining to independent probabilities of elements of dataset.

<https://www.itshared.org/2015/03/naive-bayes-on-apache-flink.html>

4.3.2 Linear Discriminant Analysis (LDA): This approach is distinctive in the ways that it is used for reduction in the dimensions of any problem that is to be classified. It is used in both supervised and unsupervised learning algorithms which enables the modelling process where separate groups or classes are modelled. It helps in the deduction of features from a higher dimension space to a lower one [11]. This method is usually used for

multidimensional space data where unobserved groups are acted upon. A new axis is created by LDA, where two key points are considered:

1. Maximize the length between the mean of the classes
2. Reduce the variation within the classes

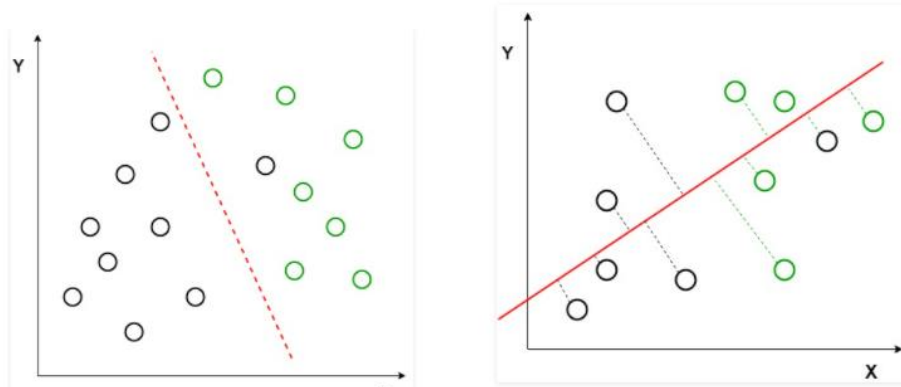


Fig 4.3.2.1 represents the axis between the two categories in the initial part and the steep axis which is at some definite distance reducing the covariance

<https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/>

4.3.3 Quadratic discriminant analysis (QDA): This modelling method is similar to that of LDA and only differentiates in that the covariance matrix is distinct for distinguishable classes. Hence, the covariance has to be calculated separately for all the classes that are considered.

4.3.4 Multilayer perceptron (MLP) : It is a part of feed-forward neural network and mainly is composed of three kinds of layers:

1. Input layer
2. Output layer
3. Hidden Layer

The input layer collects the signal of the input and processes it. The desired skills are done by the output layer like classification and prediction. There is an arbitrary amount of layers that are hidden which lies in between the input and the output layer which act as a computational core of the machine for the MLP. As it is seen in the feed-forward loop the flow of the data is in the forward direction, the same is the case with MLP as the data flows from the input layer to the output layer. It also helps in resolving answers to problems that are difficult to separate through the linear method [11]. MLP is structured in such a way

that it can estimate any continuous functions with rough accuracy. Pattern classification is one of the most executed applications of MLP.

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k$$

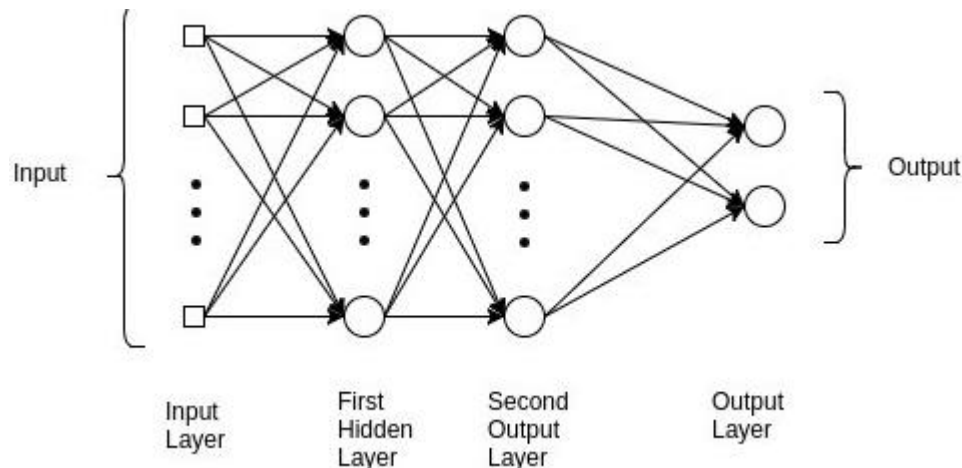


Fig 4.3.4.1 This represents all the three layers of a multi-layer perceptron and data flow through them

<https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/>

4.3.5 Stochastic Gradient Descent (SGD): This method uses multiple iterations for optimizing an objective function with certain properties like differentiability and sub-differentiability. It replaces the calculated gradient of the given dataset with the gradient descent of a randomly chosen subset of the data that is how it does stochastic approximation. Gradient descent is a popular method in ML and also in DL (deep learning) and if needed can be utilized with any or every ML approach. It is a function that represents the gradient of the slope and is the result of partial derivatives of the range of criteria of the input.

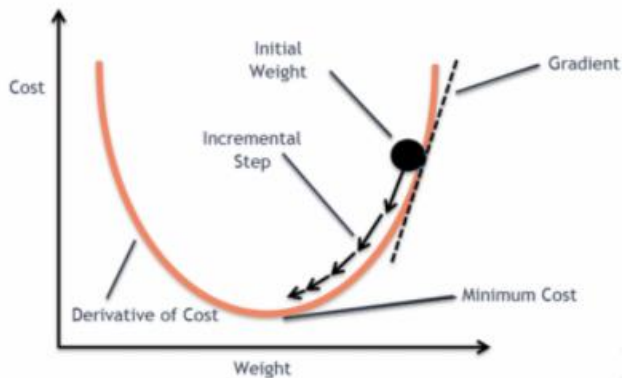


Fig 4.3.5.1 This shows the stochastic gradient descent and the weight steps pertaining to maximum cost and derivative cost

<https://towardsdatascience.com/implementing-sgd-from-scratch-d425db18a72c>

4.3.6 eXtreme gradient boosting (XG Boost): This is a method as suggested by the name that helps in boosting or elevating the performance of the model. The most effective models and the models with the weak result are combined for better prediction. If there is some incorrect evaluation then some higher weights are added. This boosting algorithm is a greedy method. It also has a stop criteria or called as early stopping or depth of tree in terms of several stages to prevent overfitting of training data.

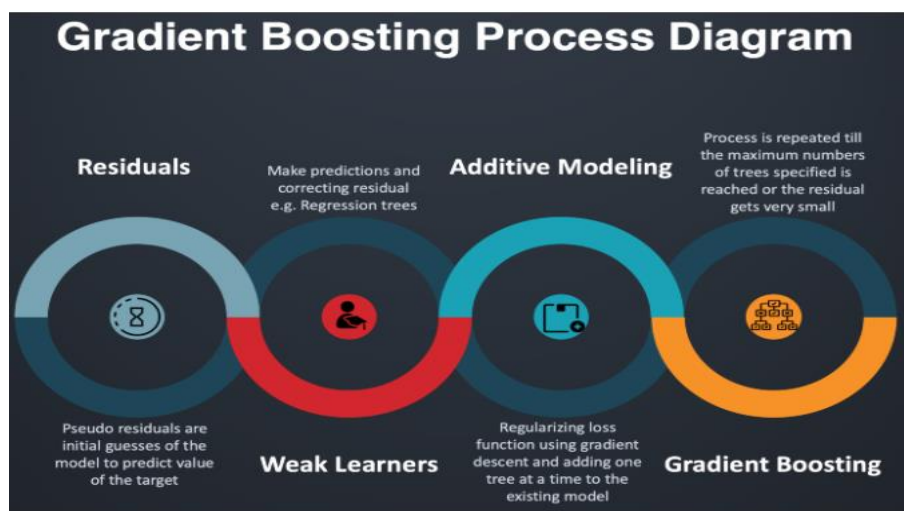


Fig 4.3.6.1 Represents the basic modelling process for XG Boost

https://dzone.com/articles/xgboost-a-deep-dive-into-boosting?edition=590295&utm_source=Zone%20Newsletter&utm_m

4.3.7 Logistics Regression (LR): It is a supervised learning technique which is used for calculating categorical dependent variables based on the range of the independent variable. The output is usually discrete values. The values may be YES, NO or 0, 1 or true and false. And it results in probabilistic values that occur between 0 and 1 instead of exact answers. It is the same as linear regression except for the steps that are used in the process. The difference is that the “S” shaped logistic function is used rather than a linear one as is the case in the linear regression technique [11]. The logistic function is used to determine the likelihood of the data. It is the most commonly used algorithm because it gives both probabilities and classification of the new data using continuous and distinct data.

