EFFICIENCY OF USING BINARY GENE EXPRESSION DATA IN LUNG CANCER SUBGROUPING

UNDERGRADUATE RESEARCH THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF BACHELOR OF THE SCIENCE OF ENGINEERING

Submitted by:

Atputharavi.R (2020/E/015)

Divyani C.P (2020/E/039)

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF JAFFNA

[SEPTEMBER] [2024]

EFFICIENCY OF USING BINARY GENE EXPRESSION DATA IN LUNG CANCER SUBGROUPING

Supervisor(s):

Supervisor Name1 : Dr. P. Jeyananthan

Supervisor Name2 : ……

Examination Committee:

Lecturer 1 .....

Lecturer 2 .....

CONTRIBUTION TO THE PROPOSAL BY THE MEMBERS IN GROUP

Sections	2020/E/015	2020/E/039
CHAPTER 1: INTRODUCTION		
1.1 Motivation and Overview		
1.2 Aims and Objectives		
1.3 Research Scope		
CHAPTER 2: LITERATURE REWIEW		
2.1 Introduction		
2.2 Forecasting Models		
2.3 Research gap		
2.4 Performance Analysis		
2.5 Available Dataset		
CHAPTER 3 : METHODOLOGY AND RESEARCH PLAN		
3.1 Methodology in brief		
3.2 Detailed Methodology		
3.2.1 Lung cancer data selection		
3.2.2 Data preprocessing		
3.2.3 Feature selection		
3.2.4 Apply machine learning methods		
3.2.5 Compare performance		
3.3 Timeline		
CHAPTER 4: PROGRESS TO DATE		
4.1 Literature Review		
4.2 Database collection		
4.2.1 Identify the Types of Lung cancer		
4.2.2 Dataset Selection		
4.3 Database Preparation		
REFERENCE		
APPENDIX		

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

AC	: Adeno Carcinoma
ACC	: Accuracy
AUC	: Area Under the ROC Curve
BASIC	: Backward Elimination Hilbert-Schmidt Independence Criterion
CNN	: Convolutional Neural Network
DNA	: Deoxyribo Nucleic Acid
DT	: Decision Tree
fRNA	: Functional RNA
GEO	: Gene Expression Omnibus
HSIC	: Hilbert-Schmidt Independence Criterion
IFS	: Incremental Feature Selection
KM	: K-Means
KNN	: K-Nearest Neighbour
LOOCV	: Leave-One-Out Cross-Validation
MCC	: Mattew's Correlation Coefficient
MCFS	: Monte Carlo Feature Selection
miRNA	: Micro RNA
NB	: Naïve Bayes
NN	: Neural Network
NSCLC	: Non-Small Cell Lung Cancer
PCA	: Principal Component Analysis
RF	: Random Forest
RNA	: Ribo Nucleic Acid
ROC	: Receiver Operating Characteristic
SCC	: Squamous Cell Cancer
SCLC	: Small Cell Lung Cancer
SMO	: Sequential Minimal Optimization
SN	: Sensitivity
SP	: Specificity
SVM	: Support Vector Machine
SVM	: Support Vector Machine
TCGA	: The Cancer Genome Atlas research network
TPM	: Transcript Per Million
UCI	: University of California, Irvine

Chapter 1: INTRODUCTION

1.1 Motivation and Overview

Lung cancer is one of the most popular cancer types worldwide, taking deaths of almost 1.8 million people annually [12]. The common term for cancerous growth in lung tissues is lung cancer [1,7]. Gene expression of cancerous cells can be compared to normal cells for understanding the pathology of cancer [7,9]. The DNA of cells in the airways becomes damaged genetically, leading to lung cancer. Inhaling dangerous substances, radiation, air pollution, smoking, and secondhand smoke can all lead to this injury [5,9,12]. Even though this condition is incredibly difficult to cure, the likelihood of recovery increases significantly with early detection [4, 5]. Early diagnosis has been proved to improve survival rate of lung cancer patients [2,9]. Lung cancer is currently very difficult to diagnose and cure because it is found in its more advanced stages [4,5,12]. Chest pain, fluid accumulation in the chest, breathing problems, hoarseness, appetite loss, blood in the cough, and an irregular new cough are common signs and symptoms of lung cancer [5]. Remarkably, several of them highlighted that the 5-year lung cancer survival rate can perhaps be realised by early-stage onset patients who received adequate treatment [2,11,12,13]. Basically, it is divided into two large groups; non-small cell lung carcinoma (NSCLC 85%) and small cell lung carcinoma (SCLC 15%) [1]. NSCLC comprises of adenocarcinomas, squamous carcinoma and large cell carcinoma [1,6]. SCLC are small cell carcinoma and Combined small cell carcinoma. Lung AC and lung SCC are two major subgroups of lung cancer with different characteristics in multiple aspects, including epidemic characteristics, pathogenesis, and genetic background [1]. Nowadays, mutations that are characteristic of lung cancer and which may enable diagnosis at the early stage of the disease are sought for [12].

