COMPARING THE PERFORMANCE OF VARIOUS MACHINE LEARNING TECHNIQUES ON CANCER PHENOTYPE ANALYSIS
TITLE OF THE RESEARCH

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

Submitted by:

Ariyakulasinghe A.P.H. (2017/E/009)

Nirmani B.G.M. (2017/E/074)

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF JAFFNA

[JANUARY] [2021]

COMPARING THE PERFORMANCE OF VARIOUS MACHINE LEARNING TECHNIQUES ON CANCER PHENOTYPE ANALYSIS

Supervisor(s):

Supervisor	:	Dr. P. Jeyananthan
External Supervisor	:	Dr. Omar Shetta

Examination Committee:

Lecturer 1 .....

Lecturer 2 .....

CONTRIBUTION TO THE PROPOSAL BY THE MEMBERS IN GROUP

Sections	2017/E/009	2017/E/074
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 Performance Analysis		✓
2.4 Available Databases		✓
CHAPTER 3 : METHODOLOGY AND RESEARCH PLAN		
3.1 Methodology in Brief	✓	✓
3.2 Detailed Methodology		
3.2.1 Breast 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	✓	
3.2 Timeline	✓	✓
CHAPTER 4: PROGRESS TO DATE		
4.1 Literature Review		✓
4.2 Database Collection		
4.2.1 Phenotype Data selection		✓
4.2.2 Dataset Selection	✓	
4.3 Database Preparation	✓	
REFERENCE	✓	✓
APPENDIX		

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

AJCC	:	American Joint Committee on Cancer
ANN	:	Artificial Neural Network
AUC	:	Area Under the Curve
CNN	:	Convolutional Neural Network
CNV	:	Copy Number Variation
DNA	:	Deoxyribonucleic Acid
GDSC	:	Genomics of Drug Sensitivity in Cancer
HER2	:	Human Epidermal Growth Factor Receptor 2
ICBC	:	Iranian Centre for Breast Cancer
INDEL	:	Insertion-Deletion
KNN	:	K Nearest Niebuhr
LSTM	:	Long Short-Term Memory
MPL	:	Multi-Layer Perceptron
mRNA	:	messenger Ribonucleic Acid
NOS	:	Not Otherwise Specified
PAM50	:	Prediction Analysis of Microarray 50
RNA	:	Ribonucleic Acid
RNN	:	Recurrent neural network
ROC	:	Receiver Operating Characteristic
RPPA	:	Reverse Phase protein Array
RvNNs	:	Recursive Neural Networks
SNP	:	Single Nucleotide Polymorphisms
SVM	:	Support Vector Machine
TCGA	:	The Cancer Genome Atlas
TNM	:	Tumour, Node and Metastases
WDBC	:	Wisconsin Diagnostic Breast Cance

Chapter 1: INTRODUCTION

1.1 Motivation and Overview

There were over 9.6 million deaths in 2018 due to cancer, becoming the second leading cause of death worldwide [13]. The rapid growth of abnormal cells that don't have usual cell growing boundaries can be identified as cancer. Activation of oncogenes, malfunctioning of tumor suppressor genes or mutagenesis due to external factors are main causes of the cancer.

There are over 200 types of cancers. A specific treatment which includes one or more steps from surgery, radiotherapy and chemotherapy is required for every cancer. These treatments vary between cancer subgroups as well [9]. Not only that but also there are severe side effects of the treatments such as Neutropenia, Lymphedema, Nausea, and Vomiting, etc, which can affect the healthy tissues or organs. If surgeons could not identify cancer subgroup correctly, the patient will be exposed to unnecessary side effects. So the correct identification of cancer subgroup has direct impact on the effectiveness of the treatment and the existence of the patient.

There is a very close relationship between molecular data and the phenotype data of the cancer patients [9]. Emerging technologies lead us to measure the molecular data very efficiently and there are dedicated data repositories such as TCGA to keep such data from various cancer patients ["GDC", Portal.gdc.cancer.gov, 2021. [Online]. Available: https://portal.gdc.cancer.gov/. [Accessed: 23- Jan- 2021]]. If we can predict the phenotype data of the patient using the available molecular data, patients can get more accurate cancer treatments [6]. An oncologist can choose the proper treatment while minimizing the negative effects associated with ineffective treatments.

