Unsupervised Techniques for Meta-analysis of Cancer Genomic Data

By

S.D.L.H. Maheeshanake

2015/CS/085

This dissertation is submitted to the University of Colombo School of Computing

In partial fulfillment of the requirements for the

Degree of Bachelor of Science Honours in Computer Science

University of Colombo School of Computing

35, Reid Avenue, Colombo 07,

Sri Lanka

July 2020

Declaration

I, S.D.L.H.Maheeshanake (2015/CS/085), hereby certify that this dissertation entitled "Unsupervised Techniques for Meta-analysis of Cancer Genomic Data" is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

…….. ……..

Date Signature of the Student

I, Mrs. M.W.A.C.R.Wijesinghe, certify that I supervised this dissertation entitled "Unsupervised Techniques for Meta-analysis of Cancer Genomic Data" conducted by S.D.L.H.Maheeshanake in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Computer Science.

…….. ……..

Date Signature of the Supervisor

I, Dr. A.R.Weerasinghe, certify that I supervised this dissertation entitled "Unsupervised Techniques for Meta-analysis of Cancer Genomic Data" conducted by S.D.L.H.Maheeshanake in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Computer Science.

…….. ……..

Date Signature of the Co-Supervisor

Abstract

Cancer is a set of diseases where abnormal cell growth can be identified. The cancer cells are not responding to the cell division or the normal signaling system of human beings. That is the main difficulty in giving treatments and control the growth of those cells. Same cancer has different molecular structures. Different people with the same cancer can show completely different behavior. Personalized medicine is essential for solving that issue. To prepare personalized medicine the identification of subtypes is really important. When considering cancer subtype identification process, the high dimensionality of genomic data is considered as an obstacle. In the biological domain, high dimensionality refers to the high number of genes when compared to the number of samples available.

The state of the art methods for dimensionality reduction sometimes does not accurately address this problem when those applied to some of the biological datasets. The performance and the suitability of them vary with the context where we apply those methods. So those techniques should be evaluated through literature and testing with our datasets. Then those techniques will be compared by analyzing their suitability with the genomic data. In this research, a pipeline has been introduced to reduce dimensionality, clustering, and validating which help in cancer subtype identification tasks. The applicability and suitability of the introduced pipeline are evaluated through a set of internal and external validation methods.

.

Preface

The results in this study rely upon endometrial cancer expression data provided by TCGA Pan Cancer study which are available at cBioPortal. The analysis of the data is entirely my own work which I carried out with the help of google colabs (for initial analysis), tensorflow, and libraries like scikit learn, matplotlib etc.

Acknowledgement

I take this moment to convey my appreciation and gratitude towards my supervisor, Senior Lecturer Mrs. Rupika Wijesinghe, co-supervisor, Senior Lecturer Dr. Ruwan Weerasinghe and advisor Ms. Amali Perera for their commitment and support throughout the research. It was a really good experience to conduct the research under their guidance .I shall be grateful to them for their proper guidance, helpful advice and time they dedicated from the very early stage of the research. Without their help and guidance, this could not have been completed properly.

I would like to express my gratitude to the bioinformatics research team, Ms. Bhagya Madhubhanie, Ms.Dinithi Rajapaksha, Mr.Mohan Anuradha for their help from the beginning. I owe a considerable amount to my family and friends who gave me support and encouragement to make this a success. Finally, I would like to thank all the people whose names have not appeared, but their untiring effort was crucial to complete this study.

Table of Contents

Declaration ii

Abstract iii

Preface iv

Acknowledgement v

Table of Contents vi

List of Figures ix

List of Tables xi

List of Acronyms xii

Chapter 1 - Introduction 1

1.1 Background to the Research 1

1.2 Research Problem and Research Questions 3

1.2.1. Research Problem 3

1.2.2. Research Questions 3

1.3 Justification for the research 4

1.4 Methodology 5

1.5 Outline of the Dissertation 6

1.6 Delimitations of Scope 6

1.7 Conclusion 7

Chapter 2 – Literature Review 8

2.1. Introduction 8

2.2. Biological Data and Bioinformatics 8

2.2.1 DNA, Gene, DNA Sequencing, Gene Expression 8

2.2.2. Protein Synthesis 9

2.3. Tumors 10

2.3.1. Endometrial Cancer 11

2.4. Dimensionality Reduction 12
2.5. Dimensionality Reduction Techniques 12
2.5.1. Feature selection methods 13
2.5.2. Clustering based approaches for feature selection 14
2.5.3. Normal Feature extraction methods 16
2.5.4. Deep learning based feature extraction methods 17
2.6. Clustering 19
2.7. Summary 20
Chapter 3 - Design 22
3.1. Introduction 22
3.2. Research Design 22
3.2.1. Data Gathering 22
3.2.2 Data Preprocessing 23
3.2.3. Reduce the high dimensionality of the data 24
3.2.4. Clustering and Subtype identification 26
3.2.5. Validating the quality of the obtained clusters. 26
3.2.5.1. Internal Validation 26
3.2.5.2. External Validations 28
3.3. High-level Design 29
Chapter 4 – Implementation 30
4.1. Data Gathering 30
4.2. Handling Missing Values 30
4.3. Dimensionality Reduction 31
4.3.1. Using highly varied genes 31
4.3.2. PCA on EC dataset. 31
4.3.2. t-SNE on EC dataset 32
4.3.4. PCA followed by t-SNE on EC dataset 33

4.3.5. Deep Learning based approach.....*33*

4.4. Clustering 33

4.4.1. Identify the optimal number of clusters .The elbow method and silhouette width was used to find the optimal number of clusters......*33*

4.5. Validation 34

Chapter 5 – Results and Evaluation 35

5.1. Overview of the endometrial cancer dataset (TCGA Pan-cancer Dataset).....35

5.2. Applying Principal Component Analysis (PCA) on Endometrial Cancer Dataset 35

5.3. Applying T-Distributed Stochastic Neighboring Entities (t-SNE) [5] method on endometrial cancer dataset 38

5.4. Applying PCA followed by T-SNE 38

5.5. Comparison of the above 3 methods with MNIST dataset 39

5.6. Comparison of the above 3 methods with endometrial cancer dataset. 40

5.7. Applying K-means clustering on endometrial cancer dataset. 41

5.8. Applying K-means clustering on EC dataset – highly varied genes (80%). 43

5.9 Applying K-means clustering on EC dataset – highly varied genes (60%). 45

5.10. Applying Deep learning based method 48

5.11 Validation 49

5.11.1. Internal Validation.....*49*

5.11.2. External Validations.....*50*

Chapter 6 - Conclusion.....52

6.1. Introduction 52

6.2. Conclusions about research questions.....53

6.3. Limitations.....55

6.4. Implications for further research 55

References.....56

List of Figures

Figure 1.1: Research methodology…….....5
Figure 2.1 - deep autoencoder method……..19
Figure 3.1 - Expression data with missing values…….…23
Figure 3.2 - Expression data after removing missing values……..……23
Figure 3.3 - architecture for reducing dimensionality…….……...…..25
Figure 3.4 - Research Design ……..…….....29
Figure 4.1 - Initial Data Format……..……...……...30
Figure 4.2 - Handling missing values..……...……...31
Figure 4.3 - PCA on EC data (i)…….……...32
Figure 4.4 - PCA on EC data (ii)…….……..32
Figure 4.5 - t-SNE on EC data…….…..32
Figure 4.6 - PCA followed by t-SNE on EC data……..33
Figure 4.7 - The Elbow method …….……..……..33
Figure 4.8 – Obtaining Silhouette score ……..…..……..…….……..34
Figure 4.9 – Obtaining validation scores ……..……...……...…..34
Figure 5.1 - Outline of the endometrial cancer dataset……..35
Figure 5.2 -1st 10 principle components with their percentage of explained variance..36
Figure 5.3 - PCA plot for the EC dataset with PCA 1 and PCA2…….…….....37
Figure 5.4 - mostly varied genes with PCA 1……...….37
Figure 5.5 - Application of t-SNE with EC dataset…….……..38
Figure 5.6 - Application of PCA followed by t-SNE on E.C dataset……39
Figure 5.7 - A comparison of the application of PCA and T-SNE to mnist dataset…..40
Figure 5.8 - A comparison of the application of PCA and T-SNE to mnist dataset…..40
Figure 5.9 - Elbow plot for the E.C dataset……...41
Figure 5.10 - K-means with K= 10……42
Figure 5.11 - K -means with K= 2……42
Figure 5.12- Elbow plot for the E.C dataset with highly varied 14000 genes…...……43
Figure 5.13 - K -means with K= 2 highly varied 14000 genes……44
Figure 5.14 - K -means with K= 17 highly varied 14000 genes……44
Figure 5.15 - Elbow plot for the E.C dataset with highly varied 10200 genes…..……45
Figure 5.16 - K -means with K= 2 -highly varied 10200 genes……..…46

Figure 5.17 - K -means with K= 5 -highly varied 10200 genes……..…46
Figure 5.18 - K -means with K= 9 -highly varied 10200 genes……..47
Figure 5.19 - K -means with K= 13 highly varied 10200 genes……47
Figure 5.20 – Identified cluster after applying the deep learning approach ……49

List of Tables

Table 5.1 – Silhouette Distances for different K values ……48
Table 5.2 – Internal Validation Scores for different K values ……...…50
Table 5.3 – External Validation ……51

List of Acronyms

PCA - Principal Component Analysis

t-SNE - t-distributed Stochastic Neighbor Embedding Algorithm

LDA - Linear discriminant analysis

EC - Endometrial Cancer

PSO - Particle Swarm Optimization

CNV - Copy Number Variation

DNA - Deoxyribonucleic acid

RNA - Ribonucleic acid

SNV - single nucleotide variations

UFS - Unsupervised feature selection

UDFS - Unsupervised Discriminative Feature Selection

SUD - Sequential backward selection method for Unsupervised Data

GnRH - Gonadotropin hormone

KNN - K nearest neighborhood

DI - Dunn Index

NMI - Normalized Mutual Information

USFSM - Unsupervised Spectral Feature Selection Method EC Endometrial cancer

Chapter 1 - Introduction

1.1 Background to the Research

The human body is created with millions of cells. The cell is the basic unit of all living tissues. In most of the cells, there's a structure called nucleus. The nucleus contains the genome. The genome is split among 23 parts of chromosomes. Each chromosome contains a long strand of DNA tightly packed around proteins called histones. Within the DNA there are sections called genes. Those genes contain the instructions for making proteins.

When a gene is switched on, an enzyme called RNA polymerase is attached to the beginning of the gene. It moves along the DNA making a strand of messenger RNA out of free basis in the nucleus. The DNA code determines the order in which the free basis is added to the messenger RNA. This process is called transcription. Before the messenger RNA can be used as a template for the production of the protein it needs to be processed. This involves removing and adding sections to the RNA. Then the messenger RNA moves out of the nucleus into the cytoplasm. Protein factories in the cytoplasm called ribosomes are bound to the messenger RNA. The ribosome reads the code in the messenger RNA in order to produce the chain made up of amino acids. There are 20 types of amino acids. Transfer RNA molecule carries the amino acids to the ribosomes. The messenger RNA is read, 3 bases at a time. As each triplet is read, transfer RNA delivers the corresponding amino acid. This is added to a growing chain of amino acids. Once the last amino acid is added the chain falls into a complex 3d shape to form the protein. Information in DNA is stored as a code made up of 4 chemical bases called adenine (A), Guanine (G), Cytosine(C) and Thymine (T).

Cell cycle regulation makes cell growth healthy. It is controlled by genes and environmental factors. Mutations are the changes in nitrogen bases in the DNA sequence. Some of these mutations are good for organisms and some are really bad, but most of the mutations are neutral. The bad mutations lead to produce cancer cells. There are two kinds of mutations. Those are point mutations and frameshift mutations. In point

mutations, a single base is changed. Those are called single nucleotide variations (SNV). This will cause to change only one amino acid. In frameshift mutations, there can be insertion or deletions of at least one nucleotide in the DNA sequence which will affect all amino acids after the mutation when synthesizing proteins. Copy number variation (CNV) is a kind of frameshift mutation. As mentioned above every cell in a particular human has the exact same set of DNA, but there are various cell types such as blood cells, skin cells, nerve cells, fat cells, etc. The process of the cell becoming specialized or cell differentiation is controlled by the way genes are expressed. In a cell, all genes are not expressed. Some may be expressed and some may not. This is like a light bulb. When the switch is turned on the bulb will emit light as the output, same way expression controls act as the switch if they present the gene will be expressed and output proteins. There can be internal factors like hormones and external factors like radiation and alcohol which will influence this process. This study will use the expression data as the data type.

Cancer is a result of the abnormal growth of cells that have the ability to invade or spread to other parts of the body [1]. For cancer to develop, genes regulating cell growth and differentiation must be altered. These mutations are then maintained through subsequent cell divisions and are thus present in all cancerous cells. Gene expression profiling is a technique used in molecular biology to query the expression of thousands of genes simultaneously. [2]

The high dimensionality of genomic data is a huge problem faced in analyzing cancer genomic data as the number of samples or patients is comparatively low but the number of genes is high. A high number of genes reduces the accuracy of the computations related to genomic data. Using proper dimensionality reduction methods that will help to improve the performance of the models which will play a significant role in data analysis and accuracy of results. Hence, this study will help in identifying the most effective dimensionality reduction techniques which will improve identifying better subtypes in cancers using genomic data.

1.2 Research Problem and Research Questions

1.2.1. Research Problem

Personalized treatments are required in order to increase the efficiency and reduce the toxicity of the cancer treatments. It is required to identification of proper subtypes to design personalized medicine. The high dimensionality of the genomic data is an obstacle for this task and also it is hard to visualize data due to noise and many other facts. Since unsupervised learning techniques have to be used in the subtypes identification task, the result can be ambiguous .It is required to identify a proper process to subtype identification with dimensionality reduction, clustering and validation. This study tries to find a pipeline which is connected with previously mentioned 3 tasks in order to help the subtype identification task.