Tumor cells can be killed, stopped from spreading, or eliminated by treatments, which may increase the body's immunity against them. Cancer is usually staged based on the size of the original tumor, the extent to which it has entered the surrounding tissue, and whether or not it has spread to lymph nodes or other organs. There are specific staging guidelines for each form of cancer. Every stage can fit into that category in a number of different size and spread combinations.

Early stages (0-2) Cancer is generally localized and may be treated with surgery, radiation, or a combination of treatments. Advanced stages (3-4) cancer has spread, requiring more aggressive treatments like chemotherapy, targeted therapy, or immunotherapy and surgery and radiation [11]. Side effects of lung cancer treatment include shortness of breath, chest wall pain, cough, pain, fatigue, vomiting, and joint pain.

The variety of genomes causing lung cancer may not be fully captured by traditional techniques of diagnosing the disease, which usually focus on histological characteristics and clinical factors. In recent years, finding the genetic types of lung cancer using gene expression profiling has become an effective tool and this knowledge can help guide specific therapies to improve patient outcomes.

Binary gene expression data presents a potentially effective and understandable method for evaluating these complex datasets. This method reduces gene activity to a binary state (expressed or not expressed) [15]. Binary data may help in the identification of different subgroups of lung cancer, resulting in more reliable and clinically significant classifications by reducing the dimensionality and noise present in continuous gene expression measurements [3].

The motivation for this research is to evaluate the effectiveness of binary gene expression data in lung cancer subgrouping. If successful, this method would simplify the analysis of gene expression data, improving its usability for clinical applications and possibly increasing in the identification of new subtypes of lung cancer. Thus, improved patient survival and quality of life may result from the development of more focused and efficient treatment plans. Furthermore, the application of binary gene expression data may open up opportunities for the regular clinical practice of molecular diagnostics, providing an affordable and scalable approach to the classification of lung cancer.

1.2 Aims and Objectives

The main goal of this study is to evaluate the performance and effectiveness of using binary gene expression data to subgroup patients with lung cancer. This includes comparing the performance of continuous gene expression data with binary data to assess the capacity to find clinically significant subgroups [11].

Lung cancer is not a single illness; it has multiple subtypes, each with different biological characteristics, useful responses, and predictions. Properly categorizing these subgroups is necessary for effective therapy, which changes a patient's therapy treatment according to certain symptoms of cancer. Proper classification can lead to better outcomes, more precisely targeted therapies, and greater understanding of the biology underlying the illness.

This research's objective is to discover distinct genetic markers essential for differentiating between various types of lung cancer and to develop effective classification models based on these markers. Next, the research is going to focus on creating effective classification algorithms that use these identified markers to precisely classify lung cancer into its several classifications.

1.3 Research Scope

Our research encompasses a comprehensive probe into the use of gene expression data for classifying lung cancer into its various subgroups. We are going to evaluate two distinct methods: binary gene expression data where the genes are either “expressed” or “not expressed,” and continuous gene expression data which reveals other levels of gene activity. This examination is imperative since if binary gene expression data works well; it would mean great things in terms of reducing complexity in diagnosis, assisting clinicians in understanding data and ultimately expediting clinical decision-making with reduced expenses when it comes to treating lung cancer patients.

Chapter 2: Literature Review

2.1 Introduction

We have cited 15 articles as per this kind now, and hence shall come across more articles touching on a topic of interest till the end of our study. From the referred articles, we have been able to get more information about binary gene expression data, an area; that we had no prior knowledge of.

Array-based gene expression analysis has rapidly grown into a versatile technology to dissect the molecular basis of cancer [4,7,9,13]. Through quantitative analysis of the relative abundance of thousands of genes at the same time the researchers are able to discover certain molecular profiles related to the specific types of cancer, actions and reactions to treatment and overall prognosis of the patient [1,3]. Specifically in lung cancer, using binary gene expression data meaning whether a gene is / is not expressed can be a simple and efficient way to categorize and sub categorize patients with lung cancer [15]. Such a binary approach can simplify gene expression data by eliminating much of the noise while preserving enough information to classify the two types of cancer [15].

It is still unknown that whether the information derived from binary gene expression data has the possibility to be applied to classify the subgrouping of lung cancer. It has been shown to be useful in sorting out molecular variants of lung cancer that seem to have implications on the patients' prognosis and their response to the therapies given to them. This literature review shall discuss current research findings on forecasting models employed for lung cancer subgrouping together with the employed performance analysis methodologies of such models and the datasets that can be used for such research.

2.2 Forecasting Models

A chromosome is a very long molecule consisting of many millions of base pairs. Most of these bases don't do too much, but certain portions of the chromosome are special. They are called genes. These are the parts that code for different things. In a human gene will be on average around 10,000 to 50,000 base pairs long, though the longest is two-and-a-half million base pairs, and when a gene is expressed, specific protein is produced.