Now the problem is accuracy of the model. As our objective is closely related to the survival of the patient, we should aim a model with highest accuracy. Now we have to achieve this aim using such a large data set with almost 20,000 genes. The proper way to reach this achievement is Machine learning.

There are a lot of machine learning techniques such as Support Vector Machine (SVM), Artificial Neural Network (ANN), K Nearest Niebuhr (KNN), deep learning, etc which could be used for this analysis. We should choose the proper machine learning technique with highest accuracy for this analysis. Selection of the machine learning algorithm is not limited to the prediction, but it is for pre-processing and feature selection as well. At the end of this research, we target to produce a combination of the machine learning techniques which yields an excellent model for the analysis of several phenotype data of cancer patients.

1.2 Aims and Objectives

Previous studies show the interest of the current researchers on neural network based methods [4]. They used neural network based machine learning techniques for their researches without any comparison, because they think that those techniques are more efficient than the classic machine learning techniques. As this study is very crucial, we cannot choose some random method of interest for this study. Hence we are doing a comprehensive comparison between machine learning methods here for cancer phenotype predictions.

Main objective of this study is comparing the performance of various machine learning techniques on cancer phenotype predictions. At the end, provide a combination of machine learning techniques for pre-processing, feature selection and prediction. Also we aim to study the factors could be influence on the performance of each machine learning methods. Hence, why performance differ between methods?

1.3 Research Scope

Scope of our study is providing a comparably better model for cancer phenotype prediction with high accuracy. This model will enhance the treatment pattern of several cancers, where millions of patients in the world would get benefit from this. Also this study will be very useful in the medical field.

Chapter 2: Literature Review

2.1 Introduction

Nowadays, cancer has become one of the leading cause of death [3,1]. Factors such as human diet and genetic factors have a significant contribution on this issue. Even though some of these factors are already revealed, there are on-going studies to reveal hidden associative factors. It means that cancer-related experiments and studies are being updated for time to time.

The earlier researches used statistical methods in their cancer related studies [4]. However, recently, for the improvisation, machine learning methods are vastly used in these studies. Like us, most of novel researchers and studies take advantage of these machine learning technologies.

In cancer-related studies, accuracy is very important. The final output of these studies helps to improve cancer treatment accuracy. It directly affects the survival of the patient. Hence, a poor or inaccurate results of these studies will negatively impact the patients.

When we proceed with this research, we should have a clear understanding of the biological criteria of cancers [7, 14], trends of machine learning technologies currently in use [1, 3, 4, 6, 8, 9, 10] and various type of predictions [1, 3, 8, 9, 10]. Recently published research papers provide better approaches to understand these issues. Hence, we went through some related research articles to gain knowledge in biology and machine learning. It was very helpful to identify the state of the art in cancer-related studies.

2.2 Forecasting Models

Gene expression is a process which is inherent and stochastic [14]. Clear understanding of this process helps to find possible cures for diseases such as cancer. Gene expression has two sequential steps called transcription and translation. As illustrated in figure 2.1, transcription is the process of state conversion from DNA to messenger RNA (mRNA). This is the initial state of this process. Promoters are accountable for the beginning of the transcription process. Next is translation. In this step, mRNA produce protein with the help of tRNA (transfer RNA) and ribosomes [7, 14].

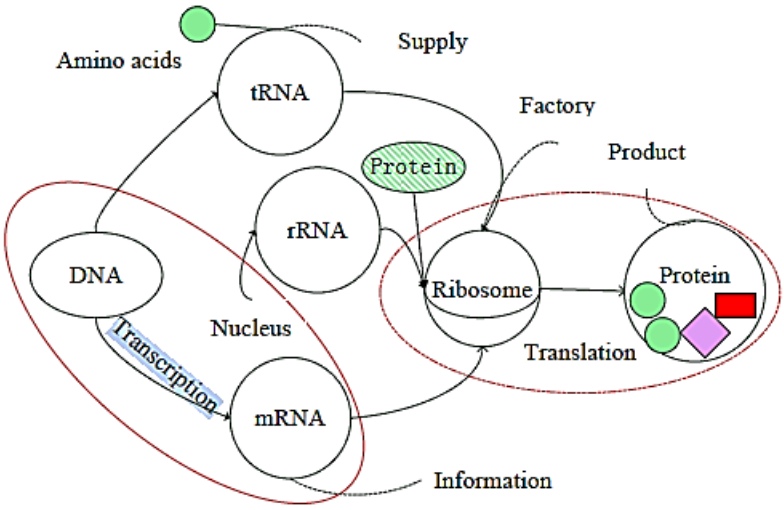

Figure 1 :Gene Transcription and translation in design [7]