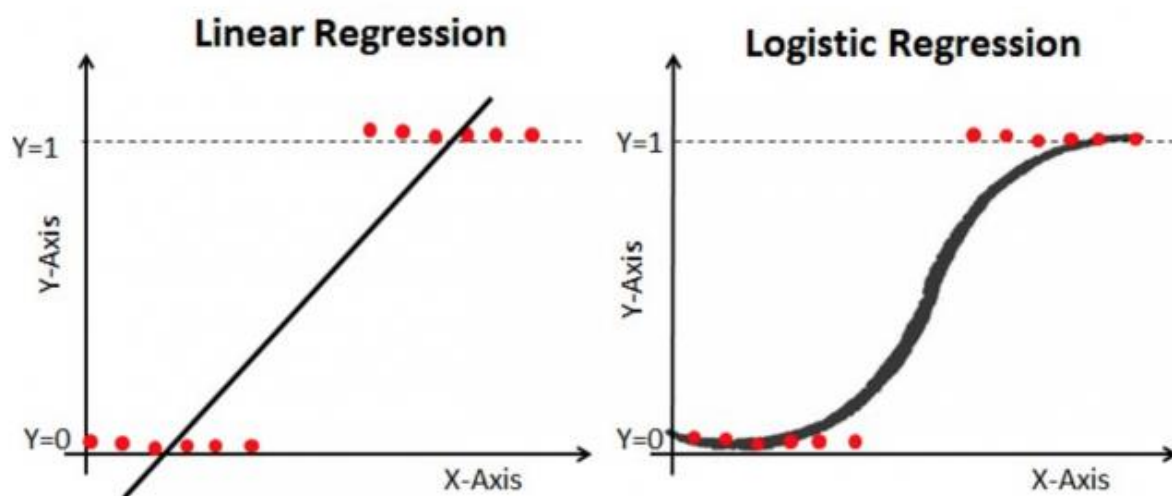


Fig 4.3.7.1 Both the diagrams represents linear and logistic regression respectively

<https://medium.com/mlearning-ai/logistic-regression-60694a973bee>

4.3.8 K nearest neighbors (KNN): This algorithm for ML is a supervised ML approach. It assumes the similarity between the existing and the new data and labels the new data as per the existing label or category of data. It acts by marking a new data point based on similar is new data with the already existing one. This neighboring data point aids in the classification of the new dataset. As this method is not based on specific parameters so it is not used to making assumptions based on any underlying data. It categorically classifies the data by considering the similarity index of the dataset [15].

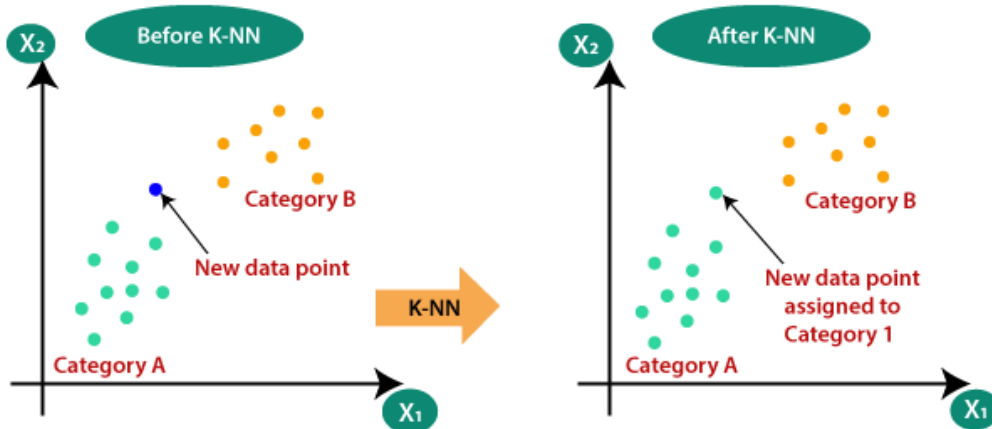


Fig 4.3.8.1 This represents the KNN approach where the nearest category is assigned the respective data point

<https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

4.3.9 Support vector machine (SVM): It is a quite famous method of ML and is a type of supervised learning which can help in solving both classification and regression problems. The primary objective of the SVM is to create the best possible decision boundary that separates the data into classes in the n-dimensional space. Now there is a category for the data and new data can be pushed into these categories based on similarity. The decision boundary that is created is called a hyperplane [22]. SVM finds the most intricate and extreme vectors to build this hyperplane and these are further called support vectors. There could be many decision boundaries to distinguish the dataset as needed and that also depends on the features of the dataset [22] [23]. The location of the hyperplane is affected by only the vector points associated with different categories.

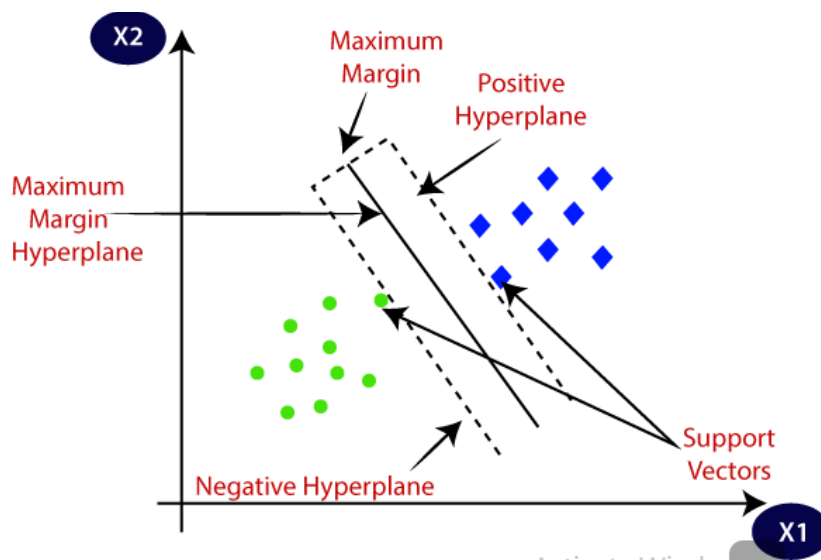


Fig 4.3.9.1 It represents the positive and negative hyperplane resulted due to SVM algorithm

<https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

4.3.10 Decision Tree: It is a supervised learning method of ML that is widely used for resolving classification problems. As the name suggests, the algorithm works in the form of the tree having root, internal node (a feature of the dataset), and leaf nodes (output). There are decision nodes that come into action when any decision has to be taken and it is composed of a lot of branches whereas leaf nodes are the outcome of the decision nodes and do not have branches. Like any other ML method, this also depends on the features of the dataset and is demonstrated in a graphical form so that all the possible solutions can be seen for a particular set of conditions. This method is used because it almost interprets how a human is supposed to address such complex problems associated with data and the core logic to build this tree is a lot simpler to understand in comparison to other techniques.

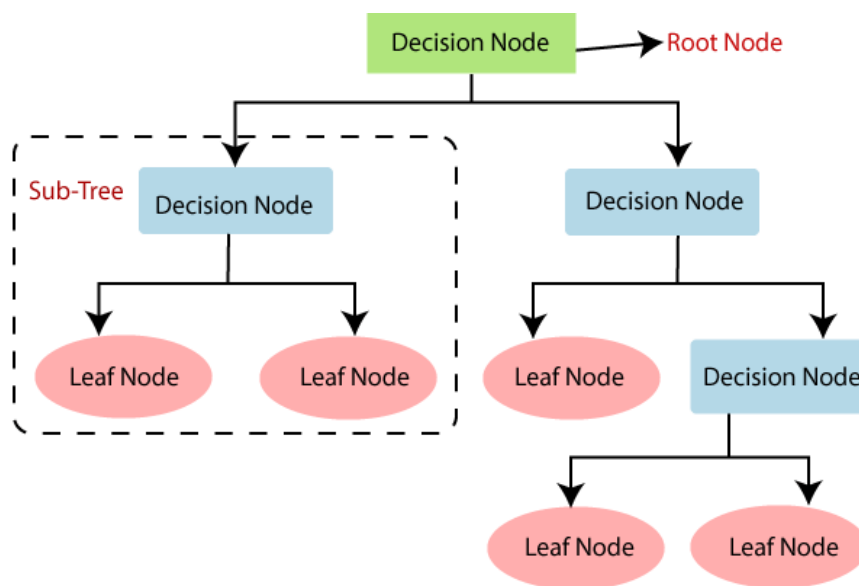


Fig 4.3.10.1 this is the diagram that clearly explain the decision tree and its elements

<https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>

4.3.11. Random Forest (RF): It is yet another tree-based approach and is quite flexible and effortless to use. It basically combines many decision trees and trains each with distinct sets of parameters or observations. It is used for classification as well as regression problems. The average of all the different predictions can give the best result in the random forest method. It has higher predictive accuracy than the decision tree because it uses the average function to solve the issue of overfitting trees. It also enables us to figure out the

most important feature in the data set. It selects random data samples and makes a decision tree and evaluates the result from independent trees and then the best result is chosen based on voting and the final prediction will be decided.