Gene expression describes the process by which functional products are made from genes [4]. There are two main stages of gene expression. The first stage is known as Transcription. It involves the production of RNA from a gene. The second stage is called Translation. It involves the use of RNA from transcription by ribosomes to produce proteins. Protein synthesis is not always the outcome of gene expression. Because RNA has a direct functional purpose in the cell, it sometimes fails to be translated into a protein. In this case RNA called functional RNA (fRNA). One example of a fRNA is microRNA (miRNA). This is used regulates gene expression rather than code for a protein. So, we can see that there are two main types of functional gene products: protein and RNA.

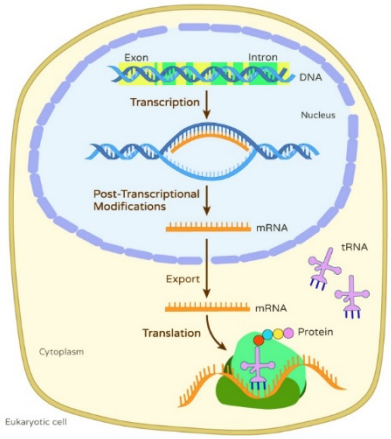

Figure 1 : Gene expression (Transcription & Translation)

Gene expression in lung cancer is a process in which specific genes in lung cells are switched on or "expressed" to make RNA and proteins that affect the behaviour of that cell [4]. In the circumstance of lung cancer, modifications in gene expression can have a paramount role on the emergence as well as progression of the sickness. Certain genes can be overexpressed (producing excessive protein), under expressed (producing inadequate protein), or may undergo changed patterns of expression because of DNA changes, habitat elements or environmental issues.

The uncontrolled proliferation and growth of lung cells because of these gene expression changes eventually leads to tumour formation [4]. Various subsets of lung carcinoma like Small Cell Lung Carcinoma (SCLC) and Non-Small Cell Lung Carcinoma (NSCLC) characteristically have different gene expression patterns [1]. It is possible for researchers to spot a particular biomarker by studying the specific gene expression profiles, which can direct diagnosis, predict the intensity of progression or choose the best treatment option for lung cancer [1,4,7]. These are important since they change fate of the patient's life. Furthermore, there is much more than meets the eye; knowing how it all started would allow us to create drugs against this disease that could interfere with the activity of oncogenes or promote activities mediated by tumour suppressors [4].

Different techniques revealed that gene expression data is central to the molecular characterization of cancer. Gene expression profiling to make probability estimation of cancer appearance, cancer categorization, and estimation of the response to some treatment. These models employ different forms of machine learning and statistical analysis to translate the large and intricate data sets obtained from gene expression experiments [4].

Forecasting models are critical tools in the study of cancer especially in the classification of tissues cancer through the gene expression. For instance, in lung cancer, gene expression data can be applied to diagnose which subtype of lung cancer it is, for example, adenocarcinoma, squamous cell carcinoma. These subtypes are may be characterized by quantitatively different molecular profiles that can be recognized employing Machine learning algorithms such as SVMs (SVM), Random forests (RF), Decision tree (DT), K-Nearest neighbour (KNN), K-Means (KM), Naïve Bayes (NB), Neural networks (NN) [1,2,3,4,5,6,7,8,9].

Apart from that, using forecasting models it is possible to guess the further development of the illness or effectiveness of treatment. For example, one can obtain some gene expression profiles that are characteristic of improved or worse outcomes in patients with lung cancer [1]. Machine learning can therefore be used to train models with historical patients' data to be used in forecasting future performance such as survival rates after certain periods, chances of relapse or chances of responding favourably to chemotherapy [4].

Forecasting models are also used in creating the concept of pharmacogenomics where medicine is adapted to the patients' genes. This is because the models in question can predict which of the treatment plans are most likely to be effective based on the expression of genes in a particular patient, helping to optimize the outcomes of cancer therapy and eliminating the possibility of a 'hit and miss' situation [12].

2.3 Research gap

Our research addresses critical issues in the available literature on lung cancer subtypes. To begin with, only a small number of researchers have employed binary gene expression data which simplifies the gene activity solely to 'on' or 'off' states rather than the much preferable continuous data. Moreover, research aimed at assessing how well binary data compare to continuous data in cancer subgroups is still wanting, and thus it is not clear which data type is superior.

Moreover, it is possible that environmental factors might affect the way genes are expressed, which makes it hard to evaluate such models with precision without considering those signal factors. Another significant challenge is that in the process of converting continuous to binary data, the making of threshold on the cut off 'on' or 'off' levels have a significant bearing on the findings yet this aspect has not been explored in most of the previous studies.

2.4 Performance Analysis

The assessment of performance in the context of our research on subgrouping lung cancer patients using binary gene expression data means determining how effective the chosen models and methods are in grouping the patients according to gene expression. Here is a summary of the presented details that need to be addressed:

1. Subgrouping Accuracy
    - We will have to see how well the models can be used to classify lung cancer patients into groups (e,g, lung cancer types or stages). This can be done, for instance, by using the concept of accuracy, which is the ratio of correct predictions to total predictions.
2. Binaries vs. Continuous Data Comparison
    - In order to reconsider the application of continuous data, performance of the models employing binary gene expression data and continuous

gene expression data should be compared. Similar to precision, that tells what portion of the predicted subgroups is correct, recall, which indicates the number of the predicted subgroups that were retrieved, and F1-score, which is a weighted average of precision and recall, can be applied. This will help determine if binary data is less effective than or equal to continuous data.

