Islamic University of Technology (IUT)

Department of Computer Science and Engineering (CSE)

Detection of Lung Adenocarcinoma Cancer based on RNA-seq gene expression data using LIMMA and TabNet

---

Authors

Faysal Bin Rahman, 170041065

Farhan Anjum, 170041045

Musaddiq Hasan Fatin Khan, 170041053

Supervisor	**Co-Supervisor**
Tareque Mohmud Chowdhury	Tasnim Ahmed
Assistant Professor	Lecturer
Department of CSE	Department of CSE

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of B.Sc.

Engineering in CSE

Academic Year: 2020-2021

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Faysal Bin Rahman, Farhan Anjum and Musaddiq Hasan Fatin Khan under the supervision of Tareque Mohmud Chowdhury, Assistant Professor of Department of Computer Science and Engineering (CSE) and Tasnim Ahmed, Lecturer of Department of Computer Science and Engineering(CSE), Islamic University of Technology (IUT), Gazipur, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Authors:

_____

Faysal Bin Rahman

Student ID: 170041065

_____

Farhan Anjum

Student ID: 170041045

_____

Musaddiq Hasan Fatin Khan

Student ID: 170041053

Supervisor:

_____

Tareque Mohmud Chowdhury

Assistant Professor

Department of CSE

Islamic University of Technology

Co-Supervisor:

_____

Tasnim Ahmed

Lecturer

Department of CSE

Islamic University of Technology

Acknowledgement

We would like express our gratitude towards IUT authority for granting us the fund and providing assistance required to implement our proposed system. We are forever indebted to our supervisor, Assistant Professor, Tareque Mohmud Chowdhury and co-supervisor, Lecturer Tasnim Ahmed for providing us with insightful knowledge and guiding us at every stage of our journey. Finally, we would like to express our heartiest appreciation towards our family members for their continuous support, motivation, suggestions and help, without which we could not have achieved the scale of implementation that we have achieved.

Abstract

Lung cancer is one of the deadliest diseases of the world to this date with the highest mortality rate amidst all other forms of cancer. Detection of cancer in early stages is crucial for cancer treatment. Progress in cancer detection has been increasingly made based on gene expression levels, giving insight into making correct and successful treatment decisions, thanks to recent advances in high-throughput sequencing technology such as RNA-seq and the use of several machine learning approaches. However, most of the work on cancer detection uses micro-array data and machine learning models. This paper presents a new methodology based on RNA-seq data which is better at detecting transcripts than micro-array along with Deep Neural Network (Tabnet) to classify human lung cancer.

Contents

1 Chapter 1: Introduction

1.1 Overview

One of the key cause of death in the world is Cancer. In 2020, cancer had caused around 10 million deaths worldwide, accounting for roughly 1 in every 6 deaths. Human body contains trillions of cell, cancer can form anywhere in those cells. Human cells normally grow and multiply through a process known as cell division [1] to form new cells as the body requires them. Cell death [2] occurs when cells become old or damaged, and new cells replace them. This systematic process breaks down from time to time, allowing abnormal. and damaged cells to grow and grow when they should not. These cells coalesce to form a tumor, a mass of tissue. Tumors may be cancerous or non cancerous. Through infestation of nearby tissues cancerous tumors form new tumors, this process is called metastasis. This is the main reason of death of cancer patients. [3] These cancerous tumors are also referenced as malignant tumors. Unlike many cancers, blood cancer(leukemia) do not form solid tumors. [4] The tumors that do not metastasize are the Benign tumors. Removing benign tumors ensure they do not reappear, whereas cancerous tumors every so often do. [5]. Regardless, the benign tumors can grow quite large at times. This can prove to be fatal in some scenarios such as, benign brain tumor. There are almost 291 cancers reported so far. [6]Lung cancer is the leading cause of cancer morbidity and accounts for about 2 million diagnoses and 1.8 million deaths worldwide. As reported by the most recent GLOBOCAN conjecture, 2,094,000 new cases of lung cancer were identified worldwide in 2018, making lung cancer take the position of leading cause of cancer death worldwide. In men, Lung Cancer is only second to prostrate cancer, with an estimated 1,369,000 cases. Similarly, for women it closely follows breast cancer, with 725,000 cases. The accumulated lifetime risk of lung cancer diagnosis after age-normalization, is 3.8 percent for men and and 1.77 percent in women. [7, 8] Histology or cytology can be used to make a pathologic diagnosis of

lung cancer. This review is primarily concerned with the histology of lung cancer. Histologic evaluation for lung cancer diagnosis can be done using using a range of biopsy specimens , in particular, bronchoscopic or needle biopsies. Another procedure is surgical biopsy, such as thoracoscopy, excisional wedge biopsy, lobectomy, or pneumonectomy. In almost all cases, the diagnosis of lung cancer is done using light microscopy, with only a few histologic types requiring histochemical stains or immunohistochemistry. The World Health Organization (WHO) and the International Association for the Study of Lung Cancer (IASLC) have proposed an international standard for histologic classification of lung tumors. [9]Squamous cell carcinoma, adenocarcinoma, small cell carcinoma (SCLC), and large cell carcinoma are the four major histologic types of lung cancer. These types can be subdivided more explicitly into subtypes, as for example the bronchioloalveolar carcinoma (BAC) adenocarcinoma. [9]Adenocarcinoma is accountable for more than 30 percent cases among all other lung cancers.[10]

1.2 Motivation and scope of research

Lung cancer claims the lives of more people each year than breast, colorectal, and prostate cancer combined. One of the main reasons for lung cancer's high mortality rate is that it is diagnosed at later stage than rest of the cancer types. Early lung cancer detection makes it possible to fully treat it, this detection is crucial for survival. Patients with initial stage small lung cancer can cured at a rate of 80 percent to 90 percent.[1]. As the tumor progresses and infests lymphatic tissues or other areas of the body, the chances of it being cured drop dramatically. Therefore, detection of lung cancer is a critical topic in medicine.

Cancer is a congenital disease, any change of how genes control cell functions, to be precise the way they divide and grow is what causes cancer. The technique through which a gene is turned on in cells to construct RNA and different proteins

[1]https://www.cancer.net/blog/2018-06/just-diagnosed-with-lung-cancer-answers-expert

is known as gene expression.By analyzing the RNA, the proteins generated from the RNA, or the functionalities of the proteins in the cells, gene expression can be determined. Gene expression monitoring can help with the accurate detection of deferentially expressed genes and early lung cancer detection. The study and interpretation of differences in gene transcript abundance within a transcriptome is known as differential gene expression. We created a binary classifier that determines whether lung tissue is cancerous or not using differential gene expression levels and a deep neural network.

The following is our contribution to this work:

1. Work on one of the most accurate datasets available for cancer study.
2. Detection of relevant genes that play a part in lung cancer.
3. Development of a classifier that employs relevant genes to determine whether or not a patient has lung cancer.

2 Background study/Literature Review

2.1 Transcriptome

The whole set of messenger RNA, or mRNA molecules expressed from the genes of an organism, is a transcriptome.[11] There are two techniques to analyze the transcriptome. The term "transcriptome" can also refer to the particular cell or tissue generated collection of mRNA transcripts. In contrast to the genome, which is stable, the transcriptome is constantly changing. In fact, the transcriptome of an organism varies depending on a variety of factors, including developmental stage and environmental conditions. RNA Seq and microarray are two techniques used to analyze the transcriptome.