Random Forest Classifier

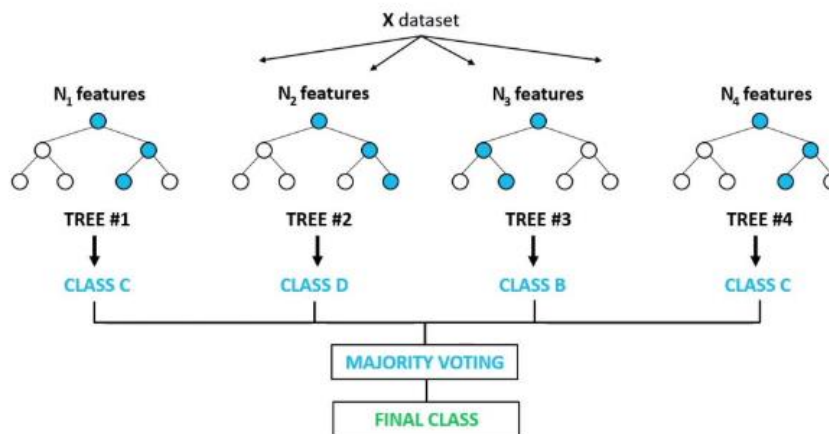


Fig 4.3.11.1 the image determine the basic outlook of the random process method

<https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/>

4.4 Performance evaluation: The results obtained by implementing all these 11 ML techniques using iLearnPlus –AutoML had these five key components that helps to decide the efficiency of the model created for the training and testing dataset.

4.4.1 Accuracy: It is defined as the most instinctive performance measure and it is a ratio of correct prediction to the total number of observations.

$$\text{Accuracy} = \text{sensitivity} * \text{prevalence} + \text{specificity} * (1 - \text{prevalence})$$

Its numerical value represents the truly positive results for the particular datasets. If we get a higher value of accuracy then our model is best.

4.4.2 Precision: It is defined as the number of true positives which is divide by the number of true positives plus the amount of false-positive values. It is also known as positive predictive value.

Basically the high value of precision responds to the low positive rate. It is called sensitivity in binary classification.

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

4.4.3 F1 Score: It is defined as the average of Precision and Recall. F1 Score is used to find the test accuracy, as it is calculated from precision and recall. The maximum value of the F1 score is 1.0, and it shows the perfect precision and recall, and the lowest value is 0.

$$\text{F1 Score} = \frac{2 * (\text{recall} * \text{precision})}{(\text{recall} + \text{precision})}$$

4.4.4 AUROC: The Receiver Operator Characteristic curve is used in binary classification for the evaluation of the matrices. It is a probability curve that plots the True Positive Rate in opposition to False Positive Rate at different threshold values and it can split the signal from the noise.

The Area under the Curve (AUC) is defined as the measure of the ability of a particular classifier so that it can distinguish between the classes and used in ROC curve.

Higher the AUC gives the improved performance of the model to distinguish between positive and negative classes.

4.4.5 AUPRC: The area under the precision-recall curve is defined as the performance matrices for the variation data in the problem where we have to find the positive data. Higher the AUPRC means it finds all of the positive data from the dataset. The average precision is one mode for the calculation of ARC. One attribute of AUPRC is that it does not use true negative data.

4.4.6 MCC: Matthews correlation coefficient is defined as the statistical rate which can produce a high score only if the prediction having good results in our confusion matrix that is true positive, false negative, true positive and false positive.

Correlation of C :1 having the perfect results between the prediction and the observation. It will returns values between -1 and +1.

As, there have been a rise in using ML algorithms for various validation and evaluation procedures, generating reliable results. After applying about 11 ML methods (Naive Bayes, LDA, QDA, MLP, SGD, XGBoost, RF, LR, SVM, KNN, Decision Tree)to 10 common descriptors of both training and testing dataset, we fetched a result consisting of values pertaining to six predicting attributes that are precision, accuracy, MCC, F1, AUROC, AUPRC. The 10 common descriptors are Kmer, Mismatch, NAC, NMBroto, Subsequence profile, RCKmer, Z_Curve 12 bit, Z_curve 36 bit, Z_Curve 48 bit an Z_Curve 144 bit.

TRAINING RESULTS									
Descriptor	Id	Sn	Sp	Pre	Acc	MCC	F1	AURO C	AUPR C
Kmer	NaiveBayes_model	45.402	67.29	58.792	56.35	0.1324	0.5069	0.6158	0.6175
	LDA_model	57.85	63.474	61.582	60.662	0.2147	0.5949	0.6524	0.6432
	QDA_model	4.42	99.432	73.206	51.934	0.106	0.0827	0.644	0.6385
	MLP_model	56.386	69.44	65.186	62.916	0.2641	0.5988	0.6801	0.67
	SGD_model	54.446	62.218	59.272	58.334	0.1698	0.5636	0.623	0.6206
	XGBoost_model	56	72.754	67.352	64.378	0.2936	0.6078	0.7114	0.7081
	RF_model	54.478	75.05	68.952	64.768	0.3047	0.6041	0.7199	0.7259
	LR_model	54.596	60.608	58.552	57.606	0.1543	0.5615	0.6191	0.6165
	SVM_model	97.94	0.508	49.578	49.216	-0.0317	0.6582	0.4793	0.5318
	SVM_model (Tuning)	57.284	72.992	67.912	65.142	0.3086	0.6177	0.7076	0.6854
KNN_model	64.866	55.594	59.328	60.23	0.2072	0.618	0.6404	0.6771	
DecisionTree_model	51.672	64.636	59.46	58.156	0.165	0.552	0.5815	0.6765	
Mismatch	NaiveBayes_model	49.224	61.172	55.984	55.202	0.1061	0.515	0.5936	0.5962
	LDA_model	53.404	59.206	56.936	56.304	0.1277	0.5474	0.5914	0.5898
	QDA_model	46.896	67.918	60.136	57.408	0.1542	0.5227	0.6141	0.622
	MLP_model	57.224	59.648	58.884	58.44	0.172	0.5743	0.6301	0.6305
	SGD_model	31.4	77.3	71.1	54.3	0.142	0.31	0.596	0.6083

		32	68	76	94	6	38		
	XGBoost_model	55.4 92	64.9 04	61.2 52	60.2	0.206 2	0.57 94	0.6503	0.6497
	RF_model	54.9 56	65.2 32	61.4 34	60.0 94	0.204 7	0.57 67	0.647	0.6506
	LR_model	53.8 22	59.0 84	57.0 42	56.4 54	0.130 8	0.55	0.5913	0.5894
	SVM_model	51.1 04	68.7 22	62.2 34	59.9 16	0.203	0.55 78	0.6328	0.624
	SVM_model	56.4 76	65.2 9	62.0 76	60.8 86	0.220 4	0.58 79	0.6517	0.6366
	KNN_model	57.7 02	57.1 76	57.4 52	57.4 38	0.149 8	0.57 39	0.5959	0.6312
	DecisionTree_model	52.8 96	60.0 4	56.9 98	56.4 7	0.130 1	0.54 75	0.5648	0.6672
NAC	NaiveBayes_model	52.3 88	60.5 74	56.7 48	56.4 86	0.131 3	0.53 4	0.6028	0.6071
	LDA_model	54.5 38	52.8 76	53.7 78	53.7 08	0.074 9	0.53 74	0.5687	0.5753
	QDA_model	0.09	100	60	50.0 5	0.016 4	0.00 18	0.6013	0.6058
	MLP_model	50.0 9	56.3 34	53.7 74	53.2 16	0.066 6	0.50 46	0.5712	0.5842
	SGD_model	60.5 98	46.6 08	51.8 08	53.6 04	0.085 4	0.54 16	0.5698	0.5763
	XGBoost_model	54.2 4	60.5 46	57.9 58	57.3 94	0.149 2	0.55 78	0.6183	0.6209
	RF_model	57.2 22	52.6 4	54.7 16	54.9 32	0.099 2	0.55 81	0.5598	0.5856
	LR_model	54.5 08	53.0 56	53.8 84	53.7 84	0.076 5	0.53 77	0.569	0.5755
	SVM_model	56.3	50.9 64	53.3 8	53.6 34	0.074 4	0.54 17	0.5772	0.5837
	SVM_model	51.4 9	58.6 92	55.5 74	55.0 96	0.103 8	0.52 27	0.5905	0.6045
	KNN_model	52.2 38	55.0 56	53.8 14	53.6 5	0.073 2	0.52 94	0.5448	0.5739
	DecisionTree_model	59.0 14	48.8 2	53.5 56	53.9 18	0.079 1	0.56 08	0.5476	0.6138
NMBroto	NaiveBayes_model	22.0 6	76.6 32	50.0 32	49.3 5	- 4	0.29 48	0.5045	0.5196
	LDA_model	51.6 72	55.5 06	54.0 36	53.5 88	0.072 2	0.52 6	0.5638	0.5442
	QDA_model	41.3 14	64.8 12	55.3 88	53.0 64	0.066	0.46 21	0.5695	0.563
	MLP_model	51.2 22	63.2 32	58.5 32	57.2 3	0.147	0.54 18	0.6074	0.6013
	SGD_model	48.6 88	57.6 28	54.5 72	53.1 58	0.065	0.50 26	0.567	0.5465
	XGBoost_model	49.8 48	62.9 04	57.3 8	56.3 78	0.129	0.53 2	0.5927	0.5964
	RF_model	50.4	65.5	59.4	57.9	0.162	0.54	0.6167	0.6273