3. Consequence of the Binarization Threshold [15]
    - The threshold used in binarization (the one below which the gene is considered to be 'off' and above which it is considered 'on') is one of the factors that affect the performance of a model. We need to discuss what effect does changing this threshold has on the performance of the model with respect to patients grouping. For instance, a very strict threshold might compromise capturing important features / patterns, while a very loose one might cause overfitting.
4. Model Efficiency
    - Binary data can frequently be processed more quickly and easier than continuous data due to the fact that it only reduces the gene expression into just two states (on and off). Evaluate how much time or computational resources our models save when using binary data.
5. Robustness to Noise
    - Since binary data may handle noise better than continuous data because it simplifies expression into clear categories. We should assess whether binary data achieves a better level of stability of results, even in the presence of errors or variations in the gene expression data.

Addressing these issues, we will be able to evaluate how effective the usage of binary gene expression data in lung cancer sub grouping is and whether it has any benefits over continuous data regarding accuracy, cost effectiveness, and biological insight.

2.5 Available Databases

1. Gene Expression Omnibus (GEO)
    - GEO is one of the most commonly used repositories for gene expression data.
    - This incorporates RNA-seq and microarray experimental data in the form of both binary and continuous gene expression datasets.
        - GSE43580 [1]
        - GSE23739 [2]
        - GSE3141 (Shedden et al) [3]
        - GSE50412 [6]
        - GDS3257 [7]
        - GSE175601 [11]
        - GSE10799 [11]

- GSE28582 [14]
- GSE13255
- GSE135304
- GSE12771
- GSE40419

2. The Cancer Genome Atlas Research Network (TCGA) [3,6,7,11,14]
   - The Cancer Genome Atlas (TCGA) is a wide-ranging repository of genomic, transcriptomic and clinical information for multiple cancer types. The lung cancer data sets consist of LUAD, and LUSC which provide an elaborate gene expression profile.
   - Gene expression data is available for extraction and can be binarized appropriately for our research needs.

3. Data world cloud-native SaaS platform [5]

4. Kent Ridge Bio-Medical Dataset repository [9,13]

5. STRING database [11]
   - It used to construct a protein-protein interaction (PPI) network

6. UCSC Xena

7. ArrayExpress
   - As a part of the European Bioinformatics Institute, ArrayExpress provides both uploaded content as well as raw data comprehended from experiments depicting gene expressions in lung cancer.

8. Lung Cancer Explorer
   - LCE is a specialized resource for gene expression and clinical data specific to lung cancer. It offers datasets that may be particularly relevant for our subgrouping study.

9. cBioPortal for Cancer Genomics
   - This platform offers cancer genotyping data sets, and in particular, lung cancer gene expression and mutation data sets. It compiles and makes available multiple data sets such as the TCGA data sets and allows comparisons across studies.