1.2.2. Research Questions

1. How the selected datasets behave with the existing Dimensionality reduction techniques?

The behavior of the biological datasets is different from one another. There is no guarantee on how a particular dataset is reacting with a particular dimensionality reduction method. So the existing algorithms, advantages, and disadvantages will be evaluated.

2. How to develop a pipeline for reduce the dimensionality in cancer genomic data and identifying proper cancer subtypes?

Considering the dimensionality reduction techniques identified from the previous question and with the knowledge of their advantages and disadvantages, a pipeline is introduced to reduce the high dimensionality of data and identify the proper subtypes of the data. The endometrial cancer data is used as a case study.

3. How to validate the results and goodness of the clusters obtained?

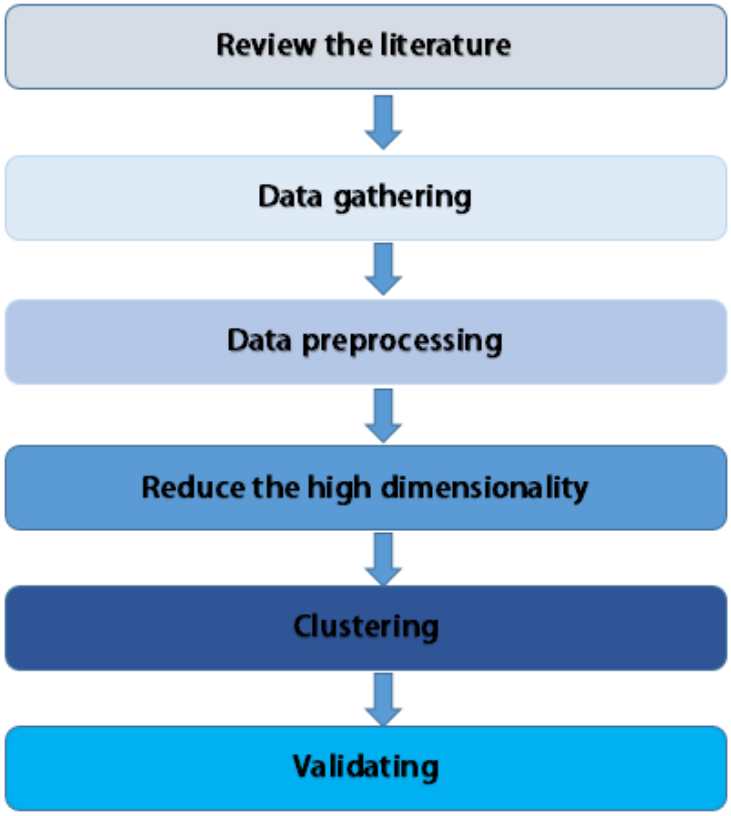
Validate the developed pipeline for identifying the subtypes of the cancer genomic data. The Goodness of the Clustering with endometrial cancer data is validated through internal validations using statistical methods such as silhouette index, and Dunn index. External validations uses the previously identified classes using clinical studies. Methods like Jaccard index and NMI score will be used as external validation methods.

1.3 Justification for the research

Cancers can't be treated through basic principles of medicine. If there are two patients with sick the same treatment can be given those people. Most of the time both of them will get cured. But if they have the same cancer, with the same treatment one patient might get cured and the other patient might die as cancer is a genetic level disease. So personalized medicine is very helpful to solve this issue to some extent. To give personalized medicine we need to identify subtypes of cancer. Through the literature it is identified that the high dimensionality of genomic data makes the subtype identification task more difficult. Therefore the identification of an effective methodology for reduce the high dimensionality and subtype identification will be very helpful for the subtype identification task. Unsupervised techniques are used to do the subtype identification task. It is not an easy task like supervised learning and there is a requirement to identify the hidden patterns of the data. A pipeline which consisting of reducing high dimensionality, Clustering and identifying statistical method to validation will help in subtype identification was proposed in this study.

1.4 Methodology

Figure 1.1: Research methodology



This is an exploratory type research. As in Figure 1.1, the literature review starts at the beginning of the research. Since this research is related to the biological domain, a broad literature review should have to be done. Then the data will be gathered and pre-processed by removing missing values. Then a broad analysis is done to find a suitable dimensionality reduction technique which suits the Endometrial cancer data set. Then the dimensionality of the data will be reduced by applying the identified most suitable dimensionality reduction technique/techniques. The clustering will be done using the selected features by applying the dimensionality reduction technique/techniques. The pipeline of the subtype identification will be described in later sections. The validity of the implemented technique will be measured using set of internal and external validation techniques.

1.5 Outline of the Dissertation

This section contains a brief outline of the overall project. The background of the project has been discussed in this chapter. Research questions are described in this section. The limitations and the scope of the project are also mentioned in this section.

In the second chapter the literature review is presented. In section 2.2 some biological background knowledge was provided. In 2.4 the term dimensionality reduction was described and in section 2.5 some of the dimensionality reduction based approaches for feature reduction and subtyping methods presented in the literature were provided.

In the section 3, the data collection process of the high level diagram of the research design was presented.

In section 4, the implementation from the initial stage was presented up to the subtype identification task.

In the section 5 the results was discussed and the evaluation part will be described.

In the last chapter, overall summary of the research will be presented

1.6 Delimitations of Scope

1. Use the most relevant and effective dimensionality reduction techniques used in genomic data, identified through the literature review.

Since there are many dimensionality reduction methods in the literature, the most effective methods will be selected and used to analyze the behavior of the selected data sets.

2. There are three types of genomic data available in cBioPortal [7] such as gene expression data, mutation data and copy number variation data (CNV). In this study gene expression data will be used for the further analysis as mutation data set is a sparse data set and CNV data represent mutations occur in the tumor cells. Gene expression

data will be useful to identify the underlying structure of the genetic level information about cancer.

3. Use endometrial cancer dataset as a case study.

Since there are some pre identified subtypes in endometrial cancer in the clinical studies there may be more subtypes which could not have been identified due to the lack of computational power and lack of data. Since there is no established ground truth for endometrial cancer, this study tries to refine the subtypes and identification of the existence of new subtypes.

1.7 Conclusion

This chapter was used to give an overall idea of the research. The background and some domain knowledge have been presented in the beginning. It introduced the research questions with a brief description. Then the justification for the research questions were given. The methodology of the research was outlined with a high level diagram and the overall outline of the research was given. The limitations of the project and scope accessed through the project were mentioned. This chapter was written to give a brief outlining of the research and more detailed information will be presented in later chapters.

Chapter 2 – Literature Review

2.1. Introduction

The literature review is presented in order to provide a general idea about the key aspects and theories about the data analyzing the process of biological data and the related works. This chapter will explain theoretical models, key concepts and related work used to build the logic in this research study.

2.2. Biological Data and Bioinformatics

2.2.1 DNA, Gene, DNA Sequencing, Gene Expression

DNA is a genetic material that can be found in every organism. DNA contains the same structure in all organisms. It stores the instructions which are used to create proteins. Each and every cell in a person contains the exact same set of DNA with minor differences. So a person can be uniquely identified by considering the DNA structure.

The gene is a structure that is contained in the DNA. Genes consists of 4 bases called Adenine, Thymine, Cytosine and Guanine that are in a paired structure. There are more than 3 billion of base pairs. The base sequence is stored in the gene which works as the instructions to create proteins. The process of determining the order of the above-mentioned nucleotides in a particular DNA part is called DNA sequencing. The DNA sequencing was previously a time consuming and expensive task. With the computational power and resources, it has become a very easy task and leads to many projects and new observations in bioinformatics.

DNA sequencing and sequence alignments that are done to compare the new sequences with the known sequences sometimes do not give the ability to infer the functionality of the gene. Around 40% of the sequenced genes, functionality cannot be directly determined by comparison with the sequences of pre-identified genes.

Microarrays [16] supports the researchers in the biological domains to infer the functionality of the genes when sequence similarity is not sufficient. The activation level of a gene (expression level) is measured by the microarray technology by measuring the amount of mRNA emitted by a particular gene. The gene expression value increases when the amount of mRNA emission increases. For this study, we work with gene expression data. Genes will be highly expressed or suppressed due to various conditions causing people sick. There can be internal factors like hormones and external factors like radiation and alcohol which will influence the gene expression process.

2.2.2. Protein Synthesis

Protein synthesis is the biological process of making proteins. It consists of 2 main phases called transcription and translation.

Steps in transcription phase:

- DNA in a gene is unwound.
- RNA polymerase or the enzyme copies the DNA strand, and then a complementary mRNA strand will be created.
- mRNA moves from the nucleus to the cytoplasm and the binds with the ribosome to start the translation.

Steps in translation phase:

- mRNA is attached to the start codon (Triplet sequence of the new nucleotides)
- Transfer RNA (tRNA) carries the amino acid which is specific to the codon on the mRNA
- Amino acids are bound together to create a polypeptide and the process continues until the last codon is reached.

2.3. Tumors

An abnormal lump or abnormal growth of cells can be considered as a tumor. When the cells in a tumor are in normal conditions, they are called benign. When the growth becomes abnormal which looks critical and produces a lump, it's called malignant [26]. Those tumors are cancers and need to take immediate action. Tumors can be divided into two basic types namely **Benign** and **Malignant.**

- **Benign -** not harmful, not cancerous, can be removed, remain clustered, no visible spread of cells, nearby tissues are not invaded, growth happens very slowly, treatments are not essential if not health threatening.

- **Malignant -** harmful, cancerous, can be metastasized to other parts of the body, cells grow rapidly, there is a chance for reoccur after the removal, requires effective treatments such as chemotherapy, radiation and immunotherapy medications.

There is a chance for a benign tumor to be transformed into a malignant tumor. In some types such as adenomatous polyps in the colon have a significant risk for converting into cancer. So those benign tumors should be removed in the benign state. Sometimes it is not very clear to state whether a tumor is malignant or benign. Those will be tested by the doctors with the consideration of a set of factors to check whether it is malignant or not.

Cancers can occur due to the cell damages due to mutations which are the changes in the sequence of DNA. Mutations can occur after somebody is born or passed down from parents. Not all mutations are harmful. Some of them are good, some of them are harmful and most of them are neutral.

2.3.1. Endometrial Cancer

According to the American Cancer Society, Endometrial cancer is the most lethal cancer related to the female reproductive organs. It is estimated that 61,880 women will get diagnosed in 2020 due to uterine cancer. It's the fourth most common cancer in the United States. 12610 estimated deaths for a year .In the United States it is considered as the sixth highest common cause of cancer death. There is a more vulnerability for white women for getting affected with endometrial cancer than the black women. But the death rate is higher among the black women than white women. When considering the time from 2006 - 2017 there is a 2% increase in deaths for both black and white women [27].

It is considered that the human endometrium is the primary recipient organ for ovarian steroid hormonal signal and is intricately responsive to these hormones. The normal human cell proliferation, generation and function are regulated by the three main classical ovarian steroid hormones estrogen, progesterone and androgens [28].The hypothalamus which stimulates pituitary gland by releasing Gonadotropin hormone (GnRH) is the starting point of the hormonal control of the endometrium. The Follicle-Stimulating hormones (FSH) are released by the pituitary gland. Those hormones circulated with blood to ovaries and stimulate follicles to grow from primary to secondary .It also leads to produce estrogen hormone which helps to stimulate the growth of the endometrium .The hypothalamus will increase the GnRH and induce the production of Luteinizing hormones (LH) which triggers ovulation and release an egg from ovaries, when the estrogen hormone is increased.

Earlier Endometrial Cancer was broadly classified into 2 groups. Those are Endometrioid Adenocarcinoma and Serous Carcinoma.

Endometrioid Adenocarcinoma is linked with the oestrogen access, obesity, hormone receptor positivity. Adjuvant radiotherapy is often used to treat early stage and Chemotherapy is often used for critical stages. To clinical identifications, postmenopausal bleeding is considered as a sign.

Serous Carcinoma is common in older and non-obese women and more aggressive type than the Endometrioid Adenocarcinoma with the worst outcome. These types of tumors are usually treated with chemotherapy for both early and critical stages.

For this type, bowel dysfunction, pelvic pressure, Bloating can be considered as signs for clinical identification.

The above classification is based on the model proposed by Bokhman in 1983 .Based on the fact whether it is estrogen dependent or estrogen independent it is shown that the current pathological classification and grading systems of high grade endometrial carcinomas are limited in both reproducibility and prognostic ability [29] .Therefore in order to select effective adjuvant therapies it is essential to have a proper classification of the subtypes in the molecule level [13].

2.4. Dimensionality Reduction

In machine learning, dimensionality is simply considered as the number of features. In this domain, dimensionality means the number of genes. In biological datasets, the number of genes is relatively very high when compared to the number of samples. That is considered as the high dimensionality of genomic data. The main reasons for that high dimensionality are noise and redundant features. So the dimensionality reduction techniques are introduced in order to remove these noise and redundant features in order to select the optimal feature subset which will show the best performance when applied to computational models.

2.5. Dimensionality Reduction Techniques

There are two main dimensionality reduction techniques. Those are Feature Selection and Feature extraction.

In feature selection, a subset is selected from the original feature space. In the feature extraction method, the existing features will be converted into a brand new feature. Filter methods, Wrapper methods, embedded methods, Hybrid methods and Ensemble methods are example types for the feature selection methods. Principal Component analysis, t-Distributed stochastic neighbor embedding (t-SNE), autoencoder are some examples for feature extraction There are also supervised feature extraction

methods like Linear discriminant analysis(LDA) [18] which cannot be used with unsupervised data.

2.5.1. Feature selection methods

In feature selection methods the subset of the original feature space will be selected. The study done for feature selection approaches used for gene expression data by Shafa Mahajana, Abhishek and Shailendra Singh [8] has stated five categories for feature selection. The study has compared the filter method, wrapper method, embedded method, hybrid method and ensemble method. The filter method or the open-loop method is really fast and consider the intrinsic properties such as distance, correlation and consistency. It is a classifier independent method. The problem with the filter methods most of them are univariate. Those methods only consider each feature independently. It does not consider the correlations among features. The wrapper method or the closed-loop method uses the help of a classifier. Although it is slower, a high accuracy will be shown by the wrapper methods when compared to filter methods and the dependencies among features are not ignored. The performance of the wrapper methods depends on its learning algorithm. Most of the wrapper methods are multivariate and it is very likely to over fit and need more computational time. Embedded methods contain an inbuilt feature selection mechanism which is embedded into the learning algorithm. It shows a better computational time when compared to wrapper methods and it is less prone to over fit [9]. Embedded methods are also multivariate. But in high dimensional data both embedded and wrapper methods shows high computational cost. The hybrid method consists of effective characteristics of both filter and hybrid methods. It shows better performance than the filter method and low time complexity than the wrapper method. But the hybrid method gives good results only with less sample size when compared to considering both methods independently [10]. The ensemble method tries to separate features into subsets and gives an aggregated output of the group. The ensemble method is considered as a robust method when dealing with high dimensional data.