A cancer can be caused by the modification of genes [11]. Problems with cell division or damage to DNA can also lead to cancer [11]. To maintain the metabolism and the proliferative state of cancer cells, transcription and translation are up-regulated in the cancer cells [11]. Also, most of the cancers are tumours. When cancer grows, distinct tumour cells might have unique genetic modifications [10]. This shows the role of transcription and translation in cancer related studies and the importance of genomic level data in cancer-related studies.

Literature shows that different levels of gene expressions have been used in cancer related studies [3, 4, 6]. RNA data include observed gene expression at transcription level. There are various technologies to measure this transcriptome data such as sequence array and RNA seq. Number of observed genes differ across technologies [4, 6]. This is general for all of the molecular data.

Another problem in these studies is combining datasets from different studies. For various reasons, researchers combined different datasets in few studies. It may cause biases in their study [6]. Anyways, nowadays researchers have paid their attention to solve this kind of issues too. A proper pipelining method can improve the quality of the combined dataset and reduce the bias in this process. Taking the union of observed gene data while including the most important characteristics lead to build a framework to combine the datasets [6]. Other than genomic data, some of the cancer-related predictions used the clinical data of patients [3].

Next we studied different machine learning algorithms which suit our study. The border between statistics and machine learning is obscure. The analysing methods used in early predictions on complex biological data relayed mainly on basic statistical methods. But now the technology has improved and moved into machine learning [4]. Even with machine learning technologies, some problems have been raised such as high dimensionality (high resolution), data integration issues (as mentioned earlier), and bias of hidden biological effects. This high dimensionality can lead to Problems such as sparsity, multi-collinearity, and overfitting. Data projection and feature selection techniques can increase the performance and reduce the dimensionality of the data. To handle the bias of hidden biological effects, the dataset

should be carefully studied and all the biological and technical facts should be considered. However, the increasing number of available samples shows great promise to increase our understanding of complex biological phenomena [4, 6].

The cancer-related studies choose different targets such as the effectiveness of drugs in cancer cells [9], cancer recurrence [1, 3, 8], death from the disease [3], cancer susceptibility [8] and cancer survival [8]. Also, researchers can choose suitable machine learning technology according to their purpose of the study. Different machine learning methods have different characteristics. However, wide choice of machine learning methods such as Support Vector Machines [1, 3, 5, 6, 8, 9, 10, 15], Naïve Bayes [6], Lasso Regression [6], Random Forests [5, 6, 8], convolutional neural networks [4], Consensus clustering [9], Decision tree [1, 5, 10], Artificial Neural Network [1, 3], Logistic Regression [5, 8], DNN [12], KNN [5], Gaussian Naïve Bayes [5], Gradient boosting [5], Linear Discriminant Analysis [5], quadratic discriminant analysis [5], ANN [15], k-means [15] and Multi-Layer Perceptron [8] were used in biological studies. Among the machine learning methods used in previous studies, we identified that SVM is the most popular one with high accuracy on classification.

Most of the researchers used more than one machine learning method for different purposes and some studies sequentially use these methods to achieve different objectives such as feature extraction, normalization, feature selection, and classification. As we mentioned earlier, the dataset should be prepared properly while selecting suitable machine learning methods [9].

There are different kinds of studies in this case such as same studies on different cancers, applies different methods on same cancers, etc. While doing the research using machine learning methods, researchers chose one type of cancer such as breast cancer, lung cancer, or oral cancer [1, 8, 9, 10]. Other than that, some researchers used sub-groups of cancers which can produce a better predictive model [3].

Some other cancer-related projects are comparing some different methods, tools, languages, etc [1]. So comparison between different machine learning methods on the same dataset to predict the same target can propose the most suitable machine learning method for a specific task. Without this kind of comparison, one cannot define a machine learning method as the most suitable machine learning method. The performance is very important in cancer-related studies because these results should help to improve the treatment effectiveness of the patients. Also, it is worthy to note that the method with higher performance could be either the oldest or the latest machine learning method [1].