2.2 Transcriptome analysis methods

2.2.1 DNA microarray, a hybridization-based technique

Microarray is one the most contemporary progress in research study of caner; it aids in the pharmacological procedure to treat a variety of diseases, oral lesions inclusive. Significant amount of both previously recorded samples or new samples can be analyzed with the microarray technology.It can also be used to check the presence of a explicit marker in tumors.[12]It is also useful in identifying cancer related chromosome-type abnormality, for example, it is particularly useful for association mapping and the linkage study to segregate chromosomal regions associated with a specific disease. This array can also be used to identify cancer-related chromosomal aberrations, such as allelic imbalance segments identified by loss of heterozygozity.[13]

The foundation of gene microarray technology is the "chip"- a small glass slide with tens of thousands of various DNA sequences.. TThe various strands of DNA are arranged in a tabular structure (in rows and columns), this makes it easy to recognize each fragment by its location on the array. Microarrays are classified into two types: gene expression microarrays and tissue microarrays (TMA). Methods, for instance, reverse trancriptase-polymerase chain reaction(RT-PCR) along with Northern blot permit the use of only a few genes for testing per experiment. Nevertheless, global expression profiling or "microarray" not only examines higher number of genes than previously possibly, it also caters to the fact that it does not get influenced by gene selection. selection.[12]

Tumor formation requires contemporaneous alterations in hundreds of cells as well as genetic variations. Microarray is a blessing to the researchers as it allows concurrent testing of a large number of gene specimens. It is particularly useful in the identification of single-nucleotide polymorphisms (SNPs) and mutations, tumor classification, identification of tumor suppressor target genes, identification of cancer

biomarkers, identification of genes associated with chemoresistance, and drug discovery. For example, we can compare the patterns of gene expression levels in a group of cancer patients and a group of normal patients to identify the gene associated with that specific cancer. Comparative genomic hybridization has been accomplished using gene microarrays. This technique uses fluorescently labeled genomic DNA to detect the presence of gene loss or amplification.[14, 15, 16] Genetic deformity in various types of cancers, such as breast carcinoma,[16], bladder carcinoma[17], fallopian tube carcinoma[18], gastric carcinoma[19], melanoma[20] and lymphoma[21] have been mapped by the array-based comparative genomic hybridization(aCGH). Whereas routine histopathologic examination does not allow for sub-classification, gene expression data can distinguish bunch of cases with substantially diverse results.

2.2.2 RNA-seq, a sequence-based approach

Sequencing in place of microarrays, in a matter of years, the measurement of trasncriptome wide gene expression convincingly shifted. RNA sequencing(RNAseq) provides an open and quantifiable system for portraying transcriptional results on a grand scale, allowing for numerous applications, comprising basic science studies, medical care, and agricultural research.

RNA-seq starts with a biological sample(cells, tissues) from which RNA is extracted. After that, particular protocols are used to separate batches of RNA molecules, for instance, the poly-A selection method for enhancing polyadenylated transcripts or for removing ribosomal RNAs using the ribo-depletion protocol. Following that, reverse transcription converts the RNA to complementary DNA(cDNA) and then cDNA end fragments are bounded with sequencing adaptors. Through PCR amplification RNA-Seq library becomes ready for sequencing.. The process is shown in figure 1.

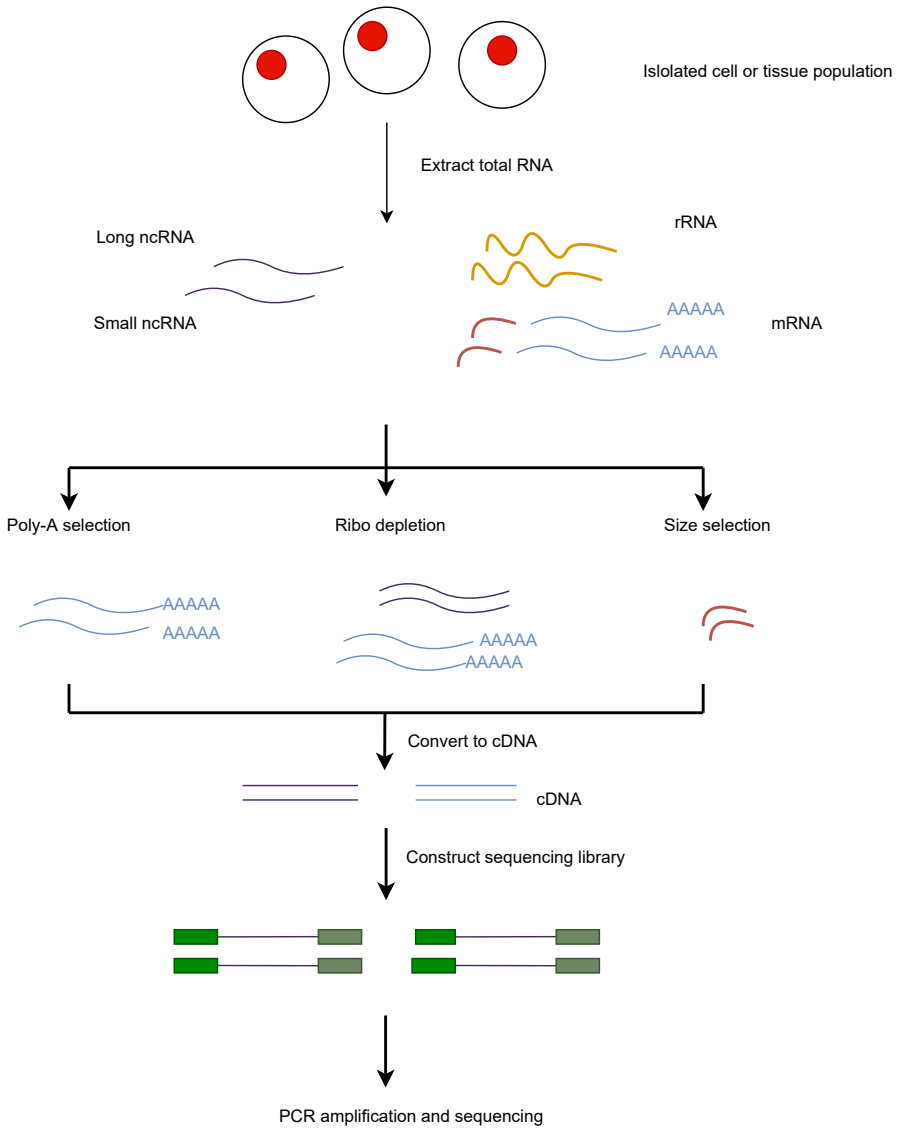

Figure 1: Overview of RNA-Seq.

2.2.3 Comparison between RNA-seq and Microarray

When the target sequences move beyond known genomic sequences, the differences in the capabilities of each technique become clear.Microarray and other hybridization-based approaches are confined within the transcripts in the chips. As excellent as the bioinformatic information for both the genome annd transcriptome of the model oraganism is, the microarray will be the same. Novel sequences, splice variants are detected by RNA-Seq along with elucidated transcripts.[22]. Experimental data can be used by RNA-Seq to characterize exon junctions, detecting ncRNA[23], to detect single nucleotide polymorphisms, and also fusion genes[24]. Furthermore, as new

sequences are annotated, existing data sets can be re-evaluated[25]. Microarrays are capable of detecting single nucleotide polymorphisms, mapping exon junctions, and detecting fusion genes, but only with arrays designed for those purposes. Before non-coding RNA can be accurately distinguished by microarray, it must be annotated and included on specialized chips. To detect non-coding RNA (small RNA) by RNA-Seq, sample preparation procedures must be modified to exclude larger RNA sequences prior to cDNA generation. Finally, microarray chips have to be updated in order to include the latest sequence data, dissimilar to RNA-Seq.

2.2.4 Study on Microarray data classifiers

First paper we found, used leukemia data which had 72 patient samples and color tumor data 62 patient samples. They used Signal to Noise ratio for feature selection and finally performed SVM on the selected genes. The classifier showed an accuracy of 94.1 percent on leukemia dataset and 90.3 percent on colon dataset.[26]

Another paper used lymphoma data which contained 4026 genes and 47 samples and colon data which contained 2000 genes and 62 samples. They used Genetic Algorithm for feature selection and then performed KNN to create a classifier which had 84.6 percent accuracy for lymphoma data and 94.1 percent accuracy for colon data. [27]

Third paper used the same dataset as the first one and second one which are leukemia data with 72 samples, lymphoma data with 47 samples and colon data with 62 samples. They performed Partial Least Square for dimentionality reduction and used Logistic discriminant and Quadratic discriminant analysis as classifiers. The Logistic Discriminant classifer showed an accuracy of 95.9 percent, 96.9 percent and 93.5 percent for Leukemia, Lympho and Colon datasets respectively. Whereas, Quadratic discriminant analysis classifier showed an accuracy of 96.4 percent, 97.4 percent and 91.9 percent for the same datasets.[28, 29]

Another paper used breast cancer data and performed Functional Link Artifical

Neural Network(FLANN) for classification which had an accuracy of 63.34 percent and integrated Particle Swarm Optimization(PSO) with FLANN together to form a classifier which had an accuracy of 92.36 percent.[30]