Rank-based feature selection methods use different criteria to obtain an ordered list by evaluating each and every feature and then select the final feature subset based

on that ordering. In the study which was done by Solorio-Fernández, Saúl Carrasco-Ochoa, J. Ariel, Martinez-Trinidad and Jose Fco, the filter methods which give results based on features ranking have been discussed. These filter methods are widely used because of their scalability and effectiveness. The most relevant filter ranking based UFS methods has been chosen in their study [11]. Those were Variance, Laplacian Score, SPEC, Unsupervised Discriminative Feature Selection method(UDFS), SVD-Entropy, Sequential backward selection method for Unsupervised Data(SUD), Unsupervised Spectral Feature Selection Method(USFSM) which give the output as a set of ranked features . 25 high dimensional datasets from the ASU Feature selection repository are used in this study. Those data consist of a large number of face images, text, biological and few other types of data. The evaluation framework was used which evaluate the above mentioned method with the use of the accuracy of the supervised classifiers .k nearest neighbor (k=3), Support Vector Machine, and Naive Bayes (NB) classifiers are used in the study and those are widely used to validate the unsupervised feature selection approaches. The results say that Variance and USFSM are best for ranking based unsupervised feature selection methods for high dimensional data sets. For the really big high dimensional data sets which consist of more than 6000, Variance is dominating with the quality and the less runtime. For datasets that consist of 6000 or less than 6000 features, USFSM is considered as the best method. Although UDFS gives statistically same results the USFSM [13] is much speedier than the UDFS method [14].

2.5.2. Clustering based approaches for feature selection

A study which was done by Guangrong Hu, Xiaohua Shen, Xiajiong Chen, Xin Li and Zhoujun proposes a novel feature selection algorithm based on clustering [12]. This algorithm doesn't need labelled data and suitable for both supervised and unsupervised methods. Features are grouped into clusters based on their similarity which means features in the same cluster are similar to each other. Unsupervised feature selection approaches play a major role not only in reducing the computational time by forming a reduced feature subset but also it improves the cluster quality. In the algorithm, the feature set is partitioned into separate homogeneous clusters using hierarchical clustering. Maximal Information Compression Index (MICI) [15] has been used as the similarity measure. A representative feature from each cluster is selected. That representative feature is chosen by considering the smallest sum of distances to all other

features in the cluster. For the selected features the clustering will be done again using K Means, SOM and Fuzzy C means clustering algorithms. The accuracy of this algorithm depends on the dataset. As an example, it has given comparatively low accuracy to the protein dataset.

Kar, Subhajit, Sharma, Kaushik Das Maitra, Madhubanti has proposed a dimensionality reduction method which will help in identifying cancer subtypes [16]. Particle Swarm Optimization (PSO) approach is tested with the use of two classifiers K-nearest neighborhood (KNN) and Support Vector Machine. The possibility of a feature to be selected is controlled by defining a threshold value. When considering the classifier performance, K-nearest neighbor classifier has shown better performance than the Support Vector Machine. In there, a candidate solution is represented by each particle. First, the optimal position for each particle will be decided and then the global best particle is found .K subset with the same size will be selected from the entire data set with the use of K-fold Cross-Validation. One subset is selected as a validation set and other K-1 subsets are kept as the training set. This step should be iterated for k times until each set is selected to the validation set. With the combination of both cross-validation accuracy and the number of genes, a fitness value will be obtained. With the findings of the study, it shows the proposed PSO-adaptive KNN based method can be used as a helpful tool for select relevant genes from the gene expression data. But the problem with this method is, the supervised classification has been used. So the unsupervised data sets like endometrial cancer dataset cannot be used with this approach.

The paper PSO Based Feature Selection for Clustering Gene Expression Data [17] suggests a wrapper based unsupervised feature selection method where Particle Swarm Optimization (PSO) is wrapped with K-mean. An initial set of particles will be selected by the PSO first. Each particle consists of a velocity and a position in the search space. The quality of the particles will be measured through a fitness function. Then the PSO and gene expression data mapping will be done. Particles are mapped to a subset of genes while fitness function is mapped to squared error function in K-means based on the velocity and the position update equations, the subsets of genes will be updated in each iteration. Then the PSO based feature selection algorithm is applied. Then the K means will be applied to selected subsets and the fitness values are calculated. To get

the initial centroid of K-means, the K-means++ algorithm will be used. An Iterative process will be run in order to minimize the squared error function .K means serves as the wrapper algorithm when the quality of the subsets of genes are evaluated. The process of updating velocity and position is executed iteratively until the best subset of features is found. This method will not be suitable for high dimensional data because of K-means is used as the algorithm in this wrapper method. Due to that, it will show a high computational cost. Because the time complexity of K-means will be increased with the number of data points.

2.5.3. Normal Feature extraction methods

Feature extraction methods create brand new components as features from the original feature space. Principal Component Analysis (PCA) [4] is a dimensionality reduction technique which is used to remove the number of dimensions in datasets. Some correlation between different dimensions is used to provide a minimum number of variables which contains the maximum number of variance about the distribution of the dataset. It's a mathematical method which makes use of eigenvalues and eigenvectors of the data matrix.

t-Distributed Stochastic Neighbor Embedding (t-SNE) [5] is another dimensionality reduction method and a really good for visualizing high-dimensional datasets. This is based on probability. PCA can only capture linear structures while t-SNE is suitable for both linear and non-linear structures. t-SNE is a technique which can be used to minimize the divergence between the two distributions.

The paper Application of t-SNE to human genetic data [19] show the advantages of T-SNE over principal component analysis. PCA is highly dependent on the outlier data samples. If most of the samples belong to a homogeneous population and minority samples from a totally different population, the appearance of those minority samples which have different genetic nature can create a huge change in the principles axes. It can also affect the distribution of the data samples along with the main principal components. The reason for that is the local data characteristics cannot be captured by PCA since it is a linear holistic dimensionality reduction methods.

The main goal of t-SNE is to preserve the pairwise distance in the high dimensional space into 2 dimensions or 3 dimensions .t-SNE has some in-built advantages such as capturing local data characteristics. The subtle data structures are also revealed in the visualization [5]. t-SNE is considered as an embedding visual analytic algorithm where the similarity and dissimilarity among the data points are measured in the low dimensional embedding space.

In T-SNE the similarity scores are added up to one in order to scale the values. The width of the normal curve deviates with the density of data near the point of interest. The less dense regions will have wider curves. If there a cluster which has half of the density compared to other clusters. The second curve of the second cluster is as wide as the other curves. The Scaling of the similarity values will make them the same for both clusters. That indicates the scaled similarity scores relatively tight clusters are the same as the relatively loose cluster. The t- distribution curve which has taller ends than the normal curve will handle the effect of outliers, unlike in the PCA.

2.5.4. Deep learning based feature extraction methods

In the study done by Young, Jonathan D. Cai, Chunhui Lu and Xinghua propose a deep learning approach using deep autoencoder/Deep Belief network to identify subtypes of cancer called glioblastoma[25]. 17 different cancer types and non-cancer organ-specific tissue controls have been used in their study. The sample size was 7528 with 15404 genes. The expression values of a gene in the tumor have been discretized to 0 or 1 based on whether or not the expression value significantly varies with the expression obtained in normal tissue. For genes that has a low expression variance in normal cells (standard deviation of expression < 0.2) used 3 fold change to check if the genes were differentially expressed in tumor cells.

The expression changes have been masked due to copy number alterations. They have discovered co-occurring of gene expression up regulation with a corresponding copy number amplification and co-occurring of gene expression down regulation with a corresponding copy number deletion. So the gene expression changes have been masked because the co-occurrences give an idea that those changes were occurred due to DNA

copy number alterations. Those are not originated by the signaling system. So the identification of genes that are expressed by the cellular signaling system is encouraged rather than the genomic alterations.

Then the feature selection is done. The genes which have a low Bernoulli variance due to their general lack of information. Then cancer-specific or tissue-specific highly correlated features have been removed. For that, they have used the Pearson correlation coefficient. If any gene with a Pearson correlation coefficient related to cancer or tissue type label, which is greater than 0.85 those will be removed .The reason for that is, they have chosen to build a generalized model. Due to the presence of cancer specific genes makes the generalized model biased. Then the training will be done with the deep learning model.

In a supervised model, there is an output. The weights are trained using back propagation considering the specific outputs. Since this is an unsupervised approach due to the unavailability of classes, deep autoencoders can only be trained using input data. The number of nodes in each layer of deep autoencoders is reduced and it leads to creating a feature bottleneck which performs a dimensionality reduction. As in figure 2.1 the input is given from the first layer and changes the weights as encoding input to the next layer. Then train the weights as decoding the input back to the first layer in order to capture the hidden patterns of the data.

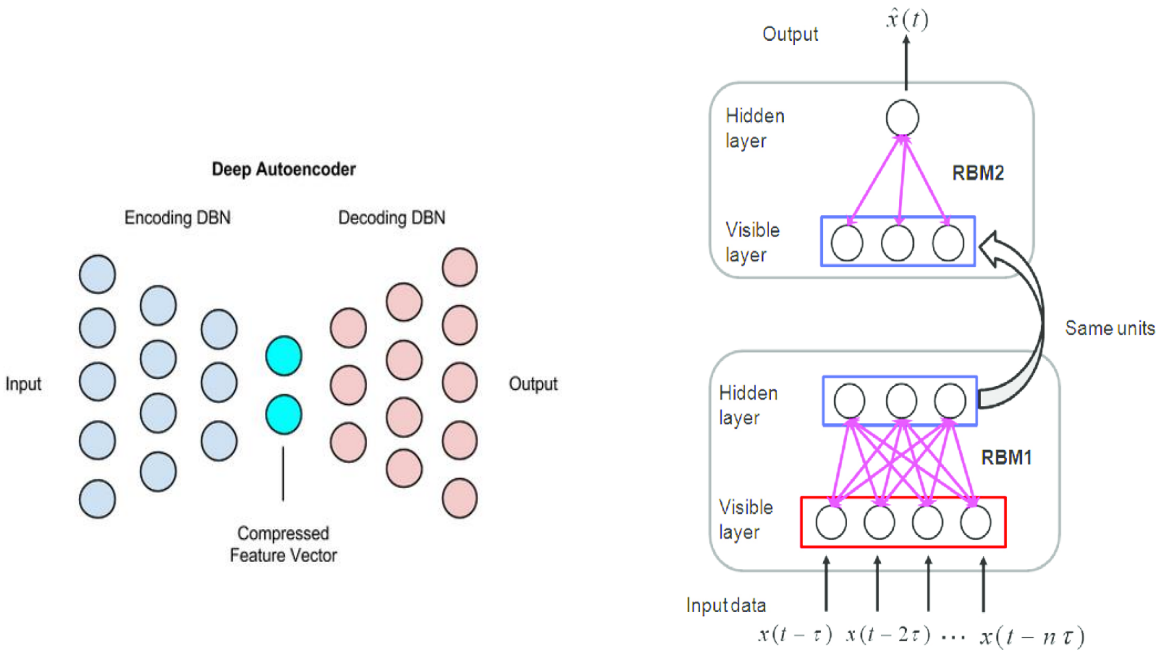


Figure 2.1 - deep autoencoder method

Another deep learning approach which uses semi supervised learning was designed by R. Chen, L. Yang, S. Goodison, [35] was used to identify the subtypes of the Breast Cancer using the METABRIC breast cancer dataset. Previously identified breast cancer subtypes were used to train their supervised model. Cancer subtypes identification can be considered as a supervised learning task where a supervised model can be trained using the existing sub classes and predict the class of new patient. There may be novel subtypes which can occur due to the genetic alterations. Supervised model does not allow to do that task. The method consists of 3 parts. The first part row genomic data is mapped to the representational space .The first part, second part uses existing classes to train weights and third part generates the subtyping results. The last layer of the representational space was used as the clustering module. They have shown that 11 cluster were obtained using heat maps survival analysis. The supervised learning model will be used in our study to reduce the dimensionality.

2.6. Clustering

Clustering plays a significant role in the process of subtype identification. Some other advantages of clustering are viewing and analyzing large amount of data as a single entity, easy interpretation and visualization of the data etc. The clustering algorithms are divided into two main groups as partitioning and agglomerative hierarchical clustering. In partitioning algorithms such as K-means, K-Medoids and CLARA algorithms, need to have predefined number of cluster. In hierarchical clustering, it does not need any pre-defined classes. The number of clusters are identified by creating a dendogram. Most of the studies uses this type of clustering methods due its ease of use and availability of

implementation rather than the importance. Work done by Souto, I. G. Costa,D. S. A. de Araujo have compared 7 clustering algorithms named hierarchical clustering with single, complete and average linkage, k-means, mixture of multivariate Gaussians, spectral clustering and a nearest neighbor-based method and 4 proximity measures, Pearson's correlation coefficient, cosine, Spearman's correlation coefficient and Euclidean distance[36].They have shown that the K-means is the most effective algorithm for gene expression data.