As mentioned above, different kinds of languages and different kinds of machine learning tools are available to implement machine learning models. Each strategy has its own characteristics, advantages, and disadvantages. As an example, some tools required more memory space and some tools are taking more time to proceed. According to the available resources and requirements, researchers should choose a better tool for modelling [8].

2.3 Performance Analysis

There are several methods in machine learning, for measuring the performance of the model. So different studies used different methods such as Area Under the ROC Curve [3, 8], sensitivity [5, 7, 8, 13], specificity [5, 7, 8, 13], accuracy [5, 6, 8, 13] and confusion matrix [6, 9]. Other than these machine learning methods, some studies used completely different and specific factors to find the performance of their models [9]. Previous studies show that sensitivity, specificity, and accuracy measures were widely used in such studies.

2.4 Available Databases

METABRIC Breast Cancer dataset [6]

- A project of Canada-UK.
- 1,980 primary breast cancer samples are available.
- Clinical and genomic data are available.

The Cancer Genome Atlas (TCGA) Data repository [6]

- A project of the National Cancer Institute and the National Human Genome Research
- There are data around 33 types of cancer
- Genomic, epigenomic, transcriptomic, and proteomic data are available

Genomics of Drug Sensitivity in Cancer (GDSC) dataset [9]

- A project at the Wellcome Sanger Institute (UK) and the Center for Molecular Therapeutics, Massachusetts General Hospital Cancer Center (USA).
- It contained drug response data and genomic markers of sensitivity
- Molecular features of cancers are available.

Iranian Centre for Breast Cancer (ICBC) dataset [1]

- A project of the National Cancer Institute.
- It contained 1189 records of cancer patients, 22 predictor variables, and one outcome variable.

BIDMC-MGH [8]

- A project of Massachusetts General Hospital and Beth Israel Deaconess Medical Centre
- It contain of 392 features
- It contain 116 breast biopsies

Wisconsin Diagnostic Breast Cancer (WDBC) datasets [8]

- A project of University of Wisconsin
- 10 real-valued features are available
- Cell nucleus data are available