		18	02	78	6	2	28		
	LR_model	51.4 64	55.5 36	53.9 12	53.5	0.070 4	0.52 44	0.5634	0.5434
	SVM_model	53.8 52	55.3 84	54.7 92	54.6 2	0.092 9	0.54 06	0.5715	0.5559
	SVM_model	46.2 38	72.2 18	62.6	59.2 3	0.191 8	0.53 06	0.6127	0.6131
	KNN_model	56.8 34	57.9 82	57.5 34	57.4 1	0.148 7	0.57 08	0.5966	0.6303
	DecisionTree_model	50.4 16	57.4 42	54.2 56	53.9 32	0.079	0.52 2	0.5393	0.6473
RCKmer	NaiveBayes_model	46.6 56	63.2 92	56.3 9	54.9 8	0.102 2	0.50 47	0.5988	0.5986
	LDA_model	57.2 52	61.2 64	59.9 22	59.2 6	0.186 5	0.58 36	0.631	0.6218
	QDA_model	1.88	99.7 6	74.0 24	50.8 28	0.068 6	0.03 66	0.6198	0.6123
	MLP_model	53.3 72	66.9 04	61.8 48	60.1 42	0.206 5	0.56 87	0.6471	0.6294
	SGD_model	48.1 5	67.5 32	59.8 7	57.8 44	0.161 6	0.52 92	0.6156	0.6064
	XGBoost_model	55.2 82	71.2 6	65.7 54	63.2 76	0.271 3	0.59 65	0.6942	0.6938
	RF_model	53.8 5	72.4 84	66.4 64	63.1 7	0.271 1	0.58 99	0.6901	0.6997
	LR_model	55.0 16	59.9 8	58.4 12	57.5 02	0.151 6	0.56 36	0.6125	0.6037
	SVM_model	59.4 92	55.5 02	57.3 28	57.4 98	0.151 5	0.57 94	0.616	0.6125
	SVM_model	54.6 26	68.9 34	63.7 2	61.7 82	0.239 3	0.58 54	0.6617	0.6506
	KNN_model	62.2 4	56.6 98	59.0 12	59.4 68	0.191 1	0.60 39	0.6255	0.6668
	DecisionTree_model	50.6 88	61.1 72	56.6 26	55.9 34	0.119 7	0.53 36	0.5593	0.6598
Subsequence	NaiveBayes_model	45.3 42	66.0 96	58.3 08	55.7 24	0.120 4	0.50 29	0.604	0.6143
	LDA_model	53.4 34	59.2 64	57	56.3 48	0.128 6	0.54 78	0.5922	0.591
	QDA_model	46.7 46	67.8 88	60.0 88	57.3 2	0.152 5	0.52 15	0.6131	0.6214
	MLP_model	48.1 2	66.1 56	58.7 92	57.1 4	0.145 8	0.52 75	0.61	0.6091
	SGD_model	51.6 12	56.1 64	53.3 62	53.8 86	0.081	0.51 46	0.556	0.5541
	XGBoost_model	51.2 52	69.1 42	62.7 4	60.2	0.209 1	0.56 15	0.6496	0.6503
	RF_model	47.1 66	70.7 84	62.2 02	58.9 76	0.186 6	0.53 41	0.6368	0.6496
	LR_model	53.7 62	59.1 14	57.0 56	56.4 38	0.130 5	0.54 98	0.5918	0.5904
	SVM_model	52.9 56	69.0 52	63.3 12	61.0 06	0.224 6	0.57 34	0.6437	0.6217

	SVM_model	52.9 56	69.0 52	63.3 12	61.0 06	0.224 6	0.57 34	0.6437	0.6217
	KNN_model	52.7 48	61.1 72	57.7 48	56.9 6	0.140 6	0.54 94	0.5909	0.6278
	DecisionTree_model	49.9 1	60.6 68	55.9 04	55.2 88	0.106 4	0.52 73	0.553	0.6543
Z_Curve_1 2bit	NaiveBayes_model	45.0 74	62.6 94	55.4 48	53.8 88	0.080 8	0.49 13	0.5795	0.589
	LDA_model	53.5 82	56.3 98	55.7 76	54.9 9	0.101 3	0.54 31	0.5739	0.5807
	QDA_model	46.1 78	65.3 52	57.7 48	55.7 66	0.119 2	0.50 83	0.5954	0.6027
	MLP_model	53.6 42	61.3 2	58.4 42	57.4 86	0.152 2	0.55 41	0.6181	0.6147
	SGD_model	49.5 82	60.4 58	56.2 28	55.0 2	0.106 1	0.51 31	0.5759	0.5811
	XGBoost_model	52.3 88	66.3 06	61.0 52	59.3 48	0.190 5	0.56 02	0.646	0.6418
	RF_model	51.4 32	69.3 5	62.8 98	60.3 94	0.213 1	0.56 25	0.6483	0.6442
	LR_model	53.3 12	56.5 78	55.7 62	54.9 46	0.100 5	0.54 15	0.5745	0.5798
	SVM_model	55.3 12	54.6 36	55.4 54	54.9 76	0.100 3	0.55 09	0.5745	0.583
	SVM_model	47.9 68	68.8 4	60.9 54	58.4 08	0.173 5	0.53 37	0.6214	0.6004
	KNN_model	56.2 68	58.2 2	57.4 62	57.2 48	0.145 5	0.56 73	0.5922	0.631
	DecisionTree_model	51.8 22	60.3 4	56.5 88	56.0 82	0.122 3	0.54 0.54	0.5609	0.6625
Z_Curve_3 6bit	NaiveBayes_model	51.9 7	60.9 96	57.7 3	56.4 84	0.131 8	0.54 22	0.5991	0.5961
	LDA_model	53.6 72	57.3 54	56.3 4	55.5 12	0.111 6	0.54 73	0.5732	0.5781
	QDA_model	46	69.8 88	60.9 3	57.9 48	0.165 9	0.51 83	0.6251	0.6172
	MLP_model	48.6 28	66.0 08	58.9 56	57.3 18	0.149 2	0.53 14	0.6034	0.5958
	SGD_model	51.8 52	58.3 06	55.9 68	55.0 82	0.103 9	0.53 48	0.5745	0.5783
	XGBoost_model	53.9 42	69.0 82	63.7 8	61.5 1	0.233 8	0.58 34	0.6627	0.6638
	RF_model	54.9 26	70.3 36	65.3 04	62.6 34	0.257 5	0.59 44	0.6779	0.6835
	LR_model	53.4	57.6 2	56.3 96	55.5 14	0.111 7	0.54 58	0.5751	0.5785
	SVM_model	56.7 78	54.2 8	55.9 76	55.5 28	0.111 5	0.56 03	0.5832	0.587
	SVM_model	47.6 14	75.6 18	66.3 96	61.6 18	0.243 8	0.55 09	0.6676	0.6726
	KNN_model	63.9 12	54.7 86	58.6 02	59.3 48	0.189	0.60 99	0.6237	0.6612
	DecisionTree_model	51.7 62	58.1 92	55.3 42	54.9 76	0.099 9	0.53 47	0.5498	0.6561