Another paper used Breast cancer dataset, Colon cancer dataset, Leukemia dataset, Lung cancer dataset and CNS dataset. They performed Best Agglomerative Ranked Subset(BARS) and Fast Correlation-Based Filter(FCBF) for selecting relevant genes. After selecting relevant genes, they used Logistic Regression (LR) and Evolutionary Generalized Radial Basis Function(EGRBF) neural network to create a classifier which had an accuracy of 91.08 percent.[31]

Another paper used lung carcinoma dataset which had 254 sample each having 8359 genes, lukemia dataset having 72 samples and each sample having 8359 genes and finally blue cell tumor dataset having 83 samples each having 2308 gene expression level. They performed partial least square for selecting relevant gene and used SVM, LDA as classifier which had accuracy of 95.5 and 98.0 percent respectively.[32]

Another paper proposed using Bayes classification for gene selection and classification which was experimented on lung cancer dataset resulting in an accuracy of 100 percent, and an accuracy of 96.3 percent on breast cancer dataset.[33]

2.2.5 Study on RNA-seq data classifiers

Another paper used three set RNA-seq data basically LUAD cancer, STAD cancer and BRCA. They used DESeq for finding out the differentially expressed genes. After selecting the DEG's they used 5-fold stacking of kNN, SVM, DT, RF and GBDT to create a multi-model ensemble which had an accuracy of 98.8 percent for LUAD, 98.78 percent for STAD and 98.41 percent for BRCA.[34]

3 Dataset

The Genomic Data Commons (GDC) data portal provided us with the data that we required[2]. The dataset included the raw RNA Sequence gene expression data for LUAD cancer. There were 592 samples in the dataset, with 533 being primary tumors, two being recurrent tumors, and 59 being solid tissue normal (the control). Raw read counts for 26208 genes were included, as well as 88 other important details about the patients, such as their age, gender, smoking status, and so on.

The dataset was found in SummarizedExperiment format. The SummarizedExperiment class stores rectangular matrices of experimental findings, which are often generated by sequencing and microarray studies. Each object contains observations of one or more samples, as well as meta-data that describes both the observations (features) and the samples (phenotypes). The synchronization of meta-data and assays while subsetting is a crucial feature of the SummarizedExperiment class. If we want to exclude a certain sample, for example, we can do so for both the meta-data and the assay in one operation, ensuring that the meta-data and observed data are in sync. This is a particularly desired quality because improperly accounting for meta and observational data has led in a lot of inaccurate results and retractions.SummarizedExperiment is a matrix-like container with rows representing features of interest (genes, transcripts, exons, and so on) and columns representing samples. Each assay is represented as a matrix-like object in numeric or other mode in the objects. A SummarizedExperiment object's rows represent the features that are of interest.

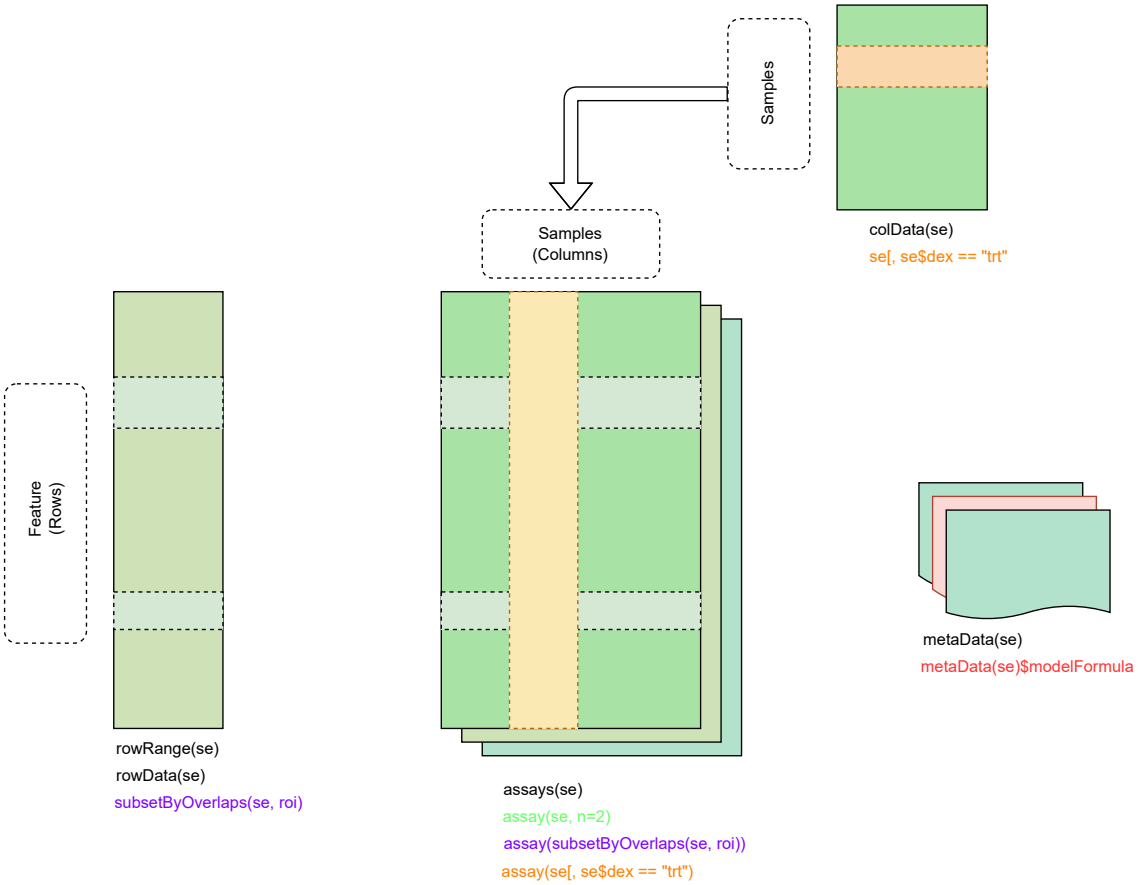

Figure 2: Structure of SummarizedExperiments

4 Methodology

4.1 Obtaining relevant genes and read counts of relevant genes

To ascertain which genes are more relevant to cause lung cancer Differential Gene Expression (DGE) was used.For statistical genomics, two packages of Bioconductor [35] which is an 'R' based open-source software, were considered for finding DGEs. The packages considered were DESeq2 and LIMMA.

4.1.1 DESeq2

RNA-seq data analysing for differential gene expression is feasible using the DESeq2 tool. DESeq2 is a package that utilizes negative binomial generalized linear models to check for differential expressions. It estimates priors for log fold change and

[2]https://portal.gdc.cancer.gov/

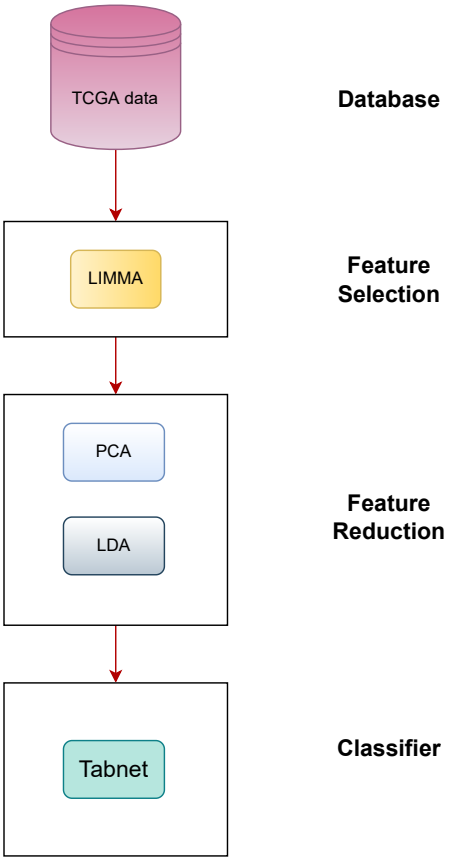

Figure 3: Brief overview of methodology

dispersion, as well as posterior estimates for these quantities, using empirical Bayes techniques [36].