Chapter 3: Methodology and Research Plan

3.1 Methodology in Brief

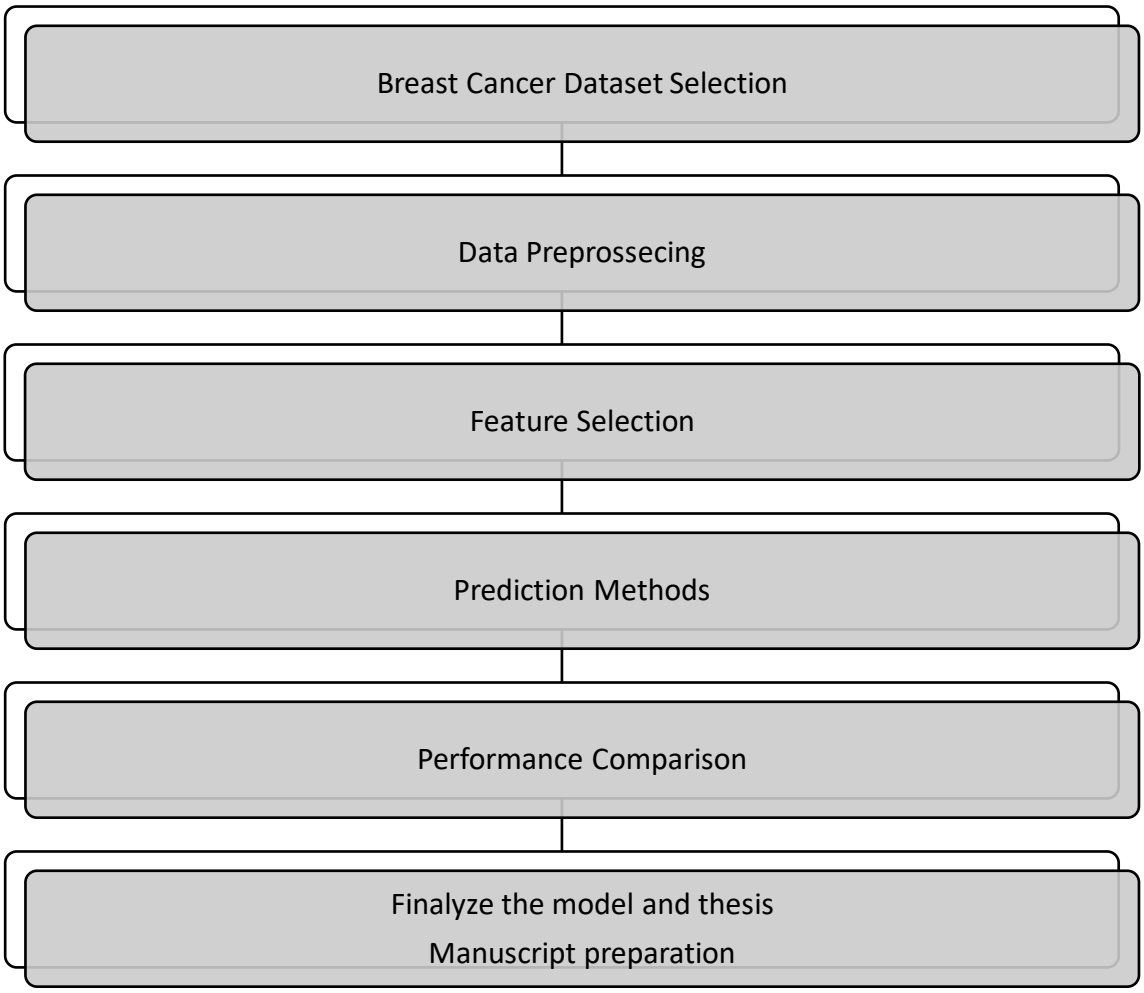

Figure 2 :Methodology in Brief

3.2 Detailed Methodology

3.2.1 **Breast cancer data selection**

From the available datasets, we choose The Cancer Genome Atlas (TCGA) data repository for our research. This data repository consists of several molecular data of around 36 cancers along with the phenotype of the patients. We choose breast cancer for our study. Among the available molecular data of breast cancer, we decided to use the following data in this study:

1. Copy number (gene-level)
2. DNA methylation
3. Exon expression RNAseq
4. Gene expression RNAseq
5. miRNA mature strand expression RNAseq
6. Protein expression RPPA
7. Somatic mutation (SNP and INDEL)

Also the phenotype data contains hundreds of clinical data related to the patients. They could be used as the target of our prediction. Among them we choose the most important ones such as:

1. Cancer stage
2. Cancer subgroups
3. Cancer historical type
4. Availability of cancer tumour ‘

3.2.2 **Data preprocessing**

Data preprocessing is the method used to clean and prepare the dataset to make it better for building and training machine learning models. This is the base to have a better performance from the study. Instead of deleting or removing some data from the dataset, applying these preprocessing methods can give more information for our model. Preprocessing steps such as filling missing values, apply binning methods, normalization, dimensionality reduction can be used in data preprocessing [12]

3.2.3 Feature Selection

Feature selection is a process of selecting the most relevant attributes from the chosen dataset for a given modelling problem. There are three types of feature selection methods such as filter methods, wrapper methods, and embedded methods. We are planning to compare between forward selection, backward elimination, correlation matrix–pearson in our model.

Forward feature selection

This belongs to the wrapper methods. It will start with a null set of features. Then the iterative process will add the selected best feature one by one to improve the model until reach a consistent accuracy.

Backward elimination

This method also an iterative method. It starts with whole features. Then the iterative process will remove the least significant feature at each iteration. This process will continue until a steady performance.

Correlation Matrix–Pearson

The relation between features and target variables is studied using Pearson Correlation. According to the correlation value the selection will be done.

Chi-squared

Relation or independence between variables can be determined using the Chi-Square method. This belongs to Filter methods.

Feature importance

Mostly this method is used with tree-based classifiers and the feature importance property of the model is used to get the importance of each feature. A score is given for each feature and the feature with the highest score is selected.

3.2.4 Apply machine learning methods

Next step is selecting proper prediction algorithms for our study. As we are comparing performance of algorithms on cancer studies, our choice of algorithms is on a wide range as follows:

SVM

SVM is one of the most commonly used machine learning algorithms which inspects data used for classification and regression analysis. It can perform classification, regression and even outlier detection and most of the biological applications take the advantage of it [1, 3, 6, 8, 9, 10].

Here, the classification mechanism is used to separate the instances into different classes based on its target value. SVM tries to find the optimal hyperplane in an N

dimensional space that split data points into separate classes. It is a supervised learning technique. As our data has discrete targets such as cancer stage, cancer subgroups, cancer historical type and availability of tumour, we have selected SVM for our study.