2.7. Summary

High dimensionality in cancer genomic data makes the cancer subtypes identification task difficult. Dimensionality reduction techniques are used to solve that problem by removing the noise and redundant data. There are two main types of dimensionality reduction methods. Those are feature selection and feature extraction methods. In feature selection, a subset is selected from the original feature space. In the feature extraction method, the existing features will be converted into a brand new feature. In [8] five main feature selection methods have been compared. Filter methods are fast but those are univariate. Wrapper methods show high accuracy compared to most of the filter methods while it interacts with a classifier. So the results depend on its learning algorithm. The embedded method contains an in-built feature selection mechanism which is embedded into the learning algorithm. It shows a good computational compared to the wrapper method. But both of those methods show high computation cost when they are dealing with high dimensional data. Hybrid method contains the strengths of both filter and wrapper methods. It shows better performance than both filter and wrapper methods. But the hybrid method gives good results only with less sample size when compared to considering both methods independently. The ensemble method is considered as the best from those methods when dealing with high dimensional data. The filter methods which give results based on features ranking have been compared in another study [11]. Those were Variance, Laplacian Score, SPEC, UDFS, SVD-Entropy, SUD, and Unsupervised Spectral Feature Selection Method (USFSM) which give the output as a set of ranked features. This study says the variance is best for datasets with more than 6000 samples and USFSM is best for datasets less than 6000 samples. Since ensemble method and variance are feature selection methods

those methods have to search the entire feature space multiple times. So the computational time is high when they are compared with feature extraction methods.

The clustering methods have also given good results as a dimensionality reduction. The method proposed in [12] has shown some good results with but it was not accurate with some datasets. In [16] Particle Swarm Optimization (PSO) approach is tested with the use of two classifiers K-nearest neighborhood (KNN) and Support Vector Machine. Since that study has used supervised classifiers, the unsupervised dataset such as EC dataset cannot be tested with them. In [17] wrapper based unsupervised feature selection method where Particle Swarm Optimization (PSO) is wrapped with K-mean was discussed. This method will not be suitable for high dimensional data because of K-means is used as the algorithm in this wrapper method. Due to that, it will show a high computational cost. Because the time complexity of K-means will be increased with the number of data points.

Feature extraction methods are better when compared to feature selection methods. When PCA and t-sne compared T-sne has shown better performance. PCA is very sensitive to outliers. t-SNE reduces that problem because of its shape of the t-distribution curve. Also, t-SNE is better for samples with varied densities. In the results section in this report contains the result when PCA and t-SNE applied to endometrial cancer dataset hose results show that both of the methods fail to obtain any cluster. In [25] the deep autoencoder method is used and it has shown better performance than all the other methods. Since it is not possible to obtain expression values of the non-cancer organic specific tissue controls for endometrial cancer this method is not applicable. A method which uses supervised and unsupervised learning was introduced by R. Chen, L. Yang, S. Goodison, uses the prior biological knowledge can be considered as a better method when compared to previously mention methods. The supervised learning model where weights are trained with the existing classes will be used in this study as the dimensionality reduction method. The clustering methods were divided into two main categories named hierarchical and partitioning method. The study [36] has shown that, K-means has high performance with gene expression data when comparing with different types of clustering algorithms.

Chapter 3 - Design

3.1. Introduction

This chapter describes the design and the methodology of the proposed solution to the Research problems.

3.2. Research Design

The Research Design is basically divided into 5 phases.

1. Data Gathering
2. Data Preprocessing
3. Reduce the Dimensionality
4. Clustering
5. Validation

3.2.1. Data Gathering

The expression data was used in this study due to its benefits, when applied to computational model other than copy number alteration data and single nucleotide variation data. The reason is the some parts of genes are not revealed to outside because it is coiled around a protein called histone .The Endometrial cancer expression data set in the TCGA pan cancer study were downloaded from cBioPoratl [7] which is an open source portal for thousands of biological relevant datasets . cBioPortal is an open platform for exploring multidimensional Cancer Genomics Data, integrated with the cancer genome atlas project (TCGA) .There were 4 identified types with clinical experiments was downloaded and the previously mentioned Endometrial cancer expression dataset was mapped to that based on sample id .

3.2.2 Data Preprocessing

First the **missing values** are removed from the data set.

TCGA-2E-A9G8-01	TCGA-4E-A92E-01	TCGA-5B-A90C-01	TCGA-5S-A9Q8-01	TCGA-A5-A0G1-01	TCGA-A5-A0G2-01	TCGA-A5-A0G5-01	TCGA-A5-A0G9-01	TCGA-A5-A0GA-01	TCGA-A5-A0GB-01	TCGA-A5-A0GE-01	TCGA-A5-A0GH-01
0.1763	3.8324	-0.5067	0.2243	-1.3706	-0.0911	-0.5700	-0.1943	-0.6534	-0.1156	-0.7175	-0.1705
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.4341	0.7793	-0.5292	1.3956	-0.7178	-0.7367	-1.0066	-0.5195	-0.7287	-0.9655	-0.2538	-0.9509
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
-0.1840	-0.1840	-0.1840	-0.1840	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3.1 – Expression data with missing values

As shown in Figure 3.1 the missing values are marked with NaN in the cells. Those rows with missing values should have to be removed and remaining genes will be used for further experiments.

TCGA-2E-A9G8-01	TCGA-4E-A92E-01	TCGA-5B-A90C-01	TCGA-5S-A9Q8-01	TCGA-A5-A0G1-01	TCGA-A5-A0G2-01	TCGA-A5-A0G5-01	TCGA-A5-A0G9-01	TCGA-A5-A0GA-01	TCGA-A5-A0GB-01	TCGA-A5-A0GE-01	TCGA-A5-A0GH-01
0.1763	3.8324	-0.5067	0.2243	-1.3706	-0.0911	-0.5700	-0.1943	-0.6534	-0.1156	-0.7175	-0.1705
0.4341	0.7793	-0.5292	1.3956	-0.7178	-0.7367	-1.0066	-0.5195	-0.7287	-0.9655	-0.2538	-0.9509

Figure 3.2 – Expression data after removing missing values

As shown in figure 3.2, the 5 rows in figure 3.1 has been reduced to 2 rows and any cell with the symbol ‘NaN’ cannot be seen.

3.2.3. Reduce the high dimensionality of the data

Reducing dimensionality of the gene expression data is also a part of the data preprocessing. But due to the importance and the critical play rolled by the high dimensionality as mentioned in previous chapters, it was considered as a separate section. The high dimensionality of the data is considered as one of the most critical things when considering the gene expression data. The problem is the number of genes are comparably high with the number of samples. This problem leads to many problems and inefficiency in computational models. In this research the most significant dimensionality reduction methods are observed and applied to the task of reducing the high dimensionality in order to make computational models efficient and to help tasks like cancer subtype identification.

First the existing dimensionality reduction methods will be identified from the literature. Then the most efficient methods for genomic data are selected. From the literature review, using highly varied genes, Principal Component Analysis (PCA)[4] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [5] and Deep learning based approaches are applied to the datasets and their functionality with those methods were measured. The first 80% of the genes which have the highest coefficient of covariance value were selected as the highly varied genes. Although it has not given well separated clusters, there is some effect on reducing the dimensionality. This highly varied data set is used in the deep learning approach for reduce the dimensionality further.

Coefficient of covariance = Standard Deviation / Mean

The supervised learning model used in the study which was done by R. Chen, L. Yang, S. Goodison, and Y. Sun[35] was chosen to reduce the dimensionality of the endometrial cancer after extracting the highly varied 14000 (80%) of the data. There were previously identified 4 subclasses for the endometrial cancer. Those 4 classes were used to train the weights of the neural network.

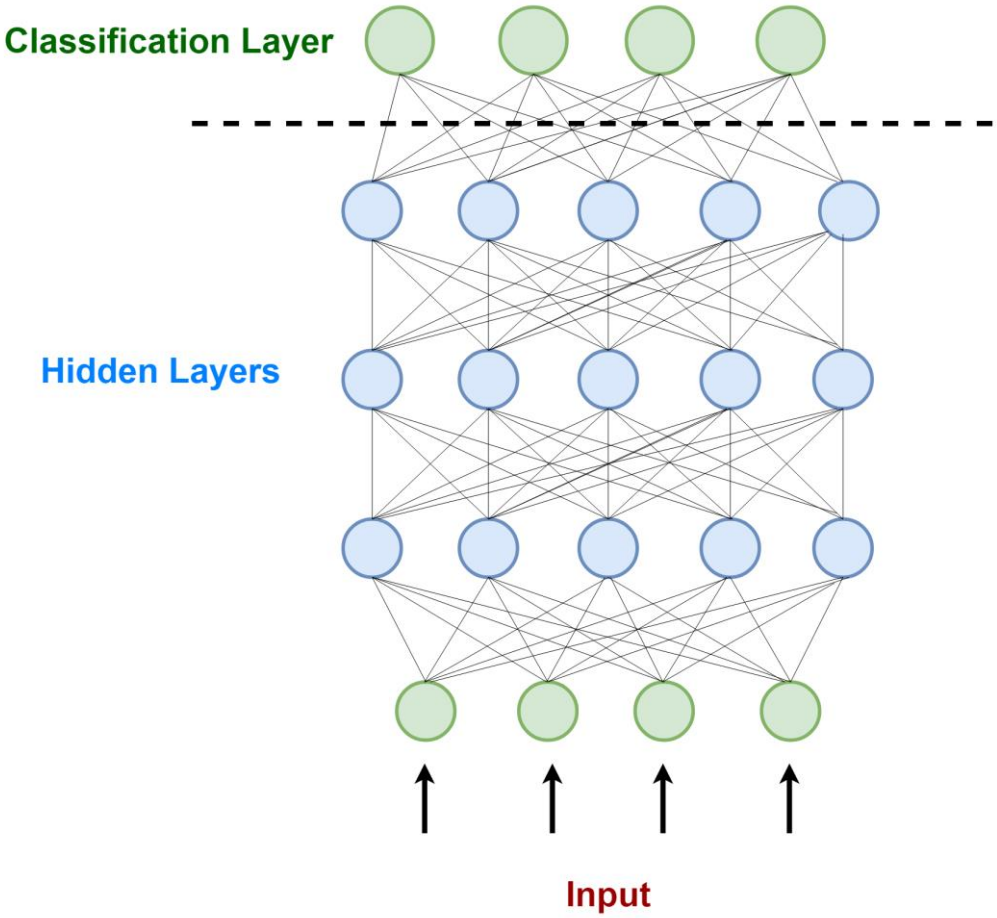


Figure 3.3 architecture for reducing the dimensionality

The genes are given as the input for the neural network. The row genomic data is mapped to the representation space. The weights are trained with the previously identified subclasses. The final layer of the representation space which consists of low dimensional representation of the input genes, are selected and it will be used to clustering which will help in cancer subtypes identification task.

3.2.4. Clustering and Subtype identification

This is one of the most critical parts of this process. After all preprocessing steps are done, then the clustering should be done in order to check whether the possibility of identifying subtypes. The K-means method is selected as the clustering algorithm to this study with its effectiveness in genomic data identified from the literature. The proper number of clusters can be estimated using the silhouette method and the elbow method [31] which will be discussed in the results section.

3.2.5. Validating the quality of the obtained clusters.

The validation part was done using both internal and external validations.

3.2.5.1. Internal Validation

Internal validations are used to measure the quality of the clustering structure without referring to any external information. Here the Dunn Index and the Silhouette score are considered as the internal validation methods. It considers the **low intra** cluster similarity (closeness of the members of the same cluster) and **high inter** cluster similarity (away from the data points of the different clusters).

Dunn Index (DI)

Dunn Index [32] was found by J.C.Dunn in 1974. It is an internal evaluation metric for measuring the goodness of the clustering. This measures the law variance among the members in the same clusters and the high variance between the two data points in 2 different clusters.

The below shown formulas can be used to measure different methods of defining the diameter of the clusters. Let cluster of vectors be Ci and x and y are assigned to the same cluster. Assume x and y are 2 n dimensional feature vectors.

- To calculate the maximum distance between 2 clusters

$$\Delta_i = \max_{x,y \in C_i} d(x, y)$$

- From the below equation, mean distance calculated among the all pairs
- The below equation calculates the distance to the all points from the mean

$$\Delta_i = \frac{\sum_{x \in C_i} d(x, \mu)}{|C_i|}, \mu = \frac{\sum_{x \in C_i} x}{|C_i|}$$

$$\Delta_i = \frac{1}{|C_i|(|C_i| - 1)} \sum_{x,y \in C_i, x \neq y} d(x, y)$$

With all above mentioned notations the Dunn Index is defined as below

Assume **δ(**Ci, Cj) is the metric for inter cluster distance when consider the clusters Ci and Cj and m is the number of clusters .

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k}.$$

Silhouette Score

Silhouette score is another internal measurement for measuring the goodness of the cluster structure. The mean silhouette coefficient for all samples are calculated here.

The mean intra cluster distance (a) and the mean nearest cluster distance (b) which is the distance between a sample and the closest cluster that the samples are not belonging to, are considered for calculating the silhouette score.

The equation **(b-a)/Max(a,b)** gives the silhouette coefficient for a sample . We use the function in the scikit-learn library [33] to get the mean silhouette coefficient over all the samples.

3.2.5.2. External Validations

The external validations refer to the validation done with the use of external information such as clinical variables or the existing classes which have been identified previously. In this study The Jaccard index and the Normal Mutual Information Score are used as the external validation measures. The pre identified class labels for endometrial cancer using clinical studies are used for this validation.

Jaccard Index

When considering the goodness of clustering, the true negative rate is considered as an obstacle for an accurate value due to most of the pairs can belong to that category. The true negatives are neglected when calculating the jaccard index. The count of pairs which are in the same class or belong to same cluster is considered as true positives.

$$Jaccard\ Index = \frac{True\ positive}{True\ positive + False\ positive + False\ negative}$$

Normalized Mutual Information (NMI)

Normalized mutual information score is considered as a good measurement for evaluate the goodness of the clustering. In NMI mutual information is scaled between 0 and 1. This function is normalized by some generalized mean of true labels and predicted labels that are determined by the average method.