Multiple steps are involved in differential expression analysis with DESeq2, as shown in the figure 4 below. DESeq2 will model the raw counts using normalization factors (size factors) to account for library depth discrepancies. It will then estimate gene-wise dispersion and decrease these estimates to obtain more accurate dispersion estimates to simulate the counts. Finally, using the Wald test or the Likelihood Ratio Test, DESeq2 will fit the negative binomial model and perform hypothesis testing.

4.1.2 LIMMA

LIMMA is an R package for analyzing RNA sequencing data for differential expression. Following pre-processing as well as normalization, the package is set up so that data from all technologies can be analyzed using the same analysis pipeline. All

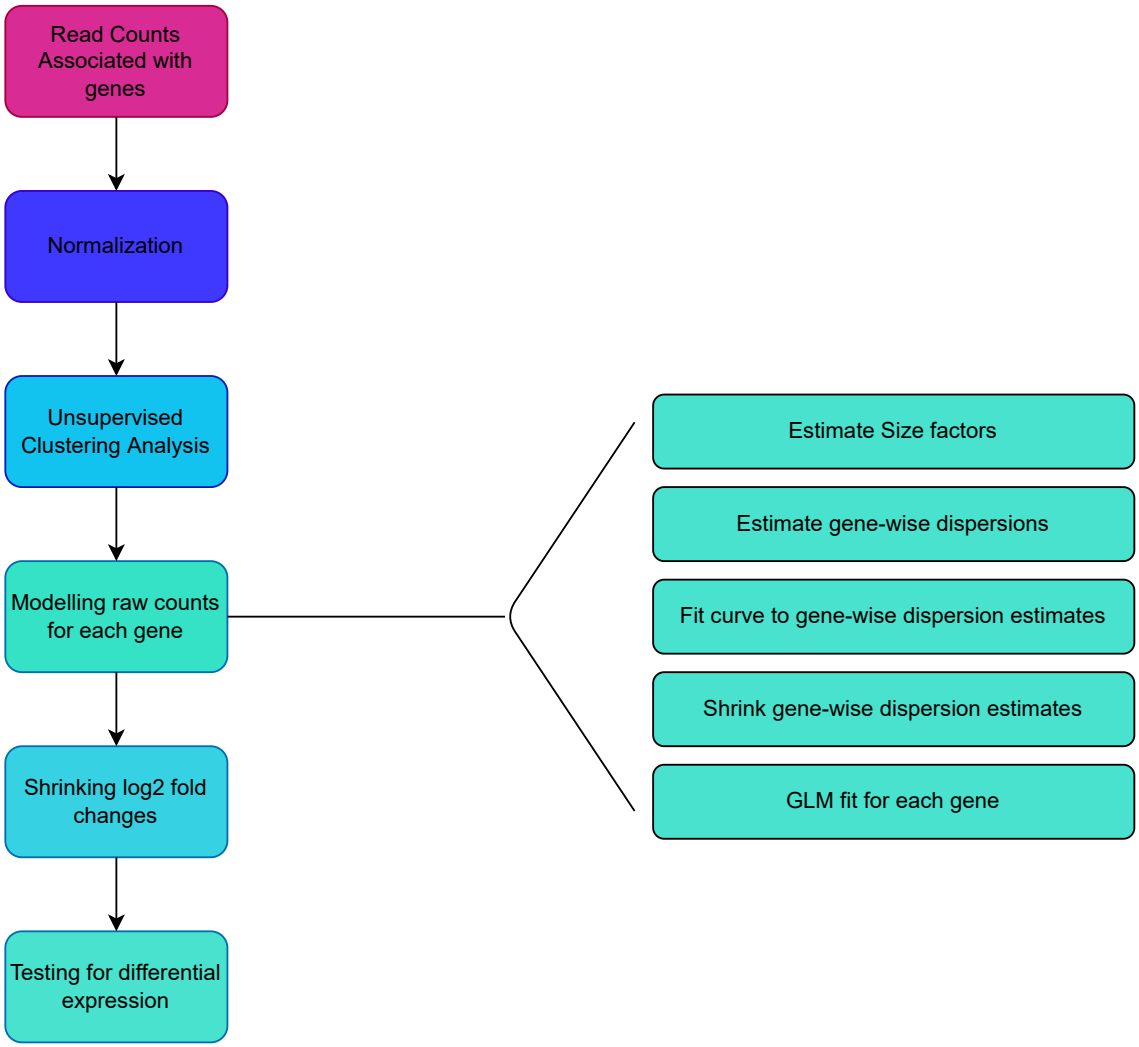

Figure 4: DESeq2 working principles

downstream tools that were previously available only for microarray data are equally available for RNA-seq data too thanks to LIMMA. [37]. We found that LIMMA was more precise in detecting genes that are more differentially expressed [38]. Therefore, we chose LIMMA for our work. With the help of the LIMMA library, we were able to identify the top 100 differentially expressed genes. We were also able to get normalized read counts of the relevant 100 genes for all of our samples with the help of LIMMA. These normalized read counts were used to train our machine-learning model.

LIMMA effectively incorporates a number of statistical principles for comprehensive expression research. It works with expression values represented in a matrix

format. Each row refers to a gene or a genetic trait appropriate to the present study. Whereas, columns refer to samples of RNA. On the one hand, LIMMA accommodates a linear model in every row of data thus using the flexibility of the model in a variety of manners, hypotheses that are highly acquiescent.Conversely, the extremely parallel structure of the gene data is leveraged to adopt potency from gene-specific samples, this allows variableness among genes also among samples thus increasing the the reliability of statistical inferences considering there is small number of samples. So, LIMMA package includes statistical methods that

- Allow information adaptation employing the empirical Bayes procedure to procure posterior variance estimators;
- Incorporate the observed weights to justify discrepancies in data traits;
- Allow variance modelling to account for methodological even biological heterogeneity;
- Pretreatment technique to reduce noise, for instance, dispersion stabilization.