Z_Curve_4 8bit	NaiveBayes_mod el	46.5 96	65.2 9	58.1 74	55.9 48	0.124	0.51 19	0.605	0.6076
	LDA_model	55.2 84	60.9 02	58.9 52	58.0 96	0.163 4	0.56 81	0.6258	0.6184
	QDA_model	45.4 02	71.3 5	61.8 08	58.3 82	0.175 3	0.51 92	0.6407	0.6346
	MLP_model	52.6 86	67.2 64	61.5 4	59.9 76	0.202 5	0.56 43	0.6385	0.6343
	SGD_model	55.4 32	59.5 34	58.6 06	57.4 84	0.152 6	0.56 47	0.6232	0.6163
	XGBoost_model	56.0 88	72.7 24	67.3 66	64.4 08	0.294 8	0.60 81	0.7102	0.6999
	RF_model	55.1 94	75.2 3	69.4 16	65.2 14	0.314 1	0.60 95	0.7128	0.7141
	LR_model	55.3 42	60.6 06	58.7 62	57.9 76	0.161 2	0.56 73	0.6235	0.6166
	SVM_model	57.0 44	57.6 82	57.9 7	57.3 62	0.149 1	0.57 11	0.6159	0.6092
	SVM_model	53.1 34	74.6 32	67.7 78	63.8 86	0.286 5	0.59 01	0.6962	0.6801
	KNN_model	68.2 68	52.7	59.0 54	60.4 84	0.213 5	0.63 24	0.6332	0.6702
	DecisionTree_mo del	51.4 64	62.6 08	57.8 28	57.0 36	0.141 8	0.54 36	0.5703	0.6678
Z_curve144 bit	NaiveBayes_mod el	47.0 74	66.3 96	59.1 7	56.7 38	0.140 4	0.51 87	0.6091	0.6037
	LDA_model	55.8 2	61.6 52	59.4 92	58.7 38	0.175 9	0.57 41	0.6287	0.6132
	QDA_model	44.8 66	74.1 86	64.3 32	59.5 3	0.202 6	0.52 41	0.6526	0.6485
	MLP_model	52.2 68	64.6 38	59.7 2	58.4 56	0.170 9	0.55 62	0.6185	0.6018
	SGD_model	51.9 08	63.5 28	58.8 6	57.7 22	0.159 5	0.54 11	0.6265	0.6128
	XGBoost_model	57.1 64	70.2 78	65.9 38	63.7 22	0.278 1	0.61 05	0.6995	0.6896
	RF_model	56.4 78	73.0 52	68.0 44	64.7 66	0.301 8	0.61 4	0.7101	0.7183
	LR_model	54.9 84	61.8 6	59.2 38	58.4 26	0.169 9	0.56 81	0.6257	0.6124
	SVM_model	100	0	49.9 92	49.9 92	0	0.66 66	0.4126	0.4989
	SVM_model	51.4 92	76.9 02	69.3 78	64.2	0.295 9	0.58 74	0.6992	0.7106
	KNN_model	79.2 54	38.6 78	56.5 7	58.9 62	0.195 9	0.65 92	0.633	0.6861
	DecisionTree_mo del	54.1 52	59.8 64	57.4 58	57.0 08	0.140 6	0.55 71	0.5701	0.6726
TESTING RESULTS									
Descriptors	Id	Sn	Sp	Pre	Acc	MCC	F1	AURO C	AUPR C
Kmer	NaiveBayes_mod el	35.5 1	63.2 54	50.5 38	49.3 68	- 0.008	0.40 84	0.5123	0.5506

						9			
	LDA_model	55.9 44	55.1 42	55.1 76	55.5 34	0.111 1	0.55 47	0.5945	0.5937
	QDA_model	4.07	99.0 42	79.6 66	51.5 26	0.091 7	0.07 56	0.5521	0.5718
	MLP_model	55.9 38	55.9 84	55.7 14	55.9 52	0.119 7	0.55 72	0.5881	0.589
	SGD_model	46.1 8	59.2 22	50.8 28	52.6 62	0.052 4	0.46 06	0.5454	0.5601
	XGBoost_model	54.3 94	59.4 46	56.5 86	56.9 08	0.138 4	0.55 25	0.591	0.6016
	RF_model	56.5 52	58.2 5	56.7 92	57.3 84	0.149 1	0.56 25	0.5949	0.6045
	LR_model	53.0 76	47.3 58	49.7 64	50.2 1	0.004 6	0.51 21	0.5021	0.5341
	SVM_model	60	40	30.0 3	50.0 3	0	0.40 03	0.5299	0.6889
	SVM_model	59.4 14	57.6 58	57.8 56	58.5 26	0.171 5	0.58 48	0.6058	0.6069
	KNN_model	55.3 4	52.3 96	53.5 34	53.8 58	0.078	0.54 26	0.5456	0.5863
	DecisionTree_model	50.3 14	55.5 08	53.0 44	52.9	0.058 4	0.51 56	0.5291	0.6411
Mismatch	NaiveBayes_model	42.3 14	50.3 42	46.5 78	46.3 14	- 0.075 2	0.43 85	0.4531	0.5007
	LDA_model	43.7 5	50.8 28	46.7 36	47.2 82	0.055 2	0.45 05	0.4621	0.4846
	QDA_model	44.4 68	60.3 94	52.7 86	52.4 22	0.049 1	0.48 01	0.5572	0.564
	MLP_model	66.9 24	38.8 82	52.3 02	52.9 02	0.062 3	0.58 6	0.5621	0.5753
	SGD_model	54.5 94	47.0 66	49.7 12	50.7 5	0.010 8	0.43 17	0.4816	0.5692
	XGBoost_model	51.0 44	54.9 02	52.5 64	52.9 64	0.059 1	0.51 6	0.5371	0.5509
	RF_model	51.0 52	54.3 04	51.8 42	52.6 62	0.052 8	0.51 15	0.5484	0.5726
	LR_model	43.3 94	50.4 68	46.2 9	46.9 24	- 0.062 7	0.44 64	0.4627	0.4855
	SVM_model	54.7 62	51.7 94	52.1 32	53.2 62	0.066	0.53 04	0.5503	0.5505
	SVM_model	55.4 78	57.7 82	55.8 46	56.6 14	0.133 0.133	0.55 31	0.5686	0.5844
	KNN_model	54.5 1	52.7 44	53.2 3	53.6 22	0.073 1	0.53 73	0.5449	0.5728
	DecisionTree_model	51.9 9	52.2 68	52.0 3	52.1 22	0.042 5	0.51 9	0.5213	0.6402
NAC	NaiveBayes_model	43.6 38	52.2 62	46.9 9	47.9 32	- 0.044 6	0.44 73	0.4733	0.5122

	LDA_model	52.9 22	33.0 02	44.4 1	42.9 68	- 0.145 8	0.48 25	0.4086	0.4613
	QDA_model	0.35 8	100	40	50.1 5	0.026 4	0.00 71	0.4914	0.5329
	MLP_model	50.9 28	53.4 68	51.4	52.1 84	0.042 8	0.50 8	0.5242	0.5487
	SGD_model	27.8 72	69.9 42	42.7 54	48.9 54	- 0.014 9	0.26 23	0.4262	0.4654
	XGBoost_model	55.9 24	53.5 84	54.6 18	54.7 5	0.095 4	0.55 18	0.5766	0.5883
	RF_model	56.0 42	50.3 64	53.0 9	53.1 96	0.064 3	0.54 45	0.5456	0.5589
	LR_model	53.6 38	30.9 72	43.9 42	42.3 14	- 0.160 1	0.48 27	0.4054	0.4547
	SVM_model	56.3 96	27.7 38	43.9 94	42.0 76	- 0.168 1	0.49 39	0.4071	0.4594
	SVM_model	57.1 36	46.0 62	51.2 58	51.5 88	0.032 7	0.53 78	0.5326	0.5585
	KNN_model	50.9	53.1 1	52.0 1	52.0 04	0.040 1	0.51 44	0.5194	0.5468
	DecisionTree_model	57.5 96	47.7 28	52.4 88	52.6 58	0.053 6	0.54 87	0.5241	0.6376
NMBroto	NaiveBayes_model	54.1 64	40.5 4	48.3 44	47.3 36	- 0.063 9	0.49 69	0.4643	0.4876
	LDA_model	55.3 4	40.8 88	48.6 38	48.1 14	- 0.040 6	0.51 37	0.4806	0.4945
	QDA_model	51.6 48	40.4 12	46.5 32	46.0 22	- 0.087 2	0.48 17	0.47	0.5074
	MLP_model	48.7 68	49.6 44	48.9 44	49.1 96	- 0.016 1	0.48 66	0.5005	0.534
	SGD_model	49.1 44	45.7 32	48.5 6	47.4 02	- 0.060 7	0.43 57	0.4728	0.4899
	XGBoost_model	50.4 26	47.3 66	48.9 38	48.8 96	- 0.022 1	0.49 65	0.4964	0.5165
	RF_model	51.2 76	46.7 68	48.9 22	49.0 16	- 0.019 7	0.49 92	0.5026	0.5264
	LR_model	55.3 42	42.3 22	49.2 9	48.8 3	- 0.026	0.51 66	0.4881	0.499
	SVM_model	61.3 36	37.2 18	49.0 26	49.2 54	- 0.011 2	0.53 73	0.5143	0.5246
	SVM_model	54.6 1	47.6 06	51.1 04	51.1 06	0.022 3	0.52 71	0.527	0.5256