Decision Tree

Decision tree methods are widely used in classification and it is a supervised learning method. The speciality of the models built with this method is, their similarity with human reasoning. A decision tree is built by iteratively splitting the data according to its features and based on the existing classes. As decision tree method is easily understandable and can be visualized, this is another choice in our study. Also this technique is used in previous cancer related studies [1, 10].

K-Nearest Neighbour

KNN is a classification algorithm under supervised learning. It classifies the data based on the class of its nearest points. This is another method used in biological studies.

Random Forests

Random Forest combines the simplicity of decision trees with flexibility, resulting in a vast improvement in accuracy. Here we have selected random forest to compare the performance with other methods. This is also used in biological studies.

Naive Bayes

Naive Bayes is a probabilistic machine learning classifier that is based on the Bayes theorem. There are three types of Naive Bayes such as Multinomial, Bernoulli, and Gaussian. We hope to use each of these methods depends on the phenotype data of our study.

Deep Learning based methods

Multi-Layer Perceptron (MPL)

Multilayer perceptron can be considered as the most basic feed forward artificial neural network based machine learning method in use. It has one or more hidden layers between input and output layers.

The number of neurons in the input and output layer are equal to the number of measurements for the pattern problem and number of classes. All neurons except input layer use activation function, and back propagation is the technique used in model training. MLP can be used on non-linearly separable data. This is used in [8] for classification of benign and malignant proliferative breast lesions.

Recurrent neural network (RNN)

Recurrent neural network is widely used in face detection, generating image description and biological predictions. Generally it is used in sequence prediction problems. In RNN model, same parameters are used in each input. The most common RNNs are long-short term memories (LSTM), which can be used for very complex problems such as translation and speech recognition.

Recursive Neural Networks (RvNNs)

RvNNs are non-linear adaptive models which is very effective in natural language processing. Direct connections are established between the neurons. The output depends on the number of neurons in each layer and the number of connections between them. As a result, the performance, robustness, and scalability of RvNNs are more powerful compared to other types of artificial neural networks. There are three variations in RvNNs such as single layer, multiple layers or a combination of layers. Hence, we can select the one which meets our requirements.

Convolutional Neural Network (CNN)

CNN has a good capability in data analysis although it is originally designed for digital image processing. Convolution and pooling are two basic operations on CNN. Features are extracted from the dataset using convolution operation and the dimensionality of feature maps is reduced using pooling operation. Less pre-processing is required for CNN compared to other algorithms.

Long Short-Term Memory (LSTM) networks

These are a special kind of RNN that has a problem in long term dependencies.

3.2.5 Compare performance

Performance measurement is a very important aspect in our study as our aim is building a better model for cancer predictions. In this study we are using several methods such as Area under the Receiver Operating Characteristic (ROC) curve and confusion matrix for measuring the performance of our model to double confirm our findings

Area under the ROC curve

ROC curve is a graph used to measure the performance of machine learning models. This method is based on varying threshold values and it is a curve of probabilities. Here, True positive rate is plotted against false positive rate. But this method would be inefficient. But using this ROC curve we can achieve better approach considering area under the ROC curve. It measures the entire area under 2-dementional ROC curve and this measurement is a measurement of separability.

It provides an aggregate measurement of all classification threshold value. Rather explaining current absolute value of each data point. AUC ranks the prediction. Also it measures the quality of prediction model. If a model has a higher AUC value, the model is better one.

Confusion matrix

A confusion matrix compares the correct prediction of the model against the wrong prediction of the model. The rows in a Confusion matrix correspond to what the machine learning algorithm predicted, and the columns correspond to the ground truth. The size of the confusion matrix is determined by the number of things we want to predict.

Accuracy

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

$$\text{Accuracy}=\frac{TP + TN}{P+N}$$

3.3 Timeline

Table 1 :Timeline

	6th semester				7th semester				8th semester	
	W 1	W 3 – 11	W 12 – 14	W 15 – 16	W 1	W 2 - 4	W 5 - 10	W 10 - 16	W 1 – 3	W 3 – 8
Literature review	■	■	■	■	■	■	■	■	■	■
Annotated Bibliography		■	■							
Research proposal writing			■	■						
Data collection					■					
Data preparation						■				
Build the model							■			
Experimenting the models								■		
Research project report writing								■	■	■
Research paper writing										■

Chapter 4: PROGRESS TO DATE

4.1 Literature Review

Extensive literature review has been conducted. Literature review will be conducted throughout the research.