3.3. High-level Design

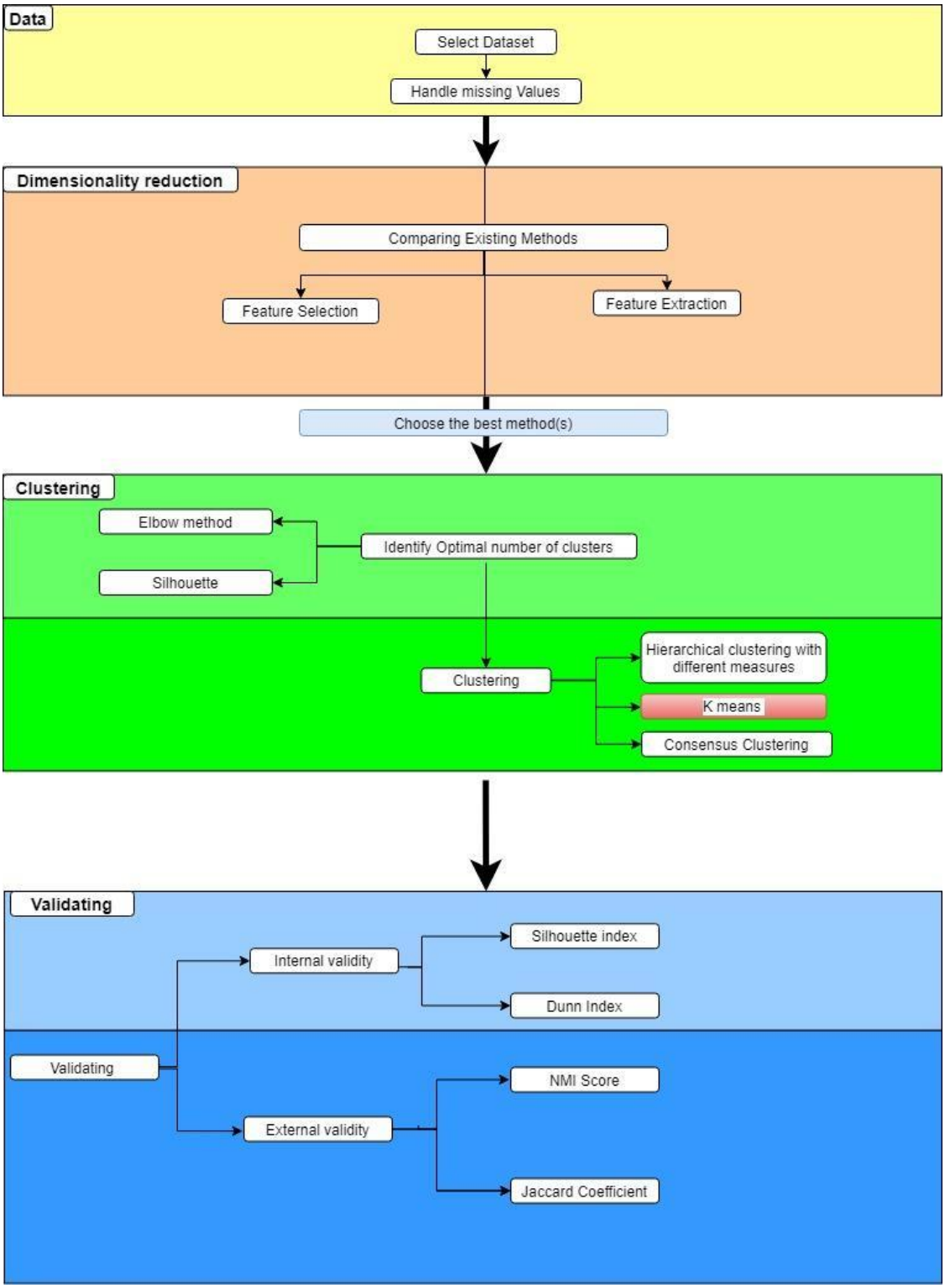


Figure 3.4 - Research Design

Chapter 4 – Implementation

4.1. Data Gathering

The data downloaded from the cBioPortal [7] initially was in txt format with space separated.

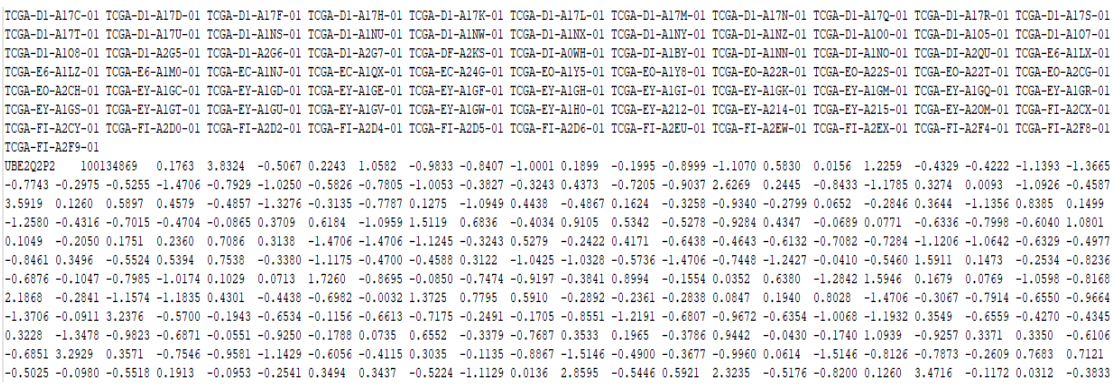


Figure 4.1 – Initial Data Format

The txt formatted file then converted into a CSV format.

4.2. Handling Missing Values

The gene expression data contains many missing values. Those missing values can occur due to experimental errors. Those missing values of the datasets should be handled in order to apply them into computational models. There are several methods to handle missing values such as replacing the missing values with 0. But it can change the patterns of the genetic information. So the missing values will be removed from the dataset. As described in previous chapter missing values can be seen with the 'NaN' symbol in the dataset. After removing the missing values, the genes (rows) will be removed from the dataset.

```
datawithnan= pd.read_csv('endometrial_with_classes_latest.csv')


data1 = datawithnan.dropna() # remove missing values

data1.head()
```

```
/usr/local/lib/python3.6/dist-packages/IPython/core/interactiveshell.py:2718: DtypeWarning: Columns (0) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

	Entrez_Gene_Id	TCGA-2E-A9G8-01	TCGA-4E-A92E-01	TCGA-5B-A90C-01	TCGA-55-A9Q8-01	TCGA-A5-A0G1-01	TCGA-A5-A0G2-01	TCGA-A5-A0G5-01	TCGA-A5-A0G9-01	TCGA-A5-A0GA-01	TCGA-A5-A0GB-01	TCGA-A5-A0GE-01	TCGA-A5-A0GH-01	TCGA-A5-A0GI-01	TCGA-A5-A0GJ-01	TCGA-A5-A0GM-01
1	100133144	12.1093	40.1548	19.6761	23.1153	8.216554	14.583662	3.265217	14.231275	24.464730	14.589216	7.588263	16.961870	25.961002	14.429018	25.792647
2	100134869	19.3502	62.3042	11.3248	19.9139	1.175402	16.207847	10.581103	14.995386	9.601383	15.920322	8.849191	15.274885	7.231641	2.955006	9.280745
3	10357	83.4812	125.2100	223.7690	58.2272	222.588260	232.384702	104.594180	175.030794	295.593088	494.316614	343.047569	638.140633	247.132870	84.600636	393.245294
4	10431	1376.4800	542.8050	729.4690	896.7300	1058.511259	966.389286	1153.505484	1321.093038	1787.289322	1381.730556	1982.624556	1117.242567	1003.367211	1340.025753	1441.973135
6	155060	405.8790	488.1600	176.1990	635.1120	131.235477	126.734391	62.382171	178.511928	128.654801	72.201305	241.875788	75.666311	842.238948	851.588409	403.636342

Figure 4.2 – Handling missing values

4.3. Dimensionality Reduction

4.3.1. Using highly varied genes

For the datasets that contains more than 6000 genes, it is considered that the variance based feature selection method is better than the other feature selection methods. So the highly varied 14000 genes highly varied 10200 for clustering. The genes with higher coefficient of covariance will be selected as the highly varied genes.

Coefficient of covariance = Standard Deviation / Mean

4.3.2. PCA on EC dataset.

After the missing values are removed the data will be scaled .The data should be transposed because the scale function considers rows as samples as in Figure 4.3.

```
[ ]  scaled_data = preprocessing.scale(data.T)

     pca = PCA(n_components = 10) # create a PCA object
     pca.fit(scaled_data) # do the math
     pca_data = pca.transform(scaled_data)

[ ]  print(pca_data)

[[ 2.51532030e+01 -2.39135643e+01 -4.27432251e+01 ...  2.88759183e-02
  -3.74130053e+00 -1.66167827e+01]
 [-4.80120735e+01  9.97227787e+00  2.82543330e+01 ...  3.22478392e+00
   2.40724015e+01 -2.96936310e+01]
 [-3.55749685e+01 -4.68465870e+01 -1.07102343e+01 ... -9.00419530e+00
  -5.11227560e+00 -1.79314732e+01]
 ...
 [-3.69737571e+01  1.12160679e+01  1.29414227e+01 ...  1.69690259e+01
   1.91542599e+01 -4.93411210e+00]
 [-1.13912018e+01 -1.33447762e+01  4.40442692e+00 ... -1.72525713e+01
  -2.25974667e+00  1.34564686e+01]
 [-4.26821029e+01 -5.20335391e+01  1.55434453e+01 ...  5.33611842e+00
  -3.31895709e+00  3.19254253e+01]]
```

Figure 4.3 – PCA on EC data (i)

Then the bar plot is obtained which shows the first 10 principle components with their variance as in Figure 4.4.

```
[ ]  per_var = np.round(pca.explained_variance_ratio_* 100, decimals=1)
     labels = ['PC' + str(x) for x in range(1, len(per_var)+1)]

[ ]  plt.bar(x=range(1,len(per_var)+1), height=per_var, tick_label=labels)
     plt.ylabel('Percentage of Explained Variance')
     plt.xlabel('Principal Component')
     plt.title('Scree Plot')
     plt.show()
```

Figure 4.4 – PCA on EC data (ii)

Then, PCA plot will be created using the first 2 principle components.

4.3.2. t-SNE on EC dataset

Figure 4.5 – t-SNE on EC data

```
[ ]  scaled_data = preprocessing.scale(data.T) # we use transpose because the scale function expects the samples as rows

[ ]
     tsne = TSNE(n_components=2, n_iter=2500, random_state=0)

     tsne_data =  tsne.fit_transform(scaled_data)

[ ]  plt.scatter(tsne_data[:,0], tsne_data[:,1])

     plt.show()
```

4.3.4. PCA followed by t-SNE on EC dataset

```
[ ]  pca = PCA(n_components = 250)
     pca.fit(scaled_data)
     pca_data = pca.transform(scaled_data)

     print('Cumulative explained variation for 250 principal components: {}'.format(np.sum(pca.explained_variance_ratio_)))
```

```
Cumulative explained variation for 250 principal components: 0.8850496691566111
```

```
[ ]
     tsne = TSNE(n_components=2, n_iter=2500, random_state=0)

     tsne_data =  tsne.fit_transform(pca_data)
```

Figure 4.6 - PCA followed by t-SNE on EC data

4.3.5. Deep Learning based approach

The dataset was divided in to 80% -20% as train set and test set. The genes will be given as the input to the supervised model. The row genomic data is mapped to the representational space as in [35] .The weights are trained with the previously identified subclass. The last layer of the representation layer is used to cluster the data in order to identify subtypes in an unsupervised manner.

4.4. Clustering

4.4.1. Identify the optimal number of clusters .The elbow method and silhouette width was used to find the optimal number of clusters.

```
[ ]  plt.figure(figsize=(10, 8))
     from sklearn.cluster import KMeans
     wcss = []
     for i in range(1, 21):
         kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
         kmeans.fit(dataset1_standardized)
         wcss.append(kmeans.inertia_)
     plt.plot(range(1, 21), wcss)
     plt.title('The Elbow Method')
     plt.xlabel('Number of clusters')
     plt.ylabel('WCSS')
     plt.show()
```

Figure 4.7. The Elbow method

```python
from sklearn.metrics import silhouette_score
from sklearn.metrics.pairwise import euclidean_distances

def silWidth(kmeans, X):

    kmeans.fit(X)
    silScore = silhouette_score(X, kmeans.labels_, metric='euclidean')

    return silScore
```

Figure 4.8. Obtaining the Silhouette score

4.5. Validation

Dunn index and the Silhouette Score was used as internal validation methods and Jaccard Score and NMI Score was considered as external validation methods.