Chapter 3: Methodology and Research Plan

3.1 Methodology in Brief

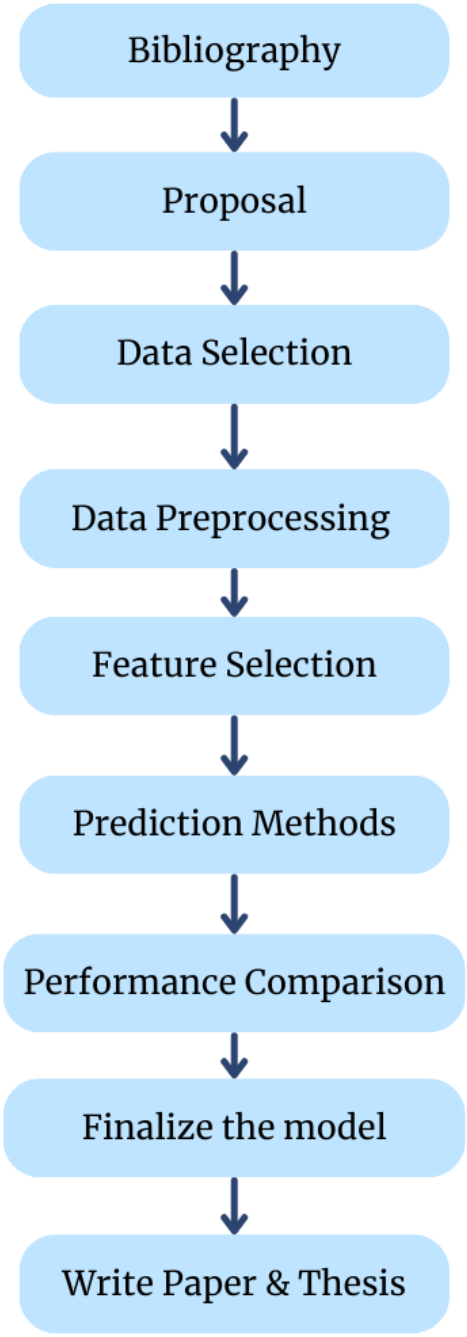

Figure 2 : Methodology in brief

3.2 Detailed Methodology

3.2.1 Lung cancer data selection

As per our research is based on gene expression cancer prediction based on subtypes, we selected lung cancer for further as it has 2 main groups and 5 subgroups. According to this We should select gene expression data, so we gather datasets from TCGA (The Cancer Genome Atlas), GEO (Gene Expression Omnibus) and NCBI (National Centre for Biotechnology Information) repositories. It makes a large collection of molecular and genomic data accessible to medical professionals, researchers, and the general public, facilitating a better understanding of the molecular and genetic mechanisms underlying health and disease. Among available datasets we decided to use Gene expression RNAseq , miRNA strand expression RNAseq and Microarray data.

Consider patient's gene expression data (transcript per million (TPM) values), as well as clinical data, such as smoking history, cancer stage, patient survival time, patient censoring event, patient age, and patient sex.[4] And analyze how they impact directly on the selected dataset. They could be used as the target of our prediction.

First, we access the gene expression profiles from GEO under accession number GSE23739 [2] , GSE135304 [16], GSE13255 [16] and GSE42830 [16].

GSE23739 [2]

The title of the dataset is "miRNA expressions in non-cancerous and cancerous human gastric tissues". The experiment type is Non-coding RNA profiling by array. MiRNA expression was profiled on Agilent Human miRNA Microarrays (V2) representing 723 human and 76 human viral miRNAs in 40 normal and 40 cancerous gastric tissues. Researchers investigated the involvement of microRNAs (miRNAs) in gastric cancers. MiRNA expression was profiled from 40 cancerous and 40 non-cancerous tissues obtained from the National Cancer Centre, Singapore, and Singhealth Tissue Repository, Singapore. They identified 80 differentially expressed miRNAs in tumors compared with normal tissues. Among these miRNAs, we identified hsa-mir-486-5p (mir-486) as a significantly downregulated miRNA in GC. Subsequent functional characterization revealed that mir-486 playing a tumor suppressor role. They also observed frequent genomic deletion of mir-486 in 20-30% of GCs. To knowledge, mir-486 represents one of the first tumor suppressor miRNAs in GC inactivated through genomic deletion.

GSE135304 [16]

The title of the dataset is "Non-Small Cell Lung Cancer, Peripheral Immunity, Extending the Tumor Macroenvironment". The experiment type is Expression profiling by array. Researchers recently established that gene expression in PAXgene stabilized blood RNA distinguishes benign (BN) from malignant (MN) pulmonary nodules in high risk candidates with an AUC of 0.84. We now expand studies to include incidental nodules identified in routine clinical settings using data

from 603 patients analyzed on Illumina microarrays. They identify 300 gene probes achieving an AUC of 0.84 for Indeterminate Pulmonary Nodules (IPN) from 6-25 mm and 0.824 for IPN from 8-20 mm, outperforming 3 prominent clinical models that achieve AUCs of 0.60-0.689. They address the basis for these differences by in silico flow cytometry using CIBERSORT and identify significant differences between MN and BN patients including proportions of γδT-cells, M0 macrophages, NK cells, B cells and exhausted CD8 T-cells.

GSE13255 [16]

The title of the dataset is "Gene Expression Profiles in Peripheral Blood Mononuclear Cells Can Distinguish Patients with NonSmall Cell Lung Cancer". The experiment type is Expression profiling by array. They compared microarray gene expression profiles in PBMC from patients with NSCLC and a control group with smoking-related non-malignant lung disease, a likely confounding factor. And found a distinguishing gene signature which was validated on 2 independent sets of hold-out samples one set from a different location. Gene expression changes were also compared between pre and post surgery samples from 18 patients.

3.2.2 Data preprocessing

First filtering missing values, Patients without both clinical and gene expression data were excluded. Furthermore, duplicate entries for patients were excluded, along with patients with missing clinical data. Data preprocessing includes data cleaning, handle duplicate values, feature scaling, missing value management, removing outliers and pattern standardization [2]. All gene expression data were scaled from 0 to 1 using binarization techniques.

Should follow a binarization, Turning the gene expression data into a simple yes/no format like 1 for "expressed" and 0 for "not expressed". This makes it easier to analyse without losing much important information. Classify data by comparing it to predefined profiles (By getting threshold value) binarize all according to that.[15]

It is important to increase the dataset's quality because poor data can largely reduce accuracy and result in erroneous predictions. Data pre-processing is the way to improve the quality of the data in a correct way. To increase the quality of data Quantile Normalization should be followed.[1]

3.2.3 Feature selection

Feature selection is the process of choosing the most important variables (features) from a dataset that will be used to build a machine learning model. We are planning to compare features(genes) using Incremental Feature Selection, Information gain attribute ranking and Backward Elimination Hilbert-Schmidt Independence Criterion in our model.