	KNN_model	48.1 7	49.7 6	48.7 42	48.9 52	- 9	0.48 32	0.4868	0.5247
	DecisionTree_model	52.2 18	48.4 44	50.3 22	50.3 32	0.006 6	0.51 24	0.5033	0.6322
RCKmer	NaiveBayes_model	35.3 8	60.2 68	47.8 66	47.8 14	- 0.043	0.40 29	0.4741	0.5231
	LDA_model	54.7 48	54.5 46	54.2 84	54.6 4	0.093 1	0.54 41	0.5762	0.5805
	QDA_model	1.79 6	99.8 8	97.7 78	50.8 1	0.082 3	0.03 47	0.5555	0.5711
	MLP_model	53.5 48	54.3 08	53.6 32	53.9 2	0.078 8	0.53 47	0.5733	0.5765
	SGD_model	65.8 38	34.9 44	50.2 18	50.3 92	0.015 5	0.56 2	0.5226	0.5411
	XGBoost_model	53.5 56	59.6 82	56.4 72	56.6 1	0.132 7	0.54 65	0.5822	0.5824
	RF_model	52.7 24	58.3 7	55.0 48	55.5 34	0.111	0.53 51	0.576	0.588
	LR_model	51.2 84	46.6 48	48.5 3	48.9 58	- 0.020 5	0.49 7	0.487	0.52
	SVM_model	60	40	30.0 3	50.0 3	0	0.40 03	0.543	0.7095
	SVM_model	58.9 26	54.0 68	55.9 72	56.4 9	0.130 9	0.57 33	0.5911	0.5854
	KNN_model	54.4 98	52.2 7	53.1 64	53.3 78	0.068	0.53 74	0.5377	0.5732
	DecisionTree_model	52.1 02	55.0 2	53.6 08	53.5 62	0.071 3	0.52 78	0.5356	0.6484
Subsequence	NaiveBayes_model	35.3 78	59.3 14	48.5 04	47.3 32	-0.05	0.39 95	0.4801	0.5321
	LDA_model	45.1 76	51.3 08	47.9 72	48.2 36	0.035 7	0.46 41	0.4708	0.4907
	QDA_model	43.6 18	62.7 9	54.0 08	53.1 98	0.065 5	0.48 07	0.5642	0.5717
	MLP_model	54.6 28	51.0 7	52.3 98	52.8 44	0.057 4	0.53 41	0.5565	0.5722
	SGD_model	29.5 56	74.3 06	50.8 18	51.8 94	0.034 4	0.35 13	0.5147	0.5335
	XGBoost_model	55.3 32	55.9 76	55.6 98	55.6 52	0.113 5	0.55 44	0.5706	0.5665
	RF_model	54.1 4	53.8 2	53.8 7	53.9 76	0.079 7	0.53 98	0.5592	0.5757
	LR_model	44.5 78	50.9 5	47.5 2	47.7 6	- 0.045 2	0.45 9	0.4688	0.4893
	SVM_model	53.5 64	50.3 6	51.0 2	51.9 46	0.039 8	0.51 95	0.5529	0.5528
	SVM_model	56.7 86	50.7 12	52.9 46	53.7 4	0.076 2	0.54 65	0.5559	0.5471
	KNN_model	56.5 26	49.6 32	52.9 02	53.0 82	0.062	0.54 61	0.5458	0.5908

	DecisionTree_model	53.7 76	49.6 4	51.6 42	51.7 06	0.034 2	0.52 64	0.5171	0.6427
Z_Curve_1 2bit	NaiveBayes_model	40.5 32	57.5 22	48.9 54	49.0 08	- 0.020 5	0.43 7	0.494	0.5353
	LDA_model	49.9 58	49.0 4	49.3 22	49.4 92	- 0.010 2	0.49 47	0.4994	0.5202
	QDA_model	42.4 44	61.5 88	52.4 7	52.0 04	0.041 5	0.46 36	0.5402	0.5586
	MLP_model	55.3 46	54.9 08	54.7 74	55.1 18	0.102 9	0.54 93	0.5796	0.5893
	SGD_model	25.6 56	72.3 36	54.0 5	49.0 14	- 0.006	0.31 4	0.5035	0.5215
	XGBoost_model	54.2 72	55.9 78	54.7 82	55.1 14	0.103 3	0.54 3	0.5824	0.5845
	RF_model	54.6 32	56.4 54	55.0 54	55.5 34	0.111 2	0.54 62	0.5752	0.5789
	LR_model	49.9 62	47.9 66	48.7 12	48.9 54	- 0.020 9	0.49 12	0.4872	0.5127
	SVM_model	53.5 52	38.6 24	46.3 46	46.0 82	- 0.080 6	0.49 48	0.4465	0.4885
	SVM_model	57.0 24	49.1 62	52.3 2	53.0 82	0.063 7	0.54 31	0.5489	0.5443
	KNN_model	54.7 32	49.8 8	52.1 46	52.3 52.3	0.046 4	0.53 33	0.5453	0.5962
	DecisionTree_model	50.9 14	49.4 02	50.1 2	50.1 52	0.003 5	0.50 34	0.5016	0.628
Z_Curve_3 6bit	NaiveBayes_model	47.8 16	44.4 86	46.5 08	46.1 38	- 0.080 4	0.46 65	0.4573	0.5006
	LDA_model	49.2 44	49.0 36	49.0 18	49.1 34	- 0.017 4	0.49 02	0.5016	0.5157
	QDA_model	46.6 14	55.3 76	51.0 52	50.9 86	0.019 8	0.48 57	0.5346	0.553
	MLP_model	53.3 08	51.7 96	52.2 56	52.5 44	0.051 2	0.52 69	0.5463	0.5601
	SGD_model	48.8 98	51.2 96	47.1 1	50.0 88	- 0.009 9	0.44 3	0.4936	0.5098
	XGBoost_model	53.6 74	56.9 4	55.0 16	55.2 94	0.106 7	0.54 06	0.5706	0.5743
	RF_model	54.3 88	51.4 28	52.4 64	52.9 52.9	0.058 5	0.53 3	0.5418	0.5432
	LR_model	48.7 64	45.2 08	47.0 06	46.9 82	- 0.060 8	0.47 75	0.476	0.4996
	SVM_model	60 60	40 40	30.0 3	50.0 3	0 0	0.40 03	0.5389	0.6923
	SVM_model	56.4 16	52.1 52	53.8 62	54.2 78	0.086 5	0.54 97	0.5534	0.5671

	KNN_model	58.0 74	52.7 46	55.1 74	55.4 12	0.108 5	0.56 55	0.5688	0.5992
	DecisionTree_model	51.0 28	49.6 34	50.3 24	50.3 32	0.006 7	0.50 64	0.5033	0.6293
Z_Cuve_48 bit	NaiveBayes_model	35.5 04	64.5 72	52.1 42	50.0 22	0.006 9	0.41 24	0.4993	0.546
	LDA_model	53.7 9	54.5 44	53.8 26	54.1 58	0.083 5	0.53 71	0.5524	0.5634
	QDA_model	43.9 94	61.2 26	53.1 6	52.5 98	0.052 9	0.47 82	0.5438	0.5673
	MLP_model	52.9 42	54.1 82	53.5 18	53.5 56	0.071 3	0.53 21	0.5613	0.5772
	SGD_model	49.1 42	58.2 36	53.4 82	53.6 82	0.075 5	0.50 4	0.551	0.5634
	XGBoost_model	53.4 36	58.2 58	55.5 76	55.8 34	0.117 1	0.54 26	0.5865	0.5949
	RF_model	55.2 3	54.9 02	54.2 94	55.0 52	0.102 1	0.54 52	0.5742	0.5917
	LR_model	54.2 74	52.1 48	52.6 64	53.2 02	0.064 5	0.53 34	0.5405	0.5584
	SVM_model	60	40	30.0 3	50.0 3	0	0.40 03	0.4876	0.6481
	SVM_model	57.1 46	54.5 44	55.0 32	55.8 32	0.117 8	0.55 89	0.5769	0.5792
	KNN_model	53.1 86	52.9 82	52.8 24	53.0 78	0.062	0.52 88	0.5449	0.5962
	DecisionTree_model	52.8 28	52.9 88	52.6 38	52.9 02	0.058 2	0.52 63	0.5291	0.6454
Z_Cuve_14 4bit	NaiveBayes_model	38.0 08	59.9 08	50.0 64	48.9 48	- 2	0.42 39	0.4902	0.5363
	LDA_model	52.4 66	51.0 7	51.6 78	51.7 66	0.035 4	0.52 03	0.5299	0.5468
	QDA_model	48.0 56	56.4 46	52.2 74	52.2 4	0.044 7	0.49 85	0.5227	0.5597
	MLP_model	51.7 48	51.3 12	51.4 96	51.5 28	0.030 7	0.51 56	0.5289	0.5317
	SGD_model	43.0 26	60.0 28	52.8 78	51.5 28	0.035 6	0.46 33	0.5289	0.551
	XGBoost_model	53.3 1	56.3 34	54.7 7	54.8 16	0.096 6	0.53 95	0.5679	0.5724
	RF_model	57.7 34	52.1 48	54.2 92	54.9 34	0.099 8	0.55 8	0.5617	0.572
	LR_model	52.4 7	48.5 64	50.2 88	50.5 1	0.010 5	0.51 28	0.5175	0.5417
	SVM_model	60	40	30.0 3	50.0 3	0	0.40 03	0.4848	0.6446
	SVM_model	54.9 9	52.5 08	53.1 62	53.7 4	0.075 5	0.53 89	0.5438	0.5619
	KNN_model	53.6 66	55.9 82	54.8 08	54.8 16	0.096 8	0.54 11	0.5537	0.5896
	DecisionTree_model	54.0 16	48.5 62	51.2 76	51.2 86	0.026	0.52 48	0.5129	0.6415

Table 5.1 The table showcases the result of all ML techniques for both the datasets

- The different technique showed variation in values for different descriptors. But the AUROC value that was almost over 0.5 in most of the cases is evident of the fact that all the descriptors were informative and conducive for the efficiency of prediction.
- On the basis of the results, RF and XG Boost turned out to be the best classifier with an average value of AUROC of 0.66194 and 0.66349. SVM and Naive Bayes also showed better results with AUROC average of about 0.655 and 0.59122 respectively.
- Rest all other ML methods which were used, showed poor performance in prediction.