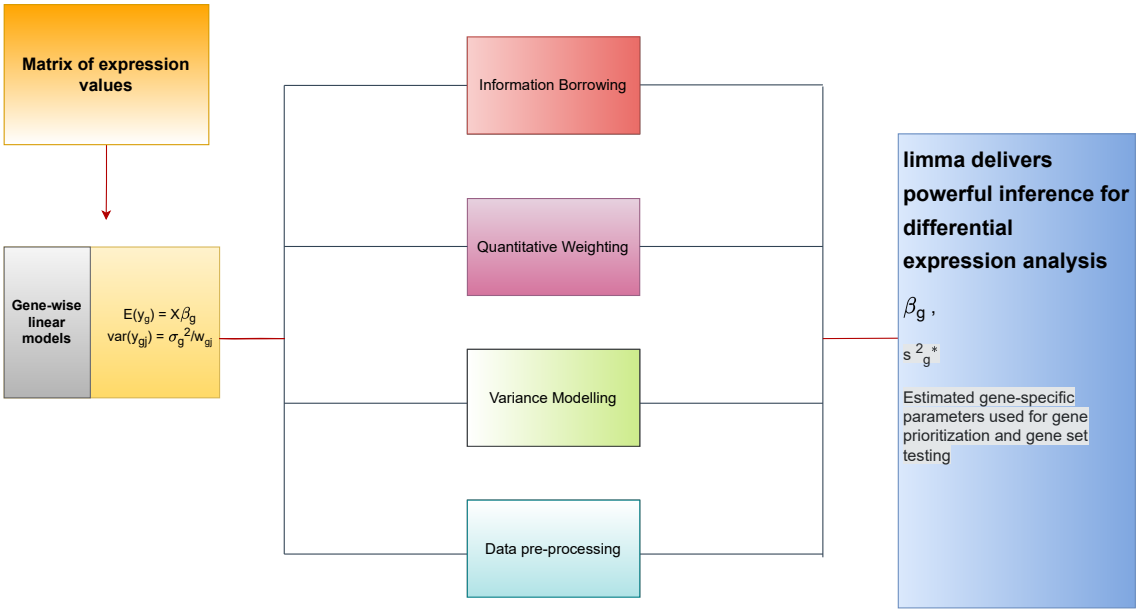

Figure 5: LIMMA Working Principles

Table 1: Top 20 DEG's after using LIMMA on LUAD dataset

external_gene_name	logFC	AveExpr	t	P.Value	adj.P.Val	B
SLC6A4	-7.66278056220821	-7.66278056220821	-39.5943360911668	6.77083924428858e-169	1.77450154914315e-164	375.709084553734
AL606469.1	-5.93088026783627	-5.93088026783627	-36.6768156365474	7.41536657771113e-155	9.71709636343266e-151	342.7687871881
SERTM1	-7.28837535777598	-7.28837535777598	-35.2448377834483	8.38255042671611e-148	7.32299605277919e-144	327.092974502947
AC095050.1	-4.94432947433946	-4.94432947433946	-34.8039014672513	1.3061178573449e-145	8.55768420132377e-142	321.131049715356
FABP4	-6.18288771340686	-6.18288771340686	-34.7244772632855	3.2504743356942e-145	1.70376862779747e-141	321.360724356119
ITLN2	-7.34232500720104	-7.34232500720104	-33.6741311110166	5.98499839255015e-140	2.6142472978659e-136	309.20407336794
GPM6A	-5.93088707374936	-5.93088707374936	-33.1922425747045	1.62218100942596e-137	6.07344569929081e-134	303.663894182379
RTKN2	-4.52761770637662	-4.52761770637662	-32.8227806384383	1.20979608778243e-135	3.96329198357524e-132	299.386569171872
LINC02016	-5.97876260090301	-5.97876260090301	-32.4613553993777	8.32760192541761e-134	2.42499768068161e-130	294.604436160873
STX11	-3.27506315749025	-3.27506315749025	-31.6880609492471	7.4429260760633e-130	1.95064206601467e-126	286.078958203958
CD300LG	-6.34162649098624	-6.34162649098624	-31.1398692690256	4.8748630580548e-127	1.16145828205e-123	279.50410352453
FAM107A	-5.00890272636113	-5.00890272636113	-31.0704428793236	1.11043458539587e-126	2.42518913450459e-123	278.789546143889
AL354714.1	-5.40820129177685	-5.40820129177685	-31.0129135734276	2.19735966439084e-126	4.42987708341194e-123	277.578606910039
AC128709.3	-4.69443829476841	-4.69443829476841	-30.9512918105351	4.56604102654795e-126	8.54762880169777e-123	276.60656229849
LINC01996	-5.80698975892672	-5.80698975892672	-30.5898561743343	3.3542033572819e-124	5.86046410584294e-121	272.760535854988
TEK	-3.48853838937666	-3.48853838937666	-30.4892925661661	1.11090467323096e-123	1.81966185475231e-120	271.895560105551
SCUBE1	-4.3268074881509	-4.3268074881509	-30.025876087029	2.80019419626607e-121	4.3169111468083e-118	266.374588604504
UPK3B	-5.79774705180907	-5.79774705180907	-29.8952232768386	1.33556605841655e-120	1.9445841810545e-117	264.81933617824
LGI3	-6.8249958028128	-6.8249958028128	-29.6910268251678	1.53903855696842e-119	2.12290118426466e-116	262.379780249971

4.2 Dimensionality reduction

The normalized read count of 100 relevant genes along with patient sex and age were considered to train the machine learning model. A number of dimensionality reduction techniques were considered for compressing the number of features. Two techniques were used in our work, Principle Componenet Analysis(PCA) along with Linear Discriminant Analysis(LDA).

4.2.1 Principal Component Analysis

PCA is a multi-faceted method for analyzing tabular data where observations are defined as set of interconnected quantifiable response variables. Its objective is to derive the key value present in the table, and showcase it as a group of new normal elements, called principal components.Afterwards, exhibit similar patterns between the observed values and the variables as if they are points on a map. [39]. The objectives of PCA are to:

- Determine major details from the tabular data;
- Reduce the amount of data package by retaining solely the major detail;
- Make portrayal of the data package more concise;
- Examine observations' as well as variables' configuration.

PCA estimates fresh variables, known as priciple components, these are derived as linear composition of the initial variables. The first principal component demands the greatest likely deviation (inertia), as a result, this factor will “explain" or “extract" the majority of the data table’s inertia. The second component is computed with the requirements of being orthogonal to the first and having the greatest achievable inertia. The remaining constituents are calculated in the same way. Factor scores are the geometric interpretations of the observations’ projections upon the primary constituent, the resultant values are the latest variables for the data.

4.2.2 Linear Discriminant Analysis

The dimensionality reduction method employed to solve problems with supervised classification is Linear Discriminant Analysis (LDA). Class variations are modelled with the LDA, for instance, multi-class division. LDA projects higher dimensional features to lower-dimensional space. The case where internal frequencies of a class are unequal, as well as, their behavior examined by generated test data is at random, is easily addressed using Linear Discriminant Analysis. Peak seperability is ensured by maximizing the relative size of in class variance to within-class variance. [40]. The objectives of LDA are:

- Determine variance between the classes, which is seperation between distinct classes (i.e the mean difference between seperate classes)
- Determine mean square deviation within-class, which is the difference of mean with each class samples.
- Construct a lower-dimensional space, it has to maximize mean square deviation between the classes while minimizing it within-class.

4.2.3 Using PCA and LDA for dimensionality reduction

Because of various challenges, it is widely assumed that it is impractical to directly use LDA solution for large data. [41]. As a result, before LDA can be used, another procedure must first be used to reduce dimensionality. We used Principal Component Analysis (PCA) and Linear discriminant analysis(LDA) to trim the dimension as these two methods are popular and widely used together [42]. In order to regularize the input and avoid over-fitting we performed Principal Component analysis first. We then used Linear Discriminant Analysis to gain the required low dimensional features and data to train our machine learning model. The most significant distinction between LDA and PCA is that PCA focuses on feature classification while LDA focuses on data classification. On one hand, using PCA the configuration of the original data sets change when the data sets are transformed to another space. On the other hand, LDA only allows for class distinction along with drawing a decision boundary among the classes [40].

4.3 Classifier network

We considered decision tree (DT) and deep neural network (DNN) for our classifier. DT is predictive pattern that utilizes ramification series of Boolean tests to make additional summarized conclusions premised on evidences. A DNN is a neural network with a high level of complexity, usually at least two layers. DNNs were chosen over DTs primarily because they perform better with large datasets [43]. Furthermore, unlike tree learning, DNNs allow for end-to-end learning for tabular data using gradient descent, which is advantageous [44]. Deep learning's ability to conceal unrefined data inside significant depictions is a prime reason for its triumph in in images, natural language, and audio. Backpropagation algorithm-based end-to-end training can effectively encrypt delineated data, thus shortening or completely removing the demand for feature handling. [45].

We used the Attentive Interpretable Tabular Learning neural network (TabNet) to create our classifier because our dataset is tabular. From a wide range of tabular datasets based on non-performance saturation, Tabnet surpasses other DNN variants. [44]. TabNet outperforms DTs while benefiting from their advantages because of careful design that:

- employs data-driven sparse instance-wise feature selection;
- develops a multi-step sequential architecture, with each stage appertaining to the chosen features contributes a part in ruling.;
- progresses the learning ability by processing selected features in a nonlinear way;
- does ensembling with higher dimensions and more steps [44].

4.3.1 Tabnet Encoder

The TabNet encoder design is primarily made up of some common elements at each decision step, they are the feature transformer, attentive transformer, as well as feature masking, as shown in Figure 6. The tabular data consists of both categorical and numerical data. TabNet maps category features to numerical features using initial statistical data along with educable embedding [46]. Each stage of decision uses identical $B \times D$ feature matrix, here B represents the batch volume and D represnets the dimension of the features. Multilevel sequential manipulation is followed in Tabnet's ciphering. The ith step takes treated data from its previous $(i-1)$th step, and and uses it to determine the features that need to be employed and also yields the refined feature model, which is then accumulated into the entire decision. Top-down attention in successive format was inspired when it was needed to look for small section of appropriate evidence from multifaceted entry, during the application of the same thing in processing visual and text data [47], and reinforcement learning [48].