Information Gain Attribute Ranking

To identify the most important genes in a lung cancer gene expression study, It measures how much each gene contributes to predicting different subgroups or outcomes of lung cancer. The higher the information gain, the more relevant the gene is considered for the study. By selecting genes with high information gain, can focus on the most informative features.[9]

Backward Elimination Hilbert-Schmidt Independence Criterion

Helps identify the most relevant genes by iteratively removing the least significant features based on their contribution to the target variable, such as cancer subtypes or patient outcomes. BASIC uses the Hilbert-Schmidt Independence Criterion (HSIC) to measure the dependency between gene expressions and the target variable, ensuring that only the most informative genes are retained.[7]

Helps to reduce the dimensionality of the dataset. This method allows for a more focused analysis on the genes that are most likely to have a biological significance in distinguishing between different lung cancer subgroups.

Incremental Feature Selection

After obtaining a list of features IFS method will be applied to extract optimal features that would allow the model to achieve the best performance in classifying lung adenocarcinoma and lung squamous cell carcinoma samples. [6,1]

Monte Carlo Feature Selection

This method will be used to analyze the gene expression profiles and extract important features from the high-dimensional data [1]. MCFS is particularly effective in handling datasets with many features and few samples.

3.2.4 Apply machine learning methods

Selecting an appropriate model is another important step to improve the accuracy of the prediction. Below models are expected to use in our research. As we are doing cancer subgrouping prediction we are expected to try on wide range of models and according to the model performance we are trying to improve and pick the best model.

Support Vector Machines (SVM)

This is a popular classification method used in many related studies, particularly with different types of kernels, including the Tanimoto kernel [7,15]. SVMs are effective for high-dimensional data like gene expression profiles [1].

Tanimoto coefficient (*T*) [15], between two binary vectors, is defined as follow:

$$T = \frac{c}{a + b - c}$$

Where,
a: the number of expressed points for gene x
b: the number of expressed points in gene y and
c: the number of common expressed points in two

Tanimoto kernel can be defined as [15]:

$$K_{Tan}(x, z) = \frac{x^T z}{x^T x + z^T z - x^T z}$$

Where,

$c = x^T z, \quad b = z^T z, \; a = x^T x$

SVM was employed as the primary classification algorithm to build an optimal classifier for distinguishing between lung adenocarcinoma (AC) and lung squamous cell carcinoma (SCC) samples.[1] The SVM was optimized using the Sequential Minimal Optimization (SMO) algorithm to accelerate the training process.

Naive Bayes

This classifier works by assuming that gene expressions are conditionally independent given the cancer subgroup, and it calculates the probability of each subgroup based on observed gene expression patterns. Despite the independence assumption, which may not fully hold in biological systems, Naive Bayes often performs well with high-dimensional data like gene expressions.
Random Forest

It handles high-dimensional data by building multiple decision trees, which improves accuracy and robustness. The method identifies important genes that differentiate between subgroups and captures complex interactions between genes. It scales well and provides accurate subgrouping, making it valuable in analysing gene expression data. It showed high sensitivity and was considered appropriate for early lung tumour diagnosis [2].

K-Nearest Neighbors (KNN)

KNN is a simple and intuitive method to work with binary gene expression data. It classifies samples based on the majority class of their nearest neighbours. While KNN is flexible and requires no training phase, it can be computationally intensive with large datasets and sensitive to noise in the data.

Decision tree

The supervised learning technique of decision trees is used in classification. Unique quality of the models created using this methodology is how closely they resemble human reasoning. By iteratively dividing the data into its features and based on the current classifications, a decision tree is constructed.

3.2.5 Compare performance

Performance measurement is a very important aspect in our study as our aim is building a better model for lung cancer subgrouping.

Sensitivity (SN) [1,2,12]

- The ratio of true positives (such like people suffering from lung cancer) to those who are detected by the model describes the level of sensitivity.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity (SP) [1,2,12]

- The model identifies the portion of real negatives (for instance, fit individuals) accurately.

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy (ACC) [1,2,4,5,6,7,9,13]

- It is the least complex measure, and it holds the ratio of correct predictions to total overall predictions of the model. Nonetheless, it can be stated that accuracy as a measure describes the capacity of an AI system to classify objects in the training set correctly while ignoring other factors, for instance, a group of images may be represented inadequately in the dataset.

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

Area Under the Curve [4,6,7]

This metric will be derived from the Receiver Operating Characteristic (ROC) curve and provides a single value that summarizes the model's performance across all classification thresholds. A higher AUC indicates better model performance

Leave-One-Out Cross-Validation

This method will be used to evaluate the performance of the classifier. In LOOCV, one sample will be removed from the training dataset for testing, and the remaining samples will be used for training. This process is repeated for each sample in the dataset, providing a robust estimate of the model's performance [6].