4.2 Database Collection

4.2.1 Phenotype Data selection

In the TCGA data they have stored hundreds of clinical data of the patients. We have selected the following data for our study:

AJCC staging of cancer

Cancers are classified according to the state of cancer tumors. AJCC Staging also a classification method that was developed by American Joint Committee on Cancer. Basically, it considers Tumor size, Lymph Nodes affected, Metastases. The method follows TNM (Tumor Nodes and Metastases) Classification of Malignant Tumors which is the accepted standard for classifying the extent of spread of cancer. It contains the stags of the cancers as below.

stage	Tumor	Nodes	Metastases
I	T1 or T2	N0	M0
IIa	T3	N0	M0
IIb	T4	N0	M0
IIIa	T1 or T2	N1	M0
IIIb	T3 or T4	N1	M0
IIIc	Any T	N2	M0
IV	Any T	Any N	M1

Table 2 : cancer stages as defined by the American Joint Committee on Cancer (AJCC) [12]

T1, T2, T3, T4 : According to the size and/or extent of the main tumor. Higher number after T means, the size/growth of the tumor.

N1, N2, N3: Refers to the number and location of lymph nodes that contain cancer. The higher the number after the N means, there are more lymph nodes contain cancer.

M0: Cancer has not spread to other parts of the body.
M1: Cancer has spread to other parts of the body.

PAM50 RNAseq

PAM50 stands for Prediction Analysis of Microarray 50 and here the tumor samples are tested for a group of 50 genes. These results come up with five main sub types of cancers based on the genes. Those subtypes are,

Luminal A**:** grow slowly and have the best prognosis

Luminal B**:** grow slightly faster than luminal A and prognosis is slightly worse.

Basal-like: more common among younger and Black women.

HER2 enriched**:** grow faster than luminal and worse prognosis

Normal-like**:** good prognosis, but slightly worse than luminal A

Neoplasm cancer status

A neoplasms means to an abnormal growth of cells in the body and simply it described as a tumor. So this states show whether the cancer exists, with or without tumor. Hence, this is a binary class problem with two classes.

Pathologic stage

There are two kinds of prognosis staging methods available in cancers. One is pathological stage and this stage applies to the patients who have undergone surgery as the initial treatment for the breast cancer. This staging method includes both clinical data and findings from surgical resections.

Stage	
O	DCIS
IA	T1N0
IB	T0-1N1M1
IIA	T0-1N1, T2N0
IIB	T2N1, T3N0
IIIA	T0-2N2, T3N1-2
IIIB	T4N0-2, T3N1-2
IIIC	Any T N3
IV	symetric

Table 3 :Cancer Stages According to Pathological Stage

Histological_type

Historical type describes a cancer tumor based on how abnormal the cancer cells and tissue look under a microscope and how quickly the cancer cells are likely to grow and spread. It contained type such as,

- Infiltrating Lobular Carcinoma
- Infiltrating Ductal Carcinoma
- Other, specify
- Mixed Histology (please specify)
- Mucinous Carcinoma
- Metaplastic Carcinoma
- Infiltrating Carcinoma NOS
- Medullary Carcinoma

4.2.2 Data set selection

Copy number (gene-level)

Copy Number Variations (CNVs) are deletions or duplications of the structural variant in the number of copies of specific regions of DNA. These alterations in DNA have been considered to be associated with various human cancers. This data is measured on 1080 patients and estimated using the GISTIC2 method.

DNA methylation

DNA methylation can be identified as an important regulator of gene transcription. There are major disruptions between normal cells and malignant cells in their DNA methylation patterns. The data of 345 patients were used. Methylation 27 K data was selected and it was measured experimentally using the Illumina Infinium HumanMethylation27 platform.

Exon expression RNAseq

Exons are important parts of RNA. It is parts of the gene sequence that are expressed in the protein. The data of 1218 patients was selected for this analysis. IlluminaHiSeq data was selected from TCGA which was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform.

Gene expression RNAseq

Gene expression is the process of directing the assembly of a protein molecule using the information encoded in a gene. There are 1218 patient's data. IlluminaHiSeq data from TCGA was selected. It was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform.

miRNA mature strand expression RNAseq

MicroRNAs are non-coding RNAs with important roles in regulating gene expression. IlluminaHiseq TCGA data was selected and data of 333 were used.