```python
print('--------------------NMI Score--------------------------')
print(normalized_mutual_info_score(arrr,kmeans.labels_))
print('-------------------------------------------------------')

print('--------------------Jaccard Score----------------------')
print(jaccard_score(arrr,kmeans.labels_,average='micro'))
print('-------------------------------------------------------')


print('-------------------- Dunn Index  ----------------------')
print(dunn_fast(data.train.data,kmeans.labels_))
print('-------------------------------------------------------')
print(silhouette_score(data.train.data,kmeans.labels_, metric='euclidean'))
```

Figure 4.9. Obtaining Validation Scores

Chapter 5 – Results and Evaluation

5.1. Overview of the endometrial cancer dataset (TCGA Pan-cancer Dataset)

	Entrez_Gene_Id	TCGA-2E-A9G8-01	TCGA-4E-A92E-01	TCGA-5B-A90C-01	TCGA-5S-A9Q8-01	TCGA-A5-A1OH-01	TCGA-A5-A2K2-01	TCGA-A5-A2K3-01	TCGA-A5-A2K4-01	TCGA-A5-A2K5-01	TCGA-A5-A2K7-01
1	100133144	12.1093	40.1548	19.6761	23.1153	18.1558	4.8869	0.0000	12.2708	0.0000	0.0000
2	100134869	19.3502	62.3042	11.3248	19.9139	29.7106	5.7258	7.4006	5.5286	19.5095	14.9341
3	10357	83.4812	125.2100	223.7690	58.2272	217.0690	149.8170	126.4290	191.6830	121.6440	264.3000
4	10431	1376.4800	542.8050	729.4690	896.7300	1173.6500	1297.2800	1946.3500	988.1340	2217.9500	1312.8000
6	155060	405.8790	488.1600	176.1990	635.1120	133.9520	67.2142	79.5560	209.2770	227.9820	146.0740

5 rows × 528 columns

Figure 5.1 - Outline of the endometrial cancer dataset

Figure 5.1 shows an overview of the data set (first 5 genes with few samples) is shown below. There are 527 samples and 20531 genes.

5.2. Applying Principal Component Analysis (PCA) on Endometrial Cancer Dataset

The missing values of the dataset will be removed before applying PCA [4] to the dataset. The dataset will be reduced to 17507 rows. Google colabs [21] has been used as the coding platform.

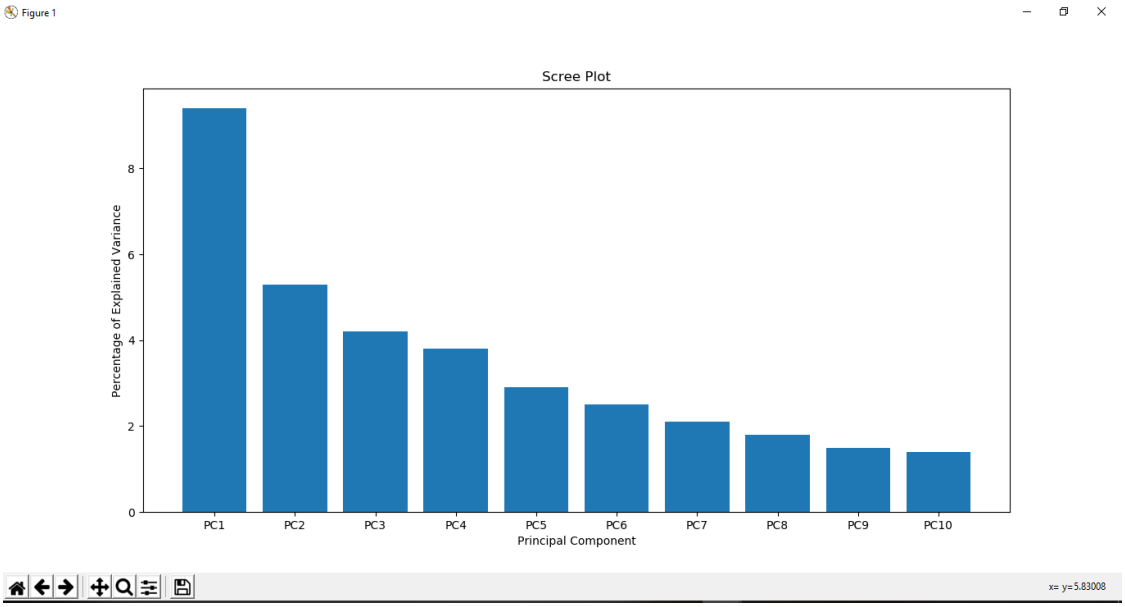


Figure 5.2 - 1st 10 principle components with their percentage of explained variance

Figure 5.2 shows the 1st 10 principle components with their percentage of explained variance .

- `Pca.explained_variance_ratio_ = [0.09423468 0.05282622 0.04238226 0.03757258 0.02940543 0.02516511 0.02146555 0.01809024 0.01540389 0.01407771 0.0125361 0.01172478 0.01069236 0.00999382 0.00931094 0.00861209 0.00759416 0.007299640.00702044 0.00673783]`

- 14.6% of the variance contained in PCA1 & PCA2 components.
- Using PCA 1 and PCA 2 clustering of samples can be analyzed as below.

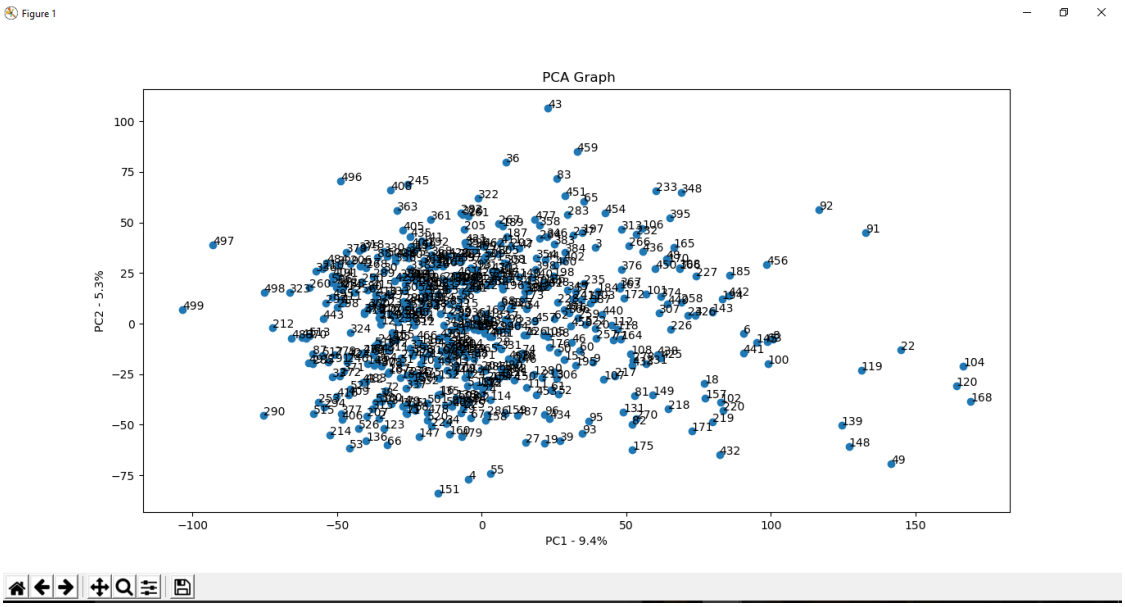


Figure 5.3 - PCA plot for the EC dataset with PCA 1 and PCA2

- Separated clusters cannot be identified as shown in Figure 5.3.

```
[28] loading_scores = pd.Series(pca.components_[0], index=data1.Entrez_Gene_Id)
     ## now sort the loading scores based on their magnitude
     sorted_loading_scores = loading_scores.abs().sort_values(ascending=False)