Confusion matrix

A confusion matrix contrasts the model's correct prediction with its incorrect prediction. A Confusion matrix's columns represent the actual data, while its rows represent what the machine learning system projected. The number of things we wish to forecast affects how big the confusion matrix is.

Always, its projected outcomes are recorded as True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) [1,2,4,5,7,9]. These following terms are crucial elements of them [2]:

- TP = Cancerous lung patients diagnosed correctly as sick people.
- FN = Noncancerous lung tumor patients erroneously assumed to be fit individuals.
- FP = Individuals not having cancer wrongly categorized as being sick.
- TN = Persons who are healthy and have been accurately categorized.

Mathew's Correlation Coefficient (MCC) [1]

- MCC does the job of assessing binary classifications' quality. The four confusion matrix categories (TP, TN, FP and FN) constitute its inputs; hence it is most often viewed as a fair metric applicable even when group sizes vary widely.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TN + FN) \times (TN + FP) \times (TP + FN) \times (TP + FP))}}$$

Recall [1,4,5,6,7,9,13]

- Similarly, recall refers to similar cases as sensitivity in that it expresses the rate at which genuine positives are detected by the model.

$$Recall = \frac{TP}{TP + FN}$$

Precision [1,5,6,7,9,13]

- Precision refers to percentage of accurate positive forecasts amidst all positive predictions made.

$$Precision = \frac{TP}{TP + FP}$$

F1-measure [1,5,6,7,13]

- Meanwhile, F1 score is a statistical measure derived from both precision and recall. It serves as a one-point metric that equilibrates trade-offs between them.

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

3.3 Timeline

Table 1 : Timeline

	Semester 6					Semester 7					Semester 8	
	Week 1-3	Week 4-6	Week 7-9	Week 10-12	Week 13-15	Week 1-3	Week 4-6	Week 7-9	Week 10-12	Week 13-15	Week 1-3	Week 4-6
Bibliography												
Literature review												
Research proposal												
Data collection												
Data preprocessing												
ML model development												
Model evaluation												
Report writing												
Research paper and thesis writing												

Chapter 4: PROGRESS TO DATE

4.1 Literature Review

Till now, we have referred to a total of 15 articles. In addition to that, we are still reading some new articles concerning our topic. Also, review of the literature will be done at various points during the research process.

4.2 Database Collection

4.2.1 Identify the types of Lung cancer

Lung cancer Subgroups

Lung cancer can be classified into various subgroups based on histological characteristics, genetic mutations, and molecular profiles. These subgroups help in determining the most effective treatment approaches. Here are some key lung cancer subgroups:

1. Histological subgroups
   a) Non-Small Cell Lung Cancer (NSCLC)
      - Adenocarcinoma
        - This is common especially among non-smokers.
        - Usually found in the lungs' outer regions.
      - Squamous cell carcinoma
        - Usually associated with suckers; it happens mostly at bronchi.
      - Large cell carcinoma
        - It is less prevalent and can be found in any part of the lung.
        - Tends to develop and progress rapidly.
   b) Small Cell Lung Cancer (SCLC)
      - It is highly aggressive and closely linked to smoking habits.
      - It begins typically from lungs' centre area and spreads quickly.

2. Clinical subgroups
   a) Early-Stage Lung Cancer
      - Often surgery, radiation, or a combination of both are used to treat this local problem either with or without chemotherapy.
   b) Locally Advanced Lung Cancer
      - With nearby tissues or lymph nodes affected by the illness but no distant organs involved.
      - Therapies can include combinations of chemotherapy, radiation and occasionally surgery.

c) Metastatic Lung Cancer
    - This type involves cancer cells that have spread from the original site to other parts in the body.
    - This condition can be treated with systemic therapies such as chemotherapy, targeted therapy or immunotherapy among others.

4.2.2 Data set selection

According to our research have to use gene expression data in lung cancer subgrouping, it is important to choose a dataset that is specific to lung cancer and includes a variety of cancer subtypes. The data should be in a binary format, meaning each gene is marked as either "on" (expressed) or "off" (not expressed). Otherwise, we can binarize the normal gene expression dataset also.

The dataset should also have a large number of genes (features) and samples to provide enough information for accurate subgrouping. We select high quality reliable datasets from publicly available repositories like Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). Additionally, the dataset should include clinical details about patients to help validate the study's findings and ensure they are relevant to real world treatment.

4.3 Database Preparation

We downloaded each data set separately and check whether it is in appropriate format (TXT or CSV) with required sample size. Then for the downloaded datasets, we added target values. After complete these steps we started data preprocessing steps. First, we merge the labels and the samples into one file if it is separate files then check whether there is any null values and duplicates in dataset. Then drop null values and duplicates. Encode the target column for further analysis. Then apply binarization using two gaussian matrix model. Then plan to perform feature selection methods which are mentioned above.