Protein expression RPPA

Protein expression is the production of proteins by cells. It will be very useful to get the information about a specific type of cancer. RPPA protein from TCGA was selected. This data was generated and processed at the MD Anderson Cancer Center using RPPA technology.

Somatic mutation (SNP and INDEL)

Somatic mutation is an alteration in DNA that occurs after conception. Gene-level non-silent mutation was selected here.

4.1 Database Preparation

As these data are stored in separate files, we downloaded them separately. Using some python code, we combined every dataset along with their corresponding phenotype data. For example, genomic data of the patient with their cancer subgroup, genomic data of the patients with their cancer stage, etc.. Patient id is the common field between files to combine them.

REFERENCES

[1] A. LG and E. AT, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence", Journal of Health & Medical Informatics, vol. 04, no. 02, 2013. Available: 10.4172/2157-7420.1000124.

[2] André, T., Sargent, D., Tabernero, J., O'Connell, M., Buyse, M., Sobrero, A., Misset, J., Boni, C. and de Gramont, A., 2006. Current Issues in Adjuvant Treatment of Stage II Colon Cancer. Annals of Surgical Oncology, 13(6), pp.887-898.

[3] C. Boeri et al., "Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation", Cancer Medicine, vol. 9, no. 9, pp. 3234-3243, 2020. Available: 10.1002/cam4.2811.

[4] C. Xu and S. Jackson, "Machine learning and complex biological data", Genome Biology, vol. 20, no. 1, 2019. Available: 10.1186/s13059-019-1689-0.

[5] Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W. and Faisal Nagi, M., 2019. "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms." *Journal of Healthcare Engineering*, 2019, pp.1-11.

[6] G. Dubourg-Felonneau et al., "A Framework for Implementing Machine Learning on Omics DataML4H/2018/102", Machine Learning for Health (ML4H) Workshop at NeurIPS, no. 42018102, p. 5, 2018.

[7] Guo, "Transcription: the epicenter of gene expression", Journal of Zhejiang University SCIENCE B, vol. 15, no. 5, pp. 409-411, 2014. Available: 10.1631/jzus.b1400113.

[8] H. Dhahri, I. Rahmany, A. Mahmood, E. Al Maghayreh and W. Elkilani, "Tabu Search and Machine-Learning Classification of Benign and Malignant Proliferative Breast Lesions", BioMed Research International, vol. 2020, pp. 1-10, 2020. Available: 10.1155/2020/4671349.

[9] M. Ding, L. Chen, G. Cooper, J. Young and X. Lu, "Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics", Molecular Cancer Research, vol. 16, no. 2, pp. 269-278, 2017. Available: 10.1158/1541-7786.mcr-17-0378

[10] M. Khawar, N. Aslam, R. Mahtab Mahboob, M. Mirza, H. Jahangir and A. Mughal, "Comparative Study of Machine Learning Algorithms in Breast Cancer Prognosis and Prediction", Comparative Study of Machine Learning Algorithms in Breast Cancer Prognosis and Prediction, no. 20, pp. 125-133, 2020.

[11] N. Laham-Karam, G. Pinto, A. Poso and P. Kokkonen, "Transcription and Translation Inhibitors in Cancer Treatment", Frontiers in Chemistry, vol. 8, 2020. Available: 10.3389/fchem.2020.00276 [Accessed 30 October 2020].

[12] Sharma, A. and Rani, R., 2017. "An Optimized Framework for Cancer Classification Using Deep Learning and Genetic Algorithm." *Journal of Medical Imaging and Health Informatics*, 7(8), pp.1851-1856.

[13] Who.int. 2020. Cancer. [online] Available at: <https://www.who.int/news-room/fact-sheets/detail/cancer> [Accessed 15 October 2020].

[14] Y. Hou and J. Linhong, "Gene Transcription and Translation in Design", Volume 7: 27th International Conference on Design Theory and Methodology, 2015. Available: 10.1115/detc2015-46128

[15] Yepes, S. and Mercedes Torres, M., 2016. Mining Datasets for Molecular Subtyping in Cancer. *Journal of Data Mining in Genomics & Proteomics*, 07(01).

APPENDIX

Table No 1. Table Title