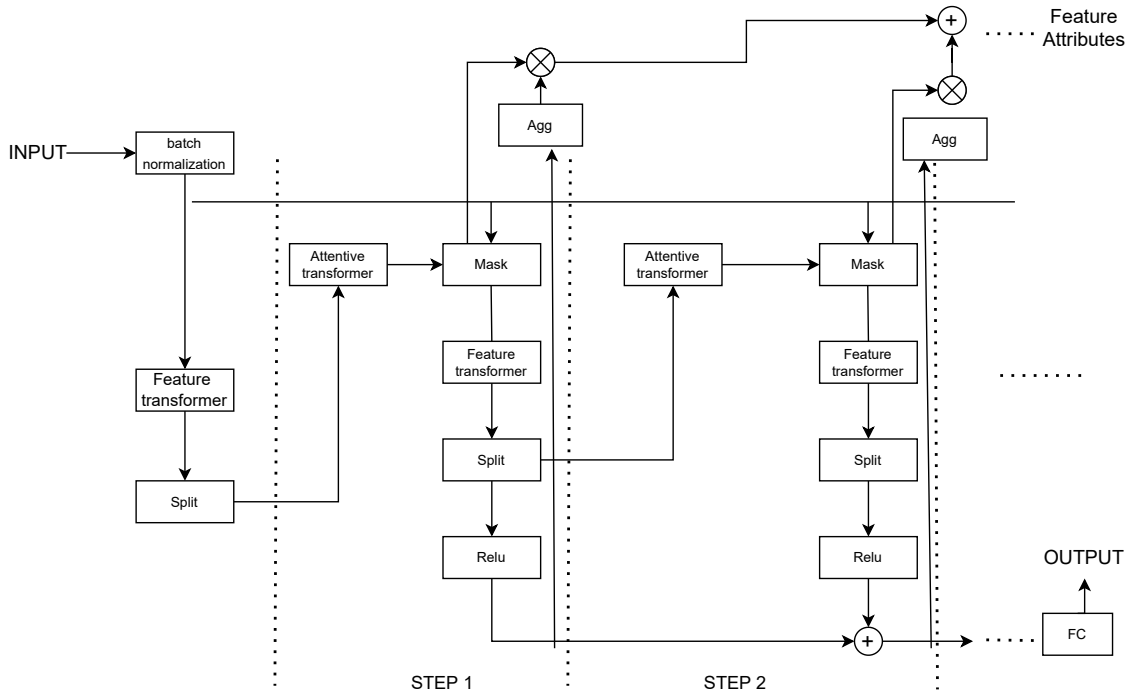

Figure 6: Tabnet Encoder Architecture

4.3.2 Feature Selection

Each decision step's Mask module is responsible for feature selection. The decision step's Attentive transformer determines the function to be implemented. The following steps occur in the Attentive Transformer Layer:

1. The preceding decision step's Feature transformer produces the tensor, afterwards transmits it to the Split unit.

2. In step 1 of the Split module, the tensor is split and a[i - 1] is obtained.

3. The h_i level constitutes of fully connected (FC) layer along with a BN layer, it is traversed by a[i - 1]. The purpose of h_i is to attain a linear amalgamation of features, which enables extraction of multi-faceted as well as conjectural features.

4. The preceding value p[i - 1] of the preceding judgment step is multiplied with the result obtained from the h_i level. The preceding scale depicts how features were used in previous decision-making phases. With the increments in the

quantity of features employed in the prior decision step the weight in the current decision step decrements.

5. Later Sparsemax [49] is employed to construct M[i]. This method to learn a mask is portrayed by the equation:

$$M[i] = Sparsemax(P[i-1] \times h_i(a[i-1]))$$

Sparsemax foments sparsity thus making feature selection more sparse through shifting the Euclidean projection atop the stochastic simplex. If D is the dimension of the features then Sparsemax can make $\sum_{j=1}^{D} M[i]_{b,j} = 1$. Sparsemax sets for all characteristics of every sample's total weights as 1, along with perfroming weight allocation of each feature j, of every sample b. This enables Tabnet to exercise the primary features for the model in all decision steps. The sparsity of the specified characteristics are adjusted using the following sparese regular term:

$$L_{sparse} = \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{j=1}^{D} \frac{-M_{b,j}[i]}{N_{steps} times B} log(M_{b,j}[i] + \epsilon)$$

The sparsity of feature selection can give superior inductive bias for convergence to a higher accuracy rate when the majority of the data set's features are redundant.

6. To update P[i] the following equation is used:

$$P[i] = \prod_{j=1}^{i} (r - M[j])$$

7. The feature selection of the ongoing phase is achieved by the product of M[i] and feature components.

8. The selected features are then sent into the current decision step's feature

transformer, and a spiral of fresh decision step is initiated.

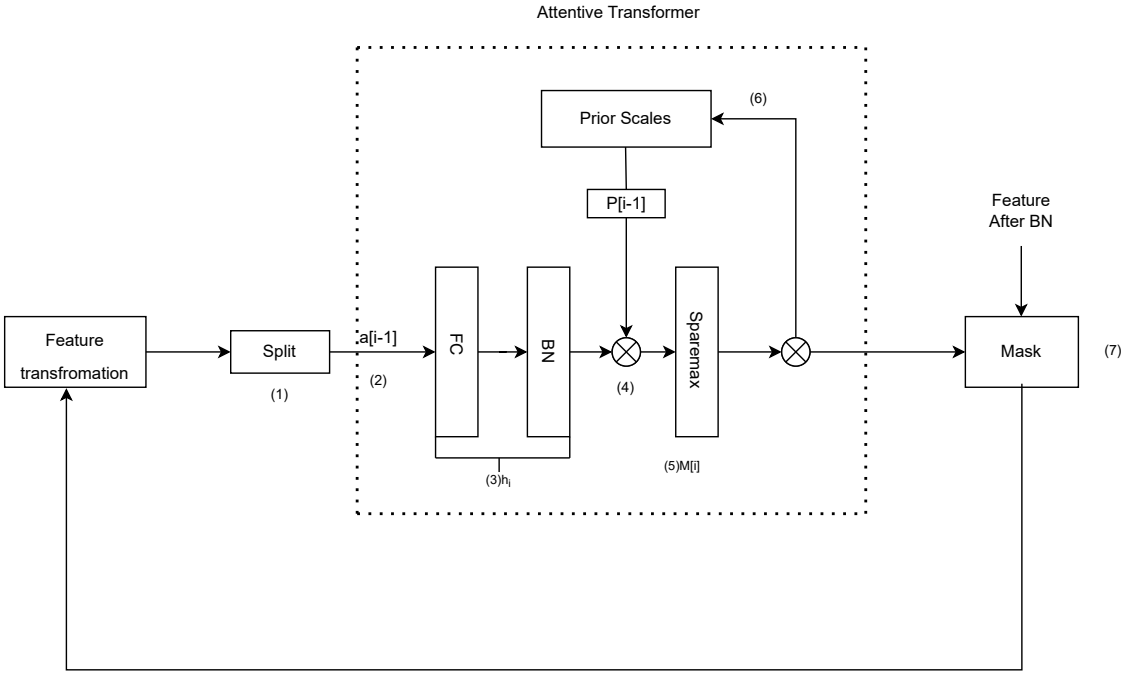

Figure 7: Structure of Attentive Transformer Layer

4.3.3 Feature Processing

Mask filters the features and passes them to the feature transformer layer in feature processing. The split module divides the processed features into two parts: one functions as the result of the working decision step and the remaining one is utilized as entry data for the following decision step. It is shown as:

$$[d[i], a[i]] = f_i(M[i] \times f)$$

Feature transformer layer has 3 layers of its own, they are : BN, gated linear unit as well as FC layers. Figure 8 depicts the structure. The Feature transformer layer is divided into two portions as can be observed.The initial one-half portion of the parameters are reciprocal, it implies that in all steps they are trained together.However, for the other half parameters are not reciprocal, also in all steps they are trained individually. Because the input for every stage are the common features, we may