     # get the names of the top 10 genes
     top_10_genes = sorted_loading_scores[0:10].index.values

     ## print the gene names and their scores (and +/- sign)
     print(loading_scores[top_10_genes])
```

```
Entrez_Gene_Id
151903    0.020673
9774     -0.020374
8636      0.019891
2647      0.019822
126003    0.019807
79176     0.019768
324      -0.019729
1654     -0.019729
27125    -0.019710
2197      0.019665
dtype: float64
```

Figure 5.4 - mostly varied genes with PCA 1

When considering PCA1, mostly varied genes can be sorted out as in Figure 5.4

5.3. Applying T-Distributed Stochastic Neighboring Entities (t-SNE) [5] method on endometrial cancer dataset

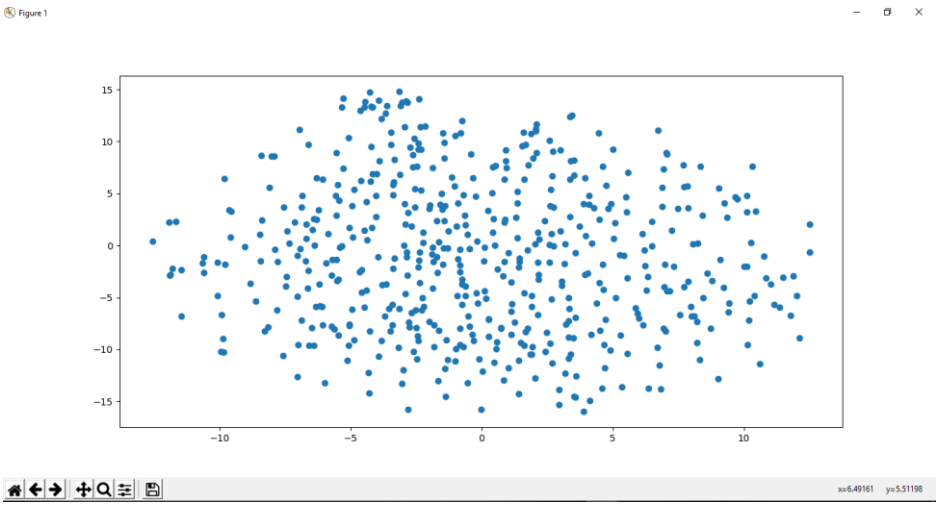


Figure 5.5 - Application of t-SNE with EC dataset

As in Figure 5.5, clearly separated clusters couldn't be found.

5.4. Applying PCA followed by T-SNE

The first 250 principal components are used here with the Cumulative explained variation 0.8849880258255189.

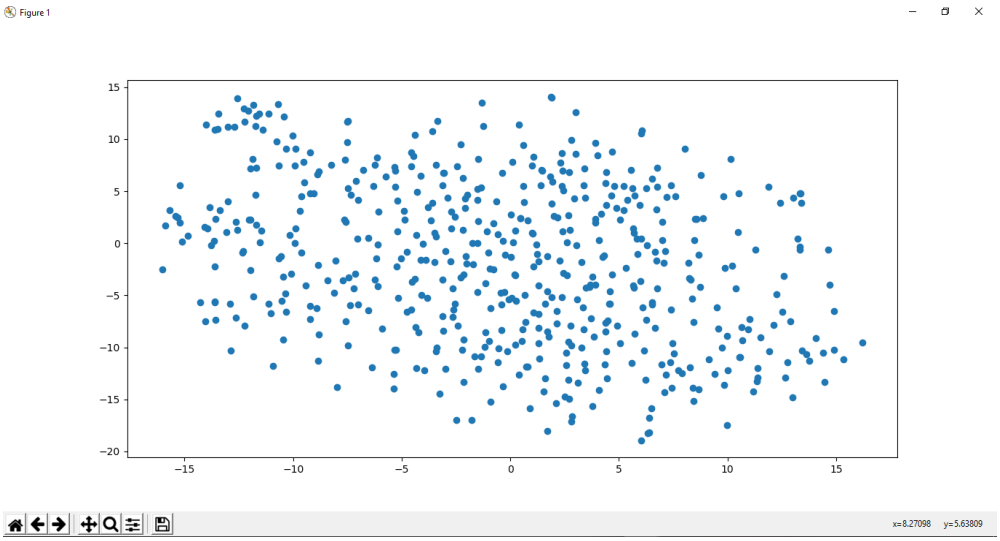


Figure 5.6 - Application of PCA followed by t-SNE on E.C dataset

As in Figure 5.6, no separate clusters could be identified using this method too.

5.5. Comparison of the above 3 methods with MNIST dataset

- In here the data frame size is (70000, 786)
- MNIST dataset[6] contains thousands of images with handwritten characters
- In the first picture, PCA has applied for n =3 situation. Then the Explained variation per principal component can be seen for PCA1 ,PCA 2 & PCA 3 as [0.09746116 , 0.07155445 , 0.06149531]
- After applying PCA and plot PCA1 against PCA2 graph 1 is obtained. It's hard to separate clusters .0-9 different numbers are shown in different colors.
- The first 10000 data sets will be considered due to the run time and the machine's performance.
- Then t-SNE will be applied for n_components = 2 (2nd graph). So 2 reduced dimensions will be given by the algorithm. Those 2 dimensions can be visualized as in graph 2. The samples have been colored according to their labels.
- It's recommended that the t-SNE is applied after applying another dimensionality reduction method. The 3rd graph in Figure 2.9 shows how MNIST dataset behaves when t-SNE applied after application of PCA with top 50 principal components.

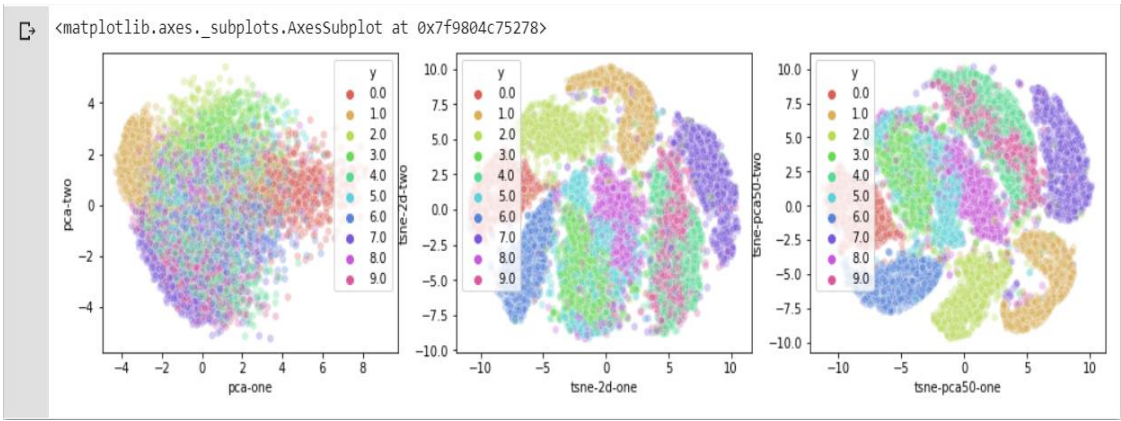


Figure 5.7 -A comparison of the application of PCA and T-SNE to mnist dataset.

It clearly shows the better accuracy of the T-SNE method.

More separation of the clusters can be seen when moving to the right side of the above graphs.

5.6. Comparison of the above 3 methods with endometrial cancer dataset.

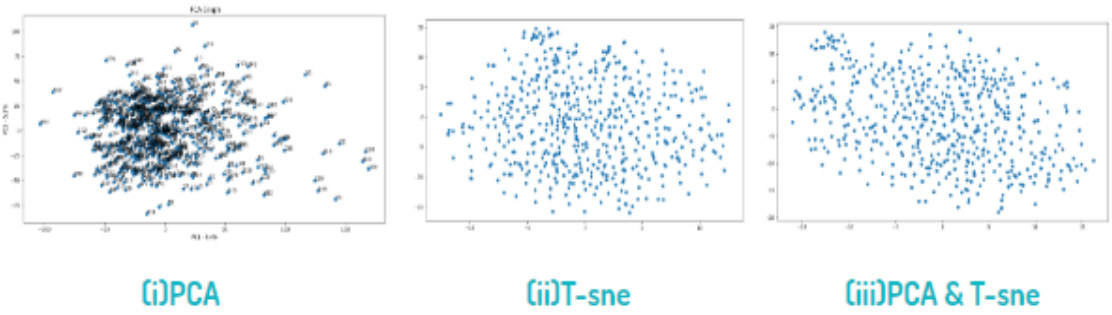


Figure 5.8 -A comparison of the application of PCA and T-SNE to mnist dataset

Even Though the above mentioned methods create good clusters with MNIST data set, no clustering could be seen with our biological dataset as shown in Figure 5.8. That proves the normal approaches can't be used when dealing with biological data most of the time.

5.7. Applying K-means clustering on endometrial cancer dataset.

The elbow graph is generated to get the number of clusters.

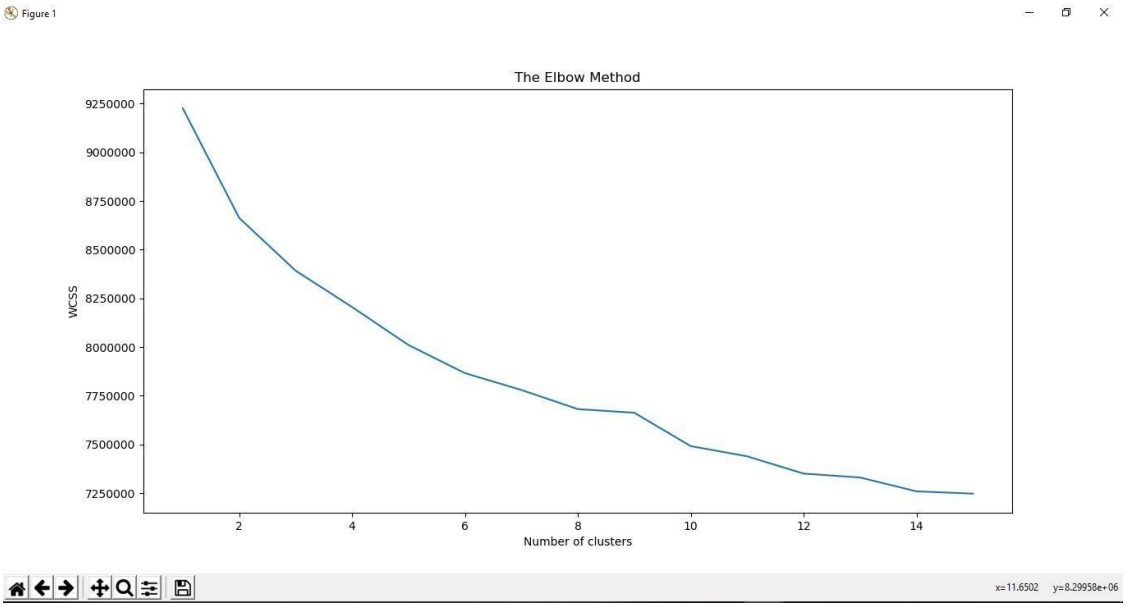


Figure 5.9 -Elbow plot for the E.C dataset

A good elbow graph cannot be seen with the elbow method as shown in Figure 5.9. Approximately 2 or 10 can be taken as the number of clusters.

i) 10 as the cluster size

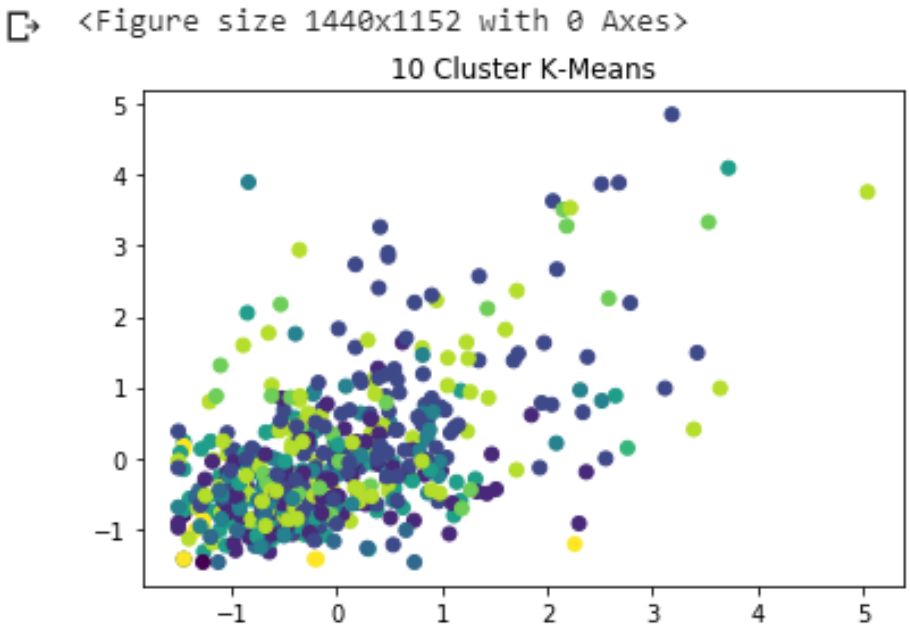


Figure 5.10 -K -means with K= 10

ii) 2 as the cluster size

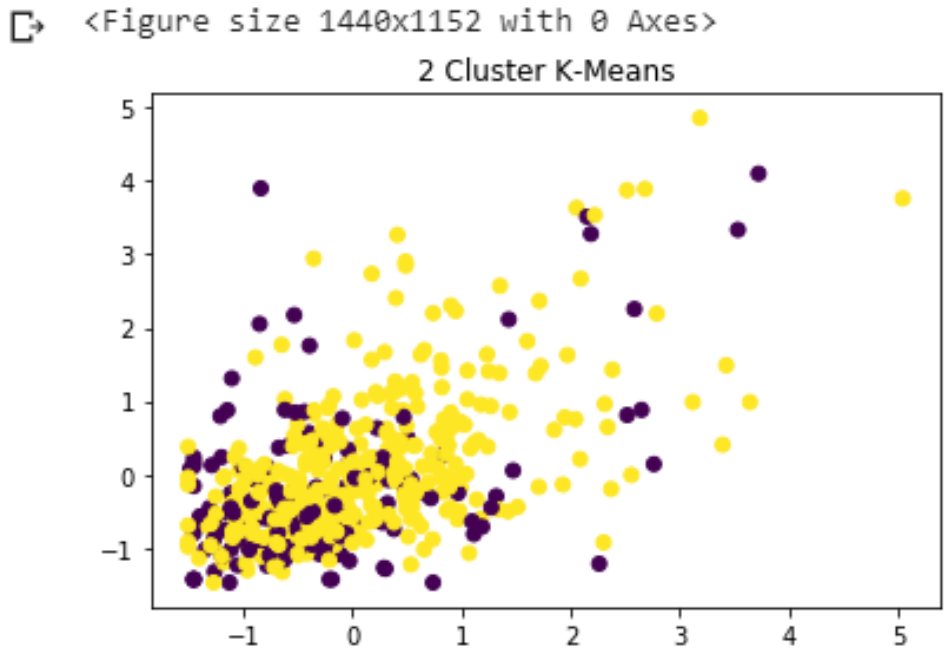


Figure 5.11 - K -means with K= 2

With the results shown in Figure 5.10 and Figure 5.11, no clustering could be obtained with the k-means clustering. In the next subsection highly varied genes will be considered.

5.8. Applying K-means clustering on EC dataset – highly varied genes (80%).

Some studies [11] have shown that for datasets that have more than 6000 features, extracting the highly varied genes based on the coefficient of covariance is a good way to reduce the high dimensionality. The method is applied to the endometrial dataset and measures the cluster tendency. The 14000 genes with highest variance (80% of the previously considered dataset after removing the missing values) from the EC dataset.

The elbow plot is obtained to see the number of clusters as in figure 5.12.

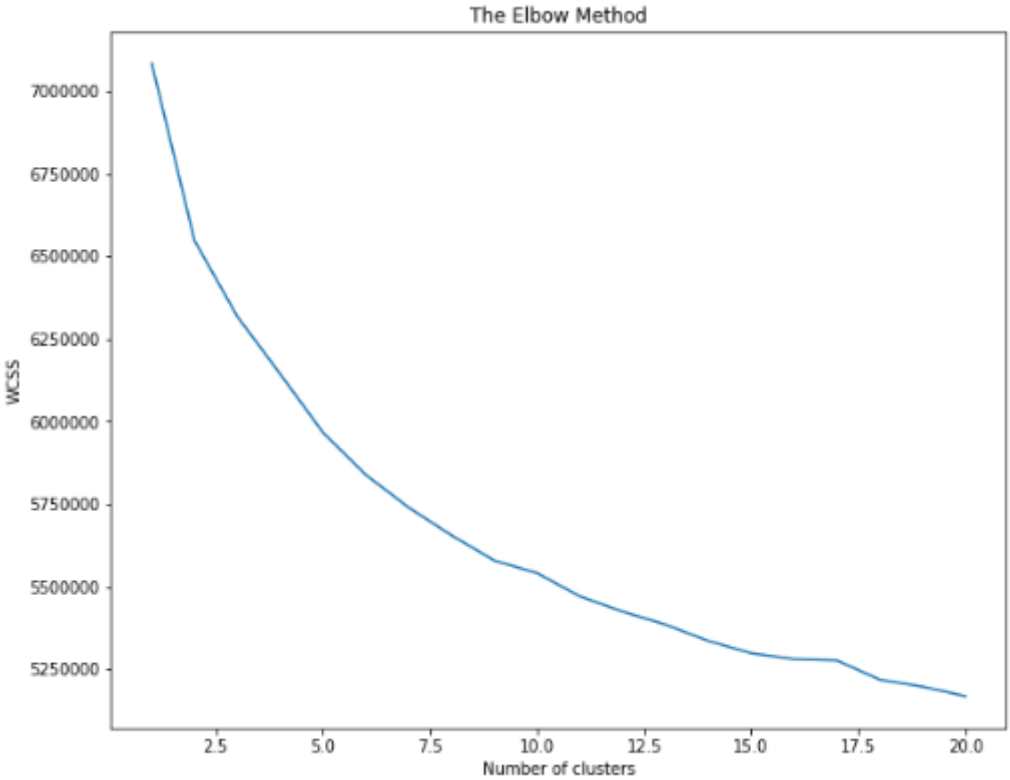


Figure 5.12 -Elbow plot for the E.C dataset with highly varied 14000 genes

According to elbow plot in Figure 5.12, elbow shapes are visible near 2 and 17.

When k = 2

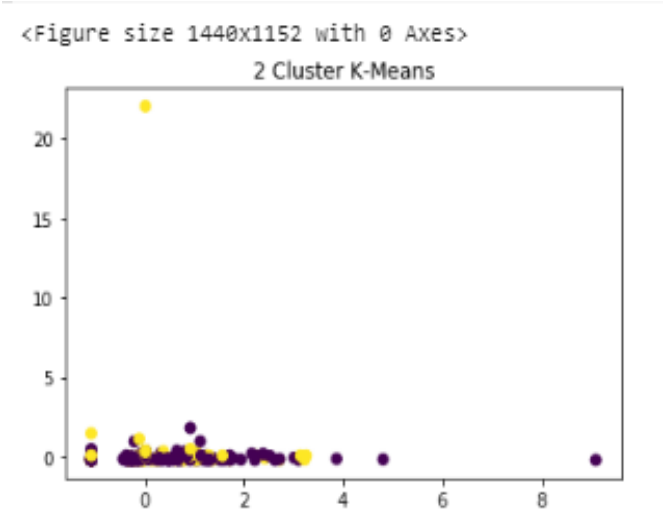


Figure 5.13 - K-means with K= 2 – highly varied 14000 genes

When K = 17

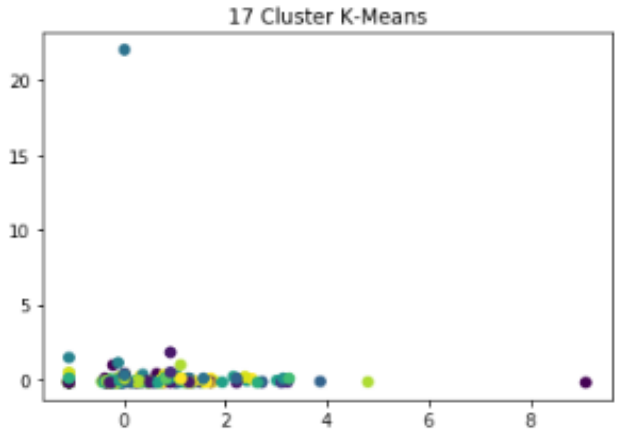


Figure 5.14 - K-means with K= 17 (highly varied 14000 genes)

.As shown in figure 5.13 and 5.14 proper clustering could not be obtained with the highly varied 14000 genes. In the next section most varied 60 % of the genes will be considered.

5.9 Applying K-means clustering on EC dataset – highly varied genes (60%).

Here the 10200 most highly varied genes (60 %) are considered for the clustering. The elbow plot is shown in the figure 5.15

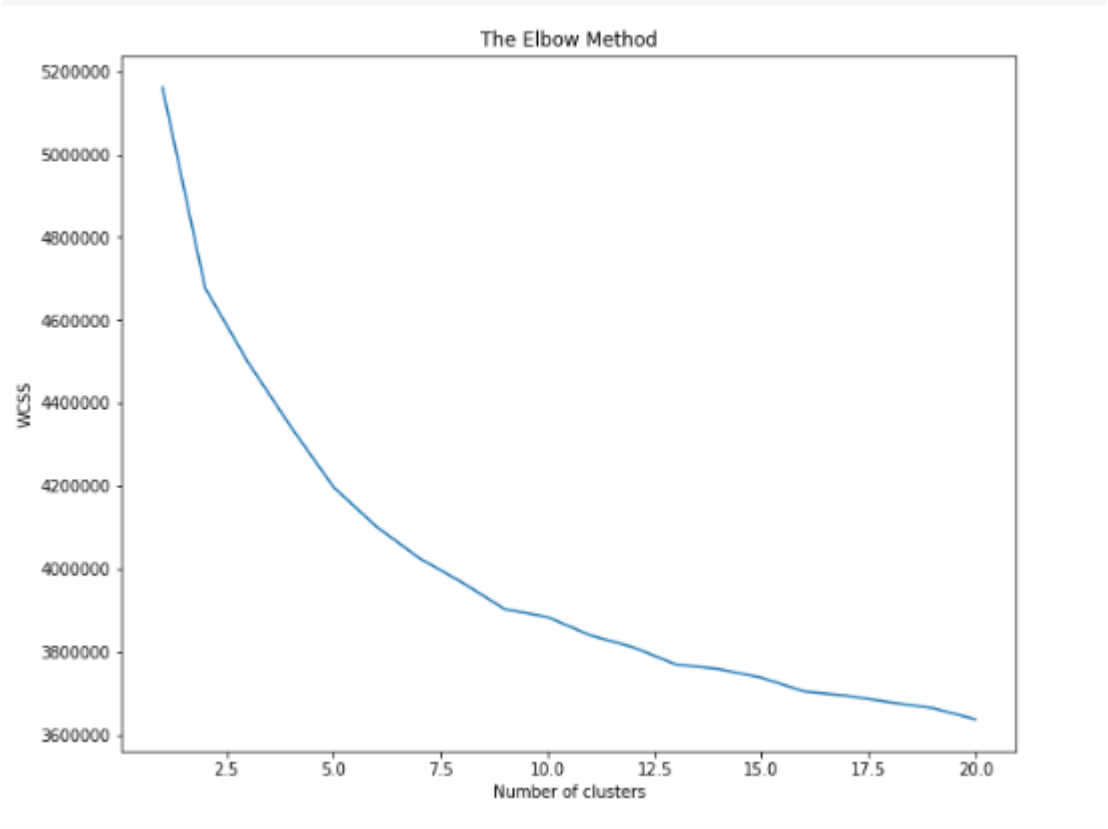


Figure 5.15 – Elbow plot for the highly varied 10200 genes (60%)

According to the elbow plot in figure 5.15 elbow shapes are visible near 2, 5, 9 and 13.

When k = 2

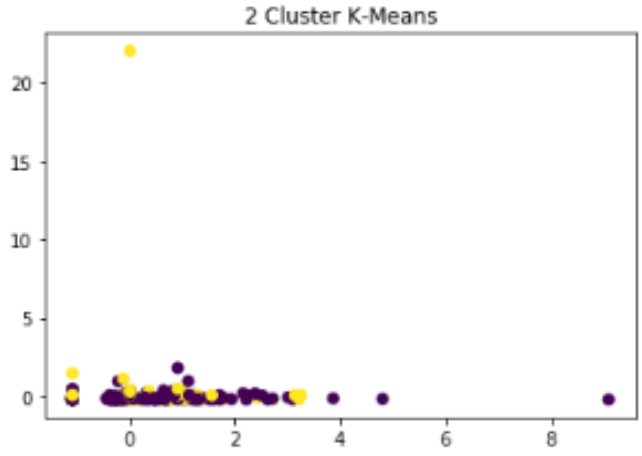


Figure 5.16 - K-means with K= 2 (highly varied 10200 genes)

As shown in figure 5.16, well separated clusters cannot be seen when the k value is 2.

When k =5

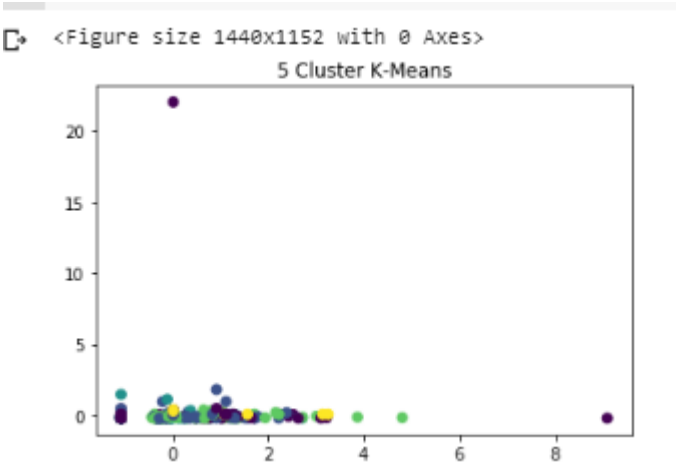
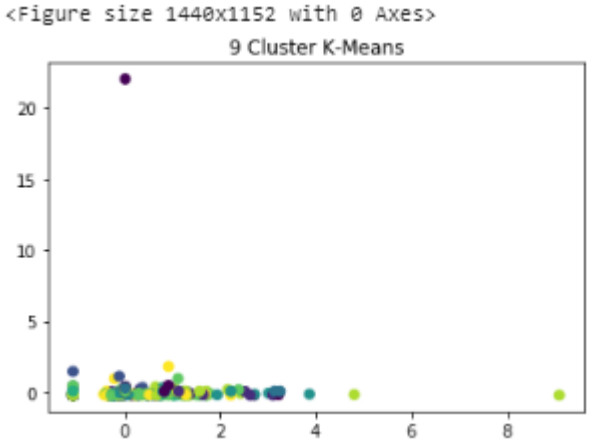


Figure 5.17 - K-means with K= 5 (highly varied 10200 genes)

As shown in figure 5.17, well separated clusters cannot be seen when the k value is 5.

When k = 9

Figure 5.18 - K-means with K= 9 (highly varied 10200 genes)



As shown in figure 5.18, well separated clusters cannot be seen when the k value is 9.

When k = 13

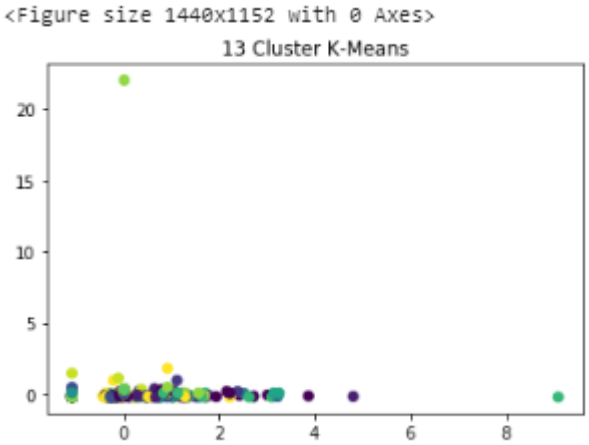


Figure 5.19 - K-means with K= 13 (highly varied 10200 genes)

As shown in figure 5.19, well separated clusters cannot be seen when the k value is 13. When considering the results obtained for most 14000 (80%) and 10200 (60%) highly varied genes fails to give well separated clusters. Since reducing the percentage further would lead to eliminate the biologically important features, this method cannot be considered as a goo method for reduce the dimensionality and the subtype identification task for endometrial cancer.