REFERENCES

[1] Fei Yuana, Lin Lub, Quan Zouc (2020), "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithm", BBA–Molecular Basis of Disease, 1866, 165822, Available: 10.1016/j.bbadis.2020.165822

[2] Ying Xie, Wei-Yu Meng, Run-Ze Li, Yu-Wei Wang, Xin Qian, Chang Chan, Zhi-Fang Yu, Xing-Xing Fan, Hu-Dan Pan, Chun Xie, Qi-Biao Wu, Pei-Yu Yan, Liang Liu, Yi-Jun Tang, Xiao-Jun Yao, Mei-Fang Wang, Elaine Lai-Han Leung (2021), "Early lung cancer diagnostic biomarker discover by machine learning methods", Translational Oncology, 14, 100907, Available: 10.1016/j.tranon.2020.100907

[3] Dingjingmu Yang (2024), "Application of machine learning in lung cancer prediction", Proceedings of the 4th International Conference on Signal processing and Machine Learning, University of Alberta, Available: 10.54254/2755-2721/53/20241290

[4] Jason Z.Zhang and Chi Wang (2023), "A comparative study of clustering methods on gene expression data for lung cancer prognosis", BMC Research Notes, 16(319), Available: 10.1186/s13104-023-06604-8

[5] Sherko H. Murad, Ardalan H. Awlla, Brzu T. Mohammed (2023), "Prediction cancer based critical factors using machine learning", Science Journal of University Zakho, Vol. 11, No. 3, pp. 447 – 452, July-September, 2023, Available: 10.25271/sjuoz.2023.11.3.1105

[6] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. Ngai and J. Shao, "Classification of lung cancer using ensemble-based feature selection and machine learning methods", Mol. BioSyst., 2014, DOI:10.1039/C4MB00659C
Available: 10.1039/c4mb00659c

[7] Mahmood Khalsan, Lee R. Machado, Eman Salih Al-Shamery, Suraj Ajit , Karen Anthony, Mu Mu, And Michael Opoku Agyeman "A survey of machine learning approaches applied to gene expression analysis for cancer prediction" 2022, DOI 10.1109/ACCESS.2022.3146312
Available: 10.1109/ACCESS.2022.3146312

[8] Tabares-Soto R, Orozco-Arias S, Romero-Cano V, Segovia Bucheli V, Rodríguez-Sotelo JL,
Jimenez-Varón CF. 2020. "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data" DOI 10.7717/peerj-cs.270 Available: 10.7717/peerj-cs.270

[9] Jayadeep Pati, 2019, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco- Genomics approach", International Journal of

Cybernetics and Informatics, vol. 7, DOI 10.1109/ACCESS.2018.2886604 Available: 10.4018/IJEHMC.2021010101

[10] Marcel Wiesweg, Fabian Mairinger, Henning Reis, Moritz Goetz, Jens Kollmeier, Daniel Misch, Susann Stephan-Falkenau, Thomas Mairinger, Robert F.H. Walter, Thomas Hager, Martin Metzenmacher, Wilfried E.E. Eberhardt, Gregor Zaun, Johannes Ko¨ ster, Martin Stuschke, Clemens Aigner, Kaid Darwiche, Kurt W. Schmid , Sven Rahmann, Martin Schuler "Machine learning reveals PD-L1-independent prediction of response to immunotherapy in non-small cell lung cancer based on gene expression context", European Journal of Cancer, vol. 140, DOI:10.1016/j.ejca.2020.09.015 Available: 10.1016/j.ejca.2020.09.015

[11] Qiaojun Hong, Haiyan Hu, Dandan Liu, Xiaojian Hu, Zhanggui Wang, Daoping Zhou, "Bioinformatic analysis of differentially expressed genes in lung cancer bone metastasis and their implications for disease progression in lung cancer patients", Journal of Thoracic disease, Vol 16, No 7 July 2024, Available: 10.21037/jtd-24-1081

[12] Katarzyna Wadowska, Iwona Bil-Lula, Lukasz Trembecki and Mariola Sliwinska-Mosson, "Genetic markers in lung cancer diagnosis", International Journal of Molecular Sciences, 2020, Available: 10.3390/ijms21134569

[13] Thangamani M, Manjula Sanjay Koti, Nagashree B.A, Geetha V, Shreyas K.P, Sandeep Kumar Mathivathanan and Gemmachis Teshite Dalu, "Lung cancer diagnosis based on weighted convolutional neural network using gene data expression", Scientific Reports, 2024, Available: 10.1038/s41598-024-54124-7

[14] Joe W. Chen & Joseph Dhahbi "Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods", Available: 10.1038/s41598-021-92725-8

[15] Salih Tuna, Mahesan Niranjan1 "Classification with binary gene expressions", School of Electronics and Computer Science, University of Southampton, Southampton, UK, Vol.2, No.6, 390-399, Available: 10.4236/jbise.2009.26056

[16] Arya Hadizadeh Mogaddam, MS, Mohesen Nayebi Kerdabadi, MS, Cuncong Zhong, PhD, Zijin Yao, PhD, "Meta-Learning on Augmented Gene Expression Profiles for Enhanced Lung Cancer Detection", 2024

APPENDIX (Optional)

Table No 1. Table Title