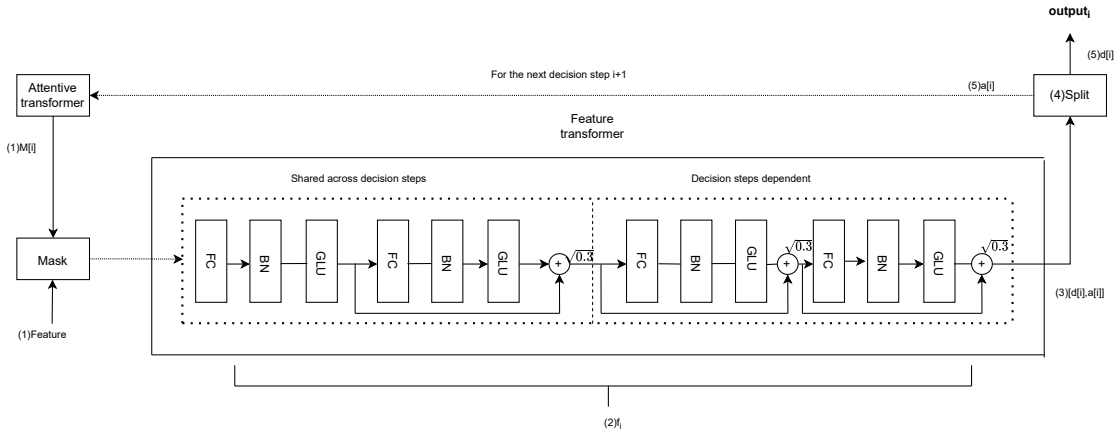

Figure 8: Structure of Feature Transformer Layer

employ the identical layer for the general component of the computational features' as well as distinct layers for the feature section of every step. Because of this configuration, this model shall be able to learn in a robust and high-capacity manner. In the layer, the residual connection is employed, and to guarantee network soundness it is multiplied with $\sqrt{0.5}$.

4.3.4 Tabnet Decoder

Figure 9 shows the encoded form without the FC layer, which is the encoders' vector sum. The decoder receives the encrypted depiction as input. The decoder reconstructs the representation vector into a feature using the Feature transformer layer. We produced the reconstructed feature after adding many steps.

5 Results and Discussion

5.1 Experimental Setup

R along with python were used in our work. R studio was used to operate LIMMA in order to obtain top 100 differentially expressed genes as well as their normalized values. After getting the read counts of 100 genes of all samples, we shifted to train our model in Python environment in Google Colab. Our samples were split

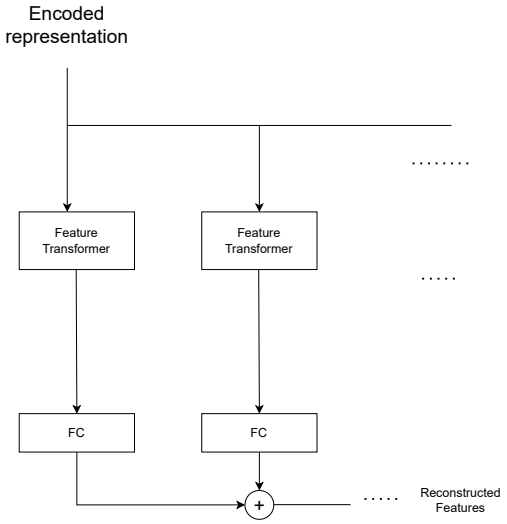

Figure 9: Tabnet Decoder Architecture

randomly into 80% for training and 20% for testing.

5.2 Evaluation Metrices

5.2.1 Accuracy

The percentile representation of the ratio of all true predictions to all predictions made is called Accuracy. If N is the overall amount of correct predictions made, with M representing the overall amount of samples, then accuracy can be defined as:

$$Accuracy = \frac{N}{M} \times 100\%$$

The accuracy of a model is calculated using samples from the test set, which are not visible to the model during training. It provides a better estimate of the model's generalization capability.

5.2.2 Precision

The amount of true positives divided by the cumulative total of positive predictions, accross all classes is precision. Precision is utilized in a multiclass problem to evaluate the accurately classified samples of each class among all the samples that were

categorized as belonging to that class. Positive Predictive Value(PPV) is another name for precision. f X is the amount of samples properly labeled in a class C, with Y representing the amount of samples incorrectly classified within C then precision of C is:

$$Precision_C = \frac{X}{X+Y}$$

5.2.3 Recall

The amount of true positives divided by the cumulative total of true positives and false negative predictions, across the classes is known as recall. In a multiclass problem, recall is used to determine how many samples were correctly classified out of all those that should have been. Sensitivity is another name for recall.If X is the number of samples correctly classified in a class C and Y is the number of samples incorrectly classified in the other class then recall of C can be defined as:

$$Recall_C = \frac{X}{X+Y}$$

5.2.4 ROC-AUC

For measuring performance of classification tasks at different cut-off values, the AUC-ROC curve is used. On one hand, AUC portrays extent or standard of disconnectedness, on the other hand, a feasibility curve is represented by ROC. It demonstrates the degree of effectiveness of the model to differentiate among the classes. AUC indicates how well 0s and 1s can be predicted as 0s and 1s by the model. By way of analogy, patients with or without disease can be discerned by the model with the higher AUC count.

To ROC curve is plotted by putting True Positive Rate (TPR) on the vertical (y) axis, along with the False Positive Rate (FPR) on the horizontal (x) axis.

5.2.5 Specificity

The amount of true negatives divided by the cumulative total to true negatives and false positive predictions, across all classes is known as Specificity. It is determined by the following way:

$$specificity = \frac{X}{X + Y}$$

Here, X indicates the amount of true negatives, and Y represents the amount of false positives.

5.2.6 Results

Our model could accurately predict between cancer and non-cancer 97.48 percent of the time. Precision of our model is 98.61 percent, recall is 97.22 percent and Specificity is 100 percent.

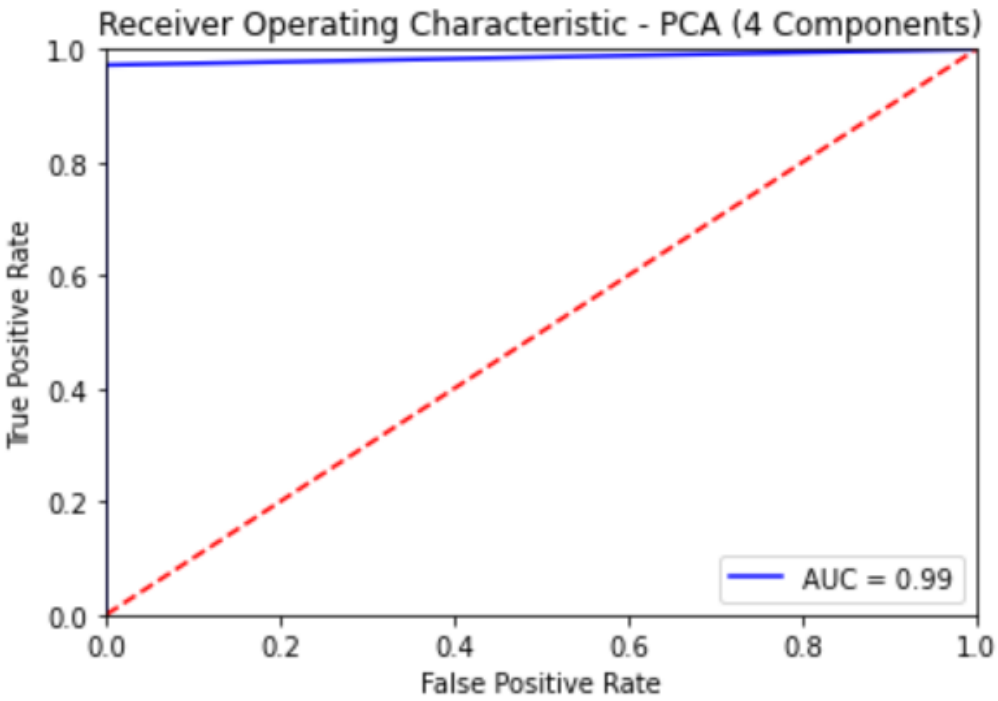

Figure 10: ROC-AUC curve

6 Conclusion

Cancer is a major public health issue all over the world. Many machine learning models have been used for cancer prediction and all of them performs almost equally. Here in this paper, we used a deep learning model for cancer prediction. We used gene expression data obtained from lung tissue. We used the top hundred differentially expressed genes obtained from normal and tumor phenotypes using LIMMA technique as feature for our deep learning model. In order to prevent overfitting in our model, we used PCA and LDA to reduce the dimensionality of our data. Finally, we used TabNet as our classifier. The accuracy of our classifier is almost equal to other classifiers, which used gene expression data of thousands of genes. We achieved almost similar results using data of only hundred genes.