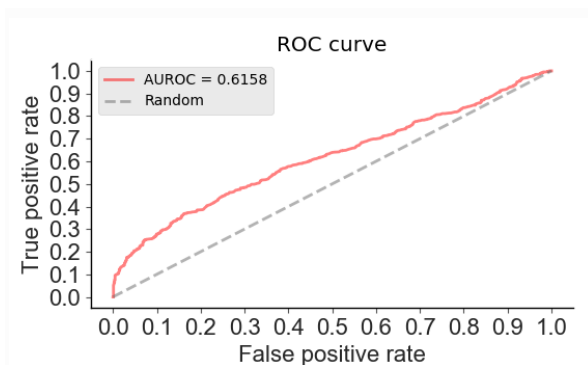


Fig 5.2 Kmer ROC curve

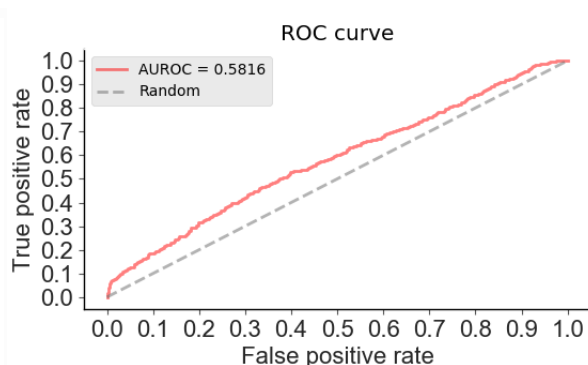


Fig 5.3 Mismatch ROC curve

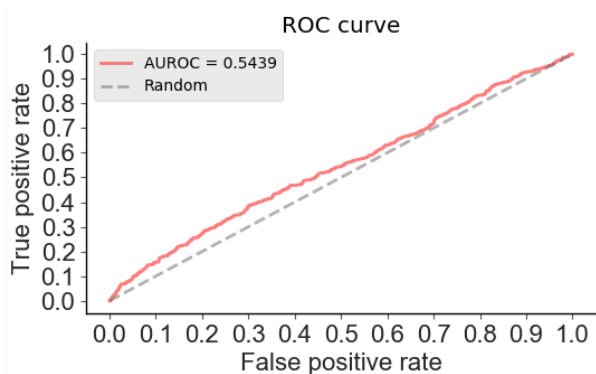


Fig 5.3 NAC ROC curve

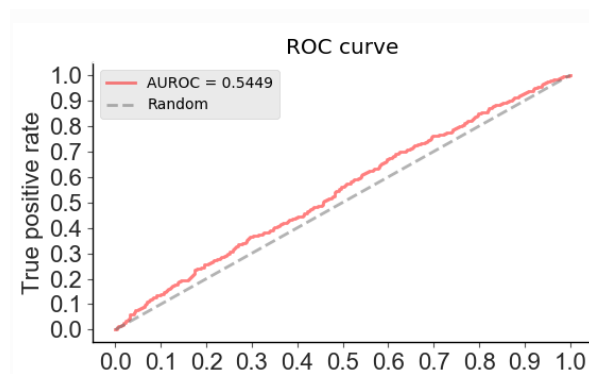


Fig 5.4 NMBroto ROC curve

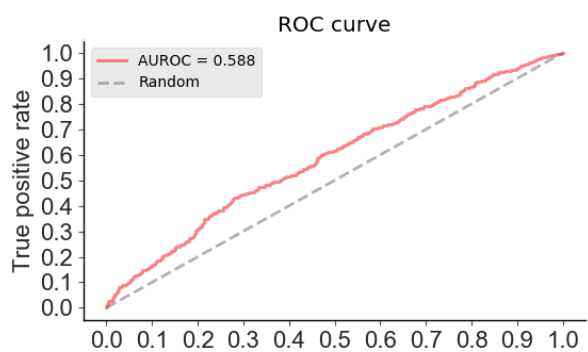


Fig 5.5 RCKmer ROC curve

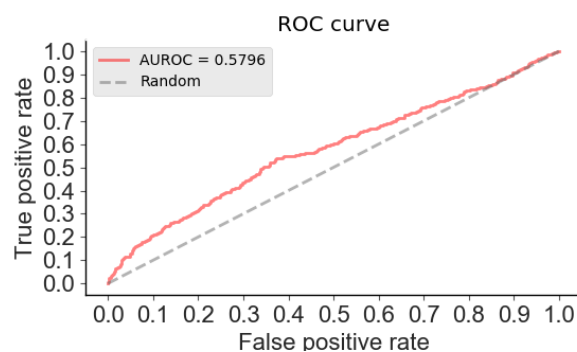


Fig 5.6 Subsequence ROC curve

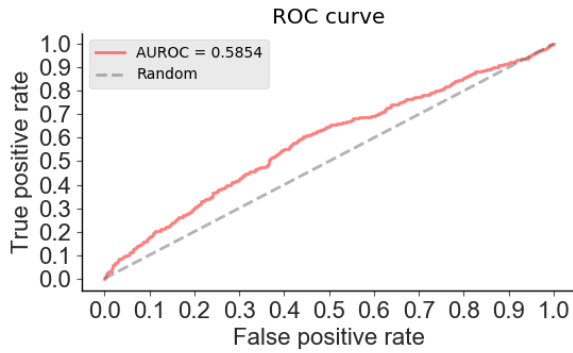


Fig 5.7 Z_Curve 12 bit ROC curve

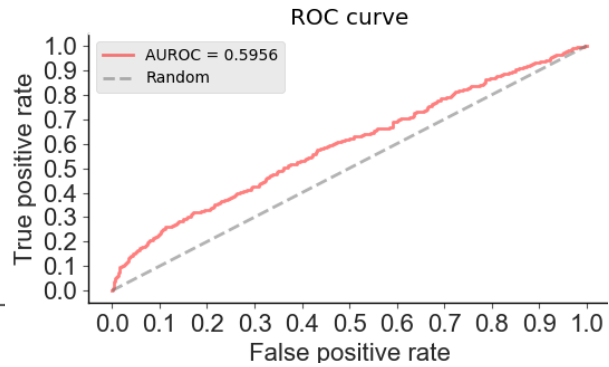


Fig 5.8 Z_Curve 36 bit ROC curve

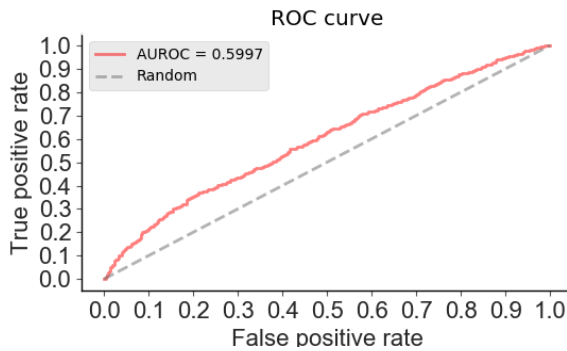


Fig 5.9 Z_Curve 48 bit ROC curve

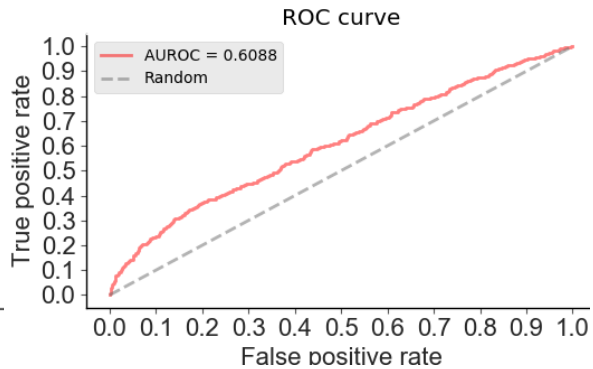


Fig 5.10 Z_Curve 144 bit ROC curve

- For the classification algorithms, SVM could prove it as the most influential classifier across 10 kinds of features, followed by LR, XG Boost, RF and AB. It is known that several features consist of a large amount of dimensions, but they are not equally crucial for the model performance.
- We trained different models of descriptors against the testing file of the descriptors to deduce the ROC curve (similarity) and accuracy where we used the SVM model for each descriptor after doing optimization that is tuning of ML approach for a particular data.
- The model prediction for Kmer descriptor predicted highest value of ROC curve that is about 0.6158, so it can be said that similarity of dataset was reliable.
- Z_Curve 144 bit also exhibited a dependable value of ROC curve that is 0.6088 signifying high similarity.