References

[1] J. M. Scholey, I. Brust-Mascher, and A. Mogilner, "Cell division," *Nature*, vol. 422, no. 6933, pp. 746–752, 2003.

[2] R. S. Hotchkiss, A. Strasser, J. E. McDunn, and P. E. Swanson, "Cell death," *New England Journal of Medicine*, vol. 361, no. 16, pp. 1570–1583, 2009.

[3] T. R. Geiger and D. S. Peeper, "Metastasis mechanisms," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1796, no. 2, pp. 293–308, 2009.

[4] M. Yan and P. Jurasz, "The role of platelets in the tumor microenvironment: from solid tumors to leukemia," *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, vol. 1863, no. 3, pp. 392–400, 2016.

[5] M. A. Furlong, J. C. Fanburg-Smith, and M. Miettinen, "The morphologic spectrum of hibernoma: a clinicopathologic study of 170 cases," *The American journal of surgical pathology*, vol. 25, no. 6, pp. 809–814, 2001.

[6] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature reviews cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[7] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[8] "Global cancer observatory: Cancer today. international agency for research on cancer. lyon, france." `https://gco.iarc.fr/today`. Accessed: 2022-04-21.

[9] W. D. Travis, T. Colby, B. Corrin, Y. Shimosato, and E. Brambilla, *Histological typing of lung and pleural tumours*. Springer Science & Business Media, 2012.

[10] W. D. Travis, L. B. Travis, and S. S. Devesa, "Lung cancer," *Cancer*, vol. 75, no. S1, pp. 191–202, 1995.

[11] V. S. Gomase and S. Tagore, "Transcriptomics," *Current drug metabolism*, vol. 9, no. 3, pp. 245–249, 2008.

[12] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, "Microarray and its applications," *Journal of pharmacy & bioallied sciences*, vol. 4, no. Suppl 2, p. S310, 2012.

[13] K. Lindblad-Toh, D. M. Tanenbaum, M. J. Daly, E. Winchester, W.-O. Lui, A. Villapakkam, S. E. Stanton, C. Larsson, T. J. Hudson, B. E. Johnson, *et al.*, "Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays," *Nature biotechnology*, vol. 18, no. 9, pp. 1001–1005, 2000.

[14] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown, "Genome-wide analysis of dna copy-number changes using cdna microarrays," *Nature genetics*, vol. 23, no. 1, pp. 41–46, 1999.

[15] J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, and P. O. Brown, "Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12963–12968, 2002.

[16] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Döhner, T. Cremer, and P. Lichter, "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances," *Genes, chromosomes and cancer*, vol. 20, no. 4, pp. 399–407, 1997.

[17] J. A. Veltman, J. Fridlyand, S. Pejavar, A. B. Olshen, J. E. Korkola, S. DeVries, P. Carroll, W.-L. Kuo, D. Pinkel, D. Albertson, *et al.*, "Array-based comparative genomic hybridization for genome-wide screening of dna copy number in bladder tumors," *Cancer research*, vol. 63, no. 11, pp. 2872–2880, 2003.

[18] A. M. Snijders, M. E. Nowee, J. Fridlyand, J. M. Piek, J. C. Dorsman, A. N. Jain, D. Pinkel, P. J. Van Diest, R. H. Verheijen, and D. G. Albertson, "Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing ccne1 in fallopian tube carcinoma," *Oncogene*, vol. 22, no. 27, pp. 4281–4286, 2003.

[19] M. M. Weiss, A. M. Snijders, E. J. Kuipers, B. Ylstra, D. Pinkel, S. G. Meuwissen, P. J. van Diest, D. G. Albertson, and G. A. Meijer, "Determination of amplicon boundaries at 20q13. 2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 200, no. 3, pp. 320–326, 2003.

[20] P. G. Buckley, K. K. Mantripragada, M. Benetkiewicz, I. Tapia-Páez, T. Diaz de Ståhl, M. Rosenquist, H. Ali, C. Jarbo, C. De Bustos, C. Hirvelä, *et al.*, "A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications," *Human Molecular Genetics*, vol. 11, no. 25, pp. 3221–3229, 2002.

[21] J. A. Martinez-Climent, A. A. Alizadeh, R. Segraves, D. Blesa, F. Rubio-Moscardo, D. G. Albertson, J. Garcia-Conde, M. J. Dyer, R. Levy, D. Pinkel, *et al.*, "Transformation of follicular lymphoma to diffuse large cell lymphoma is associated with a heterogeneous set of dna copy number and gene expression alterations," *Blood, The Journal of the American Society of Hematology*, vol. 101, no. 8, pp. 3109–3117, 2003.

[22] B. E. Howard, Q. Hu, A. C. Babaoglu, M. Chandra, M. Borghi, X. Tan, L. He, H. Winter-Sederoff, W. Gassmann, P. Veronese, *et al.*, "High-throughput rna sequencing of pseudomonas-infected arabidopsis reveals hidden transcriptome complexity and novel splice variants," *PLoS One*, vol. 8, no. 10, p. e74183, 2013.

[23] K. B. Arnvig, I. Comas, N. R. Thomson, J. Houghton, H. I. Boshoff, N. J. Croucher, G. Rose, T. T. Perkins, J. Parkhill, G. Dougan, *et al.*, "Sequence-based analysis uncovers an abundance of non-coding rna in the total transcriptome of mycobacterium tuberculosis," *PLoS pathogens*, vol. 7, no. 11, p. e1002342, 2011.

[24] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan, "Transcriptome sequencing to detect gene fusions in cancer," *Nature*, vol. 458, no. 7234, pp. 97–101, 2009.

[25] A. Roberts, L. Schaeffer, and L. Pachter, "Updating rna-seq analyses after re-annotation," *Bioinformatics*, vol. 29, no. 13, pp. 1631–1637, 2013.

[26] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[27] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method," *Bioinformatics*, vol. 17, no. 12, pp. 1131–1142, 2001.

[28] D. V. Nguyen and D. M. Rocke, "Classification of acute leukemia based on dna microarray gene expressions using partial least squares," in *Methods of Microarray Data Analysis*, pp. 109–124, Springer, 2002.

[29] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[30] J. Dev, S. K. Dash, S. Dash, and M. Swain, "A classification technique for microarray gene expression data using pso-flann," *International Journal on Computer Science and Engineering*, vol. 4, no. 9, p. 1534, 2012.

[31] A. Castaño, F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutiérrez, "Neuro-logistic models based on evolutionary generalized radial basis function for the microarray gene expression classification problem," *Neural processing letters*, vol. 34, no. 2, pp. 117–131, 2011.

[32] S. Student and K. Fujarewicz, "Stable feature selection and classification algorithms for multiclass microarray data," *Biology direct*, vol. 7, no. 1, pp. 1–20, 2012.

[33] A. Sharma and K. K. Paliwal, "A gene selection algorithm using bayesian classification approach," *American Journal of Applied Sciences*, vol. 9, no. 1, pp. 127–131, 2012.

[34] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Computer methods and programs in biomedicine*, vol. 153, pp. 1–9, 2018.

[35] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, pp. 1–16, 2004.

[36] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.

[37] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[38] Y. Tong, "The comparison of limma and deseq2 in gene analysis," *E3S Web of Conferences*, vol. 271, p. 03058, 01 2021.

[39] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[40] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

[41] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional data—with application to face recognition," *Pattern recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[42] J. Yang and J.-y. Yang, "Why can lda be performed in pca transformed space?," *Pattern recognition*, vol. 36, no. 2, pp. 563–566, 2003.

[43] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017.

[44] S. O. Arık and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *AAAI*, vol. 35, pp. 6679–6687, 2021.

[45] J. Yan, T. Xu, Y. Yu, and H. Xu, "Rainfall forecast model based on the tabnet model," *Water*, vol. 13, no. 9, p. 1272, 2021.

[46] M. Grbovic and H. Cheng, "Real-time personalization using embeddings for search ranking at airbnb," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 311–320, 2018.

[47] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," *arXiv preprint arXiv:1803.03067*, 2018.

[48] A. Mott, D. Zoran, M. Chrzanowski, D. Wierstra, and D. J. Rezende, "S3ta: A soft, spatial, sequential, top-down attention model," 2018.

[49] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *International conference on machine learning*, pp. 1614–1623, PMLR, 2016.