MicroRNAs are non-coding RNA molecules which help in the regulation of gene expression. Mainly the under and over expression of miRNAs has been related to the treatment or diagnosis of the specific cancer type. In this we use different ML algorithms that can predict the similarity in both the datasets positive and negative. We concluded that Kmer was the best descriptor to draw similarity between the training and testing datasets after five fold cross validation and turned out to be the most reliable descriptor. Kmer also showed the maximum accuracy in SVM model prediction that was about 68.34% and 0.6158 value of ROC curve determining the similarity of training and testing dataset and Z curve also showed 0.688 value indicating great similarity. Also, it was interesting to note that the SVM appeared to be the most useful classifier in predicting values of various attributes for both testing and training dataset followed by LR and XG Boost and RF. But other models exhibited poor performance in terms of accuracy of the descriptors. So, the comparison analysis of transcriptomic data with the protein coding data helped in deducing the significance of important descriptors such as Kmer and Z_Curve 144bit and laid emphasis on the most reliable classification methods such as SVM and LR.

1. V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra, “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques,” vol. 4, p. 7, 2013.
2. A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, Mar. 2011, doi: [10.3322/caac.20107](https://doi.org/10.3322/caac.20107).
3. N. G. Bediaga *et al.*, “A microRNA-based prediction algorithm for diagnosis of non-small lung cell carcinoma in minimal biopsy material,” *Br J Cancer*, vol. 109, no. 9, pp. 2404–2411, Oct. 2013, doi: [10.1038/bjc.2013.623](https://doi.org/10.1038/bjc.2013.623).
4. A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, Mar. 2011, doi: [10.3322/caac.20107](https://doi.org/10.3322/caac.20107).
5. S. S. Ramalingam, T. K. Owonikoko, and F. R. Khuri, “Lung cancer: New biological insights and recent therapeutic advances,” *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 91–112, Mar. 2011, doi: [10.3322/caac.20102](https://doi.org/10.3322/caac.20102).
6. N. J. Pastis, “The American College of Chest Physicians Lung Cancer Guidelines (3rd Edition),” *Chest*, vol. 143, no. 5, pp. 1193–1195, May 2013, doi: [10.1378/chest.12-3108](https://doi.org/10.1378/chest.12-3108).
7. P. Gasparini *et al.*, “microRNA classifiers are powerful diagnostic/prognostic tools in *ALK-*, *EGFR-*, and *KRAS-* driven lung cancers,” *Proc Natl Acad Sci USA*, vol. 112, no. 48, pp. 14924–14929, Dec. 2015, doi: [10.1073/pnas.1520329112](https://doi.org/10.1073/pnas.1520329112).
8. M. Frye, B. T. Harada, M. Behm, and C. He, “RNA modifications modulate gene expression during development,” *Science*, vol. 361, no. 6409, pp. 1346–1349, Sep. 2018, doi: [10.1126/science.aau1646](https://doi.org/10.1126/science.aau1646).
9. Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, “Integration of A Deep Learning Classifier with A Random Forest Approach for Predicting miRNAs,” *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 6, pp. 451–459, Dec. 2018, doi: [10.1016/j.gpb.2018.08.004](https://doi.org/10.1016/j.gpb.2018.08.004).
10. H. Dweep and N. Gretz, “miRWalk2.0: a comprehensive atlas of microRNA-target interactions,” *Nat Methods*, vol. 12, no. 8, pp. 697–697, Aug. 2015, doi: [10.1038/nmeth.3485](https://doi.org/10.1038/nmeth.3485).

11. S. Rashid Ahmed Ahmed, I. Al Barazanchi, A. Mhana, and H. R. Abdulshaheed, "Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set," *PEN*, vol. 7, no. 2, p. 438, Jun. 2019, doi: [10.21533/pen.v7i2.483](https://doi.org/10.21533/pen.v7i2.483).
12. M. Vidyasagar, "Identifying Predictive Features in Drug Response Using Machine Learning: Opportunities and Challenges," *Annu. Rev. Pharmacol. Toxicol.*, vol. 55, no. 1, pp. 15–34, Jan. 2015, doi: [10.1146/annurev-pharmtox-010814-124502](https://doi.org/10.1146/annurev-pharmtox-010814-124502).
13. C. Huang, R. Mezencev, J. F. McDonald, and F. Vannberg, "Open source machine-learning algorithms for the prediction of optimal cancer drug therapies," *PLoS ONE*, vol. 12, no. 10, p. e0186906, Oct. 2017, doi: [10.1371/journal.pone.0186906](https://doi.org/10.1371/journal.pone.0186906).
14. Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007, doi: [10.1093/bioinformatics/btm344](https://doi.org/10.1093/bioinformatics/btm344).
15. P. Bhuvaneshwari and A. B. Therese, "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm," *Procedia Materials Science*, vol. 10, pp. 433–440, 2015, doi: [10.1016/j.mspro.2015.06.077](https://doi.org/10.1016/j.mspro.2015.06.077).
16. J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen, "Principles of MicroRNA–Target Recognition," *PLoS Biol*, vol. 3, no. 3, p. e85, Feb. 2005, doi: [10.1371/journal.pbio.0030085](https://doi.org/10.1371/journal.pbio.0030085).
17. S. E. McGeary *et al.*, "The biochemical basis of microRNA targeting efficacy," *Science*, vol. 366, no. 6472, p. eaav1741, Dec. 2019, doi: [10.1126/science.aav1741](https://doi.org/10.1126/science.aav1741).
18. A. N. Khan, A. A. Ihalage, Y. Ma, B. Liu, Y. Liu, and Y. Hao, "Deep learning framework for subject-independent emotion detection using wireless signals," *PLoS ONE*, vol. 16, no. 2, p. e0242946, Feb. 2021, doi: [10.1371/journal.pone.0242946](https://doi.org/10.1371/journal.pone.0242946).
19. M. Reczko, M. Maragkakis, P. Alexiou, G. L. Papadopoulos, and A. G. Hatzigeorgiou, "Accurate microRNA Target Prediction Using Detailed Binding Site Accessibility and Machine Learning on Proteomics Data," *Frontiers in genetics*, vol. 2, p. 103, 2011, doi: [10.3389/fgene.2011.00103](https://doi.org/10.3389/fgene.2011.00103).
20. Y. Li *et al.*, "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1070–D1074, Jan. 2014, doi: [10.1093/nar/gkt1023](https://doi.org/10.1093/nar/gkt1023).
21. V. K. Lam, M. Miller, L. Dowling, S. Singhal, R. P. Young, and E. C. Cabebe, "Correction to: Community Low-Dose CT Lung Cancer Screening: A Prospective Cohort Study," *Lung*, vol. 197, no. 5, pp. 685–685, Oct. 2019, doi: [10.1007/s00408-019-00269-6](https://doi.org/10.1007/s00408-019-00269-6).

22. Singh T.R. (2014) Machine Learning with Special Emphasis on Support Vector Machines (SVMs) in Systems Biology: A Plant Perspective. In: P.B. K., Bandopadhyay R., Suravajhala P. (eds) *Agricultural Bioinformatics*. Springer, New Delhi. https://doi.org/10.1007/978-81-322-1880-7_16
23. A. Gupta, S. Chandra, and T. R. Singh, “HLAB27Pred: SVM-based precise method for predicting HLA-B*2705 binding peptides in antigenic sequences,” *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 3, no. 1, p. 56, Apr. 2014, doi: [10.1007/s13721-014-0056-z](https://doi.org/10.1007/s13721-014-0056-z).
24. M. Sehgal, R. Gupta, A. Moussa, and T. R. Singh, “An Integrative Approach for Mapping Differentially Expressed Genes and Network Components Using Novel Parameters to Elucidate Key Regulatory Genes in Colorectal Cancer,” *PLoS ONE*, vol. 10, no. 7, p. e0133901, Jul. 2015, doi: [10.1371/journal.pone.0133901](https://doi.org/10.1371/journal.pone.0133901).
25. A. K. Yadav and T. R. Singh, “Novel structural and functional impact of damaging single nucleotide polymorphisms (SNPs) on human SMYD2 protein using computational approaches,” *Meta Gene*, vol. 28, p. 100871, 2021, doi: <https://doi.org/10.1016/j.mgene.2021.100871>.