5.10. Applying Deep learning based method

Since any clusters could not be obtained with the highly varied genes as described in section 5.8 and 5.9, there is some impact than the normal data. The deep learning approach was tested with both normal and highly varied dataset. The highly varied genes with deep learning approach has shown better performance than the normal dataset (Based on the results in table 5.2 in section 5.11.1). So it's recommended for using highly varied genes with the deep learning approaches.

In this approach, the Silhouette Score was used for selecting the optimal number of clusters. The average silhouette distance was measured for different K values. The highest average silhouette distance was obtained for K= 6 as shown in table 5.1.

K value	**Silhouette Distance**
K = 4	0.292849303342213
K = 5	0.281232543312390
K = 6	**0.301900951748101**
K = 7	0.233913979140884
K = 8	0.248731460163114
K = 9	0.205346685569161
K = 10	0.161523254041652
K = 11	0.147436512098472

Table 5.1 – Silhouette Distances for different K values

When considering K value as 6 well separated clusters could be obtained from the EC dataset as in figure 5.20.

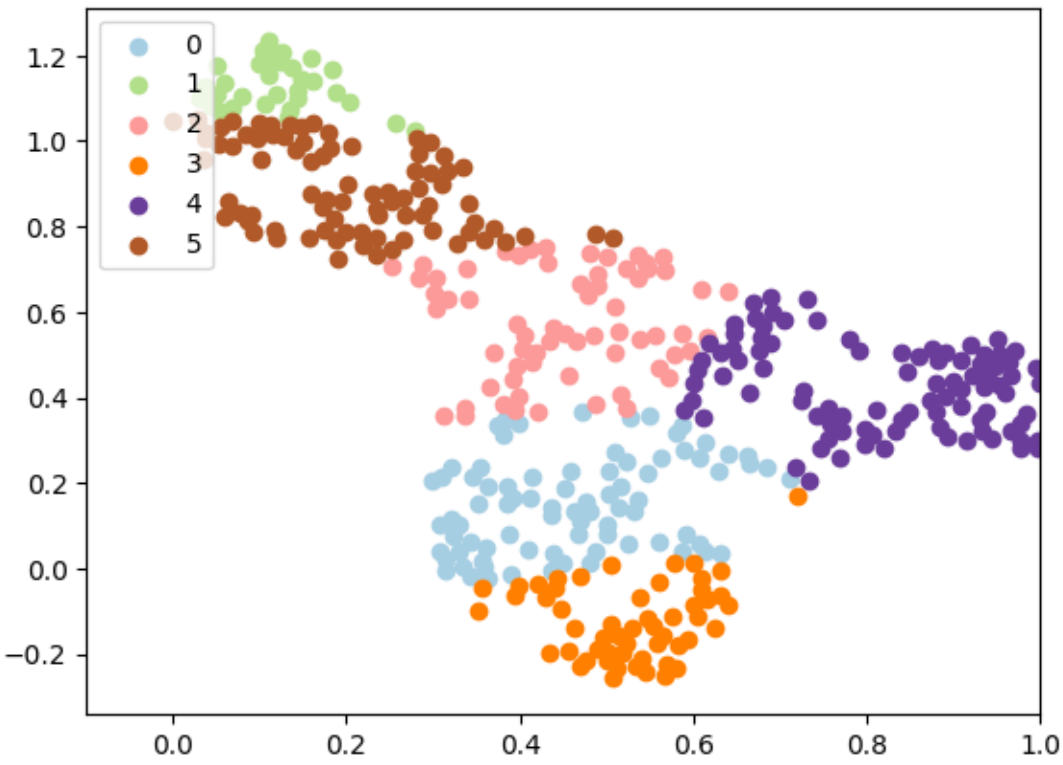


Figure 5.20 – Obtained clusters after applying the deep learning approach for highly varied genes.

5.11 Validation

5.11.1. Internal Validation

In this section the internal measures which are described in chapter 3 are measured and compared with the different dimensionality reduction methods. The Dunn index and NMI scores are calculated as the internal validation measures. The validations are done to different dimensionality reduction techniques alone and to different combinations among those methods when they are applied to the endometrial cancer dataset with K-means algorithm. Those methods are

- Validating the results after applying PCA
- Validating the results after applying t-SNE

- Validating the results after applying both PCA and t-SNE
- Validating the results after applying deep learning based approach
- Validating the results after applying deep learning based approach for the highly varied 14000 genes.

Method	PCA	t-SNE	PCA and t-SNE	Deep learning approach	Deep Learning Approach with highly varied 80% genes
DUNN Index	0.108224	0.020123	0.024976	**0.235940**	**0.240582**
Silhouette Score	0.015088	0.230294	0.240529	**0.289361**	**0.301900**

Table 5.2 – Internal Validation Scores

The above mentioned scores are calculated, when k = 6 which has been identified as the optimal number of clusters as mentioned in section 5.10.

When considering the results in table 5.2, the scores obtained by deep learning based approach is higher when compared with the other existing methods. When the Deep learning method applied to highly varied genes, it shows a greater value rather than applying to the entire (normal) dataset. This result says deep learning approach with highly varied dataset can be considered as the best method out of the all the methods in the table.

5.11.2. External Validations

External Validations are done with the use of pre-identified 4 classes for endometrial cancer from the clinical studies using the k value as 4.

Method	PCA	t-SNE	PCA and t-SNE	Deep learning approach	Deep Learning Approach with highly varied 80% genes
NMI Score	0.058987	0.024133	0.007742	0.005697	0.014364
Jaccard Score	0.144796	0.179487	0.207637	0.076923	0.076923

Table 5.3 – External Validation Scores

As in table 5.2, in table 5.3 the scores which were obtained by the deep learning approaches not always get the highest values. To do the external validation with existing subclasses we need consider cluster size as 4. But the original number of clusters we got from the silhouetted distance was 6. That might be a reason for the ambiguity of the result. But with the internal validations it is possible to say the deep learning approach with highly varied data is the best among the above considered methods for endometrial cancer dataset.

Chapter 6 - Conclusion

6.1. Introduction

Cancer is a disease which kills thousands of people yearly. Cancers cannot be treated through basic principles of medicine. It is a disease in genetic level. Due to this reason, two patients might get different results with the same treatment. One might get cured and the other one might die. So there is a requirement for identify the cancer subtypes to design personalized medicine.

Different methods have been tried in the past to identify cancer subtypes. Due to the lack of data and lack of computational power, the subtype identification task was hard and the identified subtypes were ambiguous. In present, gene sequencing techniques have resulted in large volumes of genomic data from multiple studies which leads to understand the causes better. In present, enough computation power is available to do an effective analysis.

When identifying cancer subtypes, there is a problem due to the high dimensionality of data. The high dimensionality means the number of genes are comparably high with the number of patients or the samples. So this high dimensionality should be reduced to get good results from the computational models which are used to subtype identification task. This study focuses on expression data of the endometrial cancer which consists of 20000+ genes and 500+ patients taken from multiple studies. This chapter includes how that problem was addressed, limitations and future works.

6.2. Conclusions about research questions

Biological data shows different behaviors to different algorithms. So the first research question is to check the behavior of the EC dataset with existing dimensionality reduction methods. Dimensionality reduction techniques are divided in to 2 main parts as feature extraction methods and feature selection methods. As discussed in chapter 2, the best methods identified through literature was considered in this study to analyze the behavior of the EC dataset.

From the feature selection methods the best method for more than 6000 genes was considered as the Variance based method [11].So the most highly varied 14000 genes (80% of the total number genes) and the most 10200 (60% of the total number genes) were considered and clustering was done with k-means to check whether there are separated clusters. As shown in section 5.8 and 5.9 there were no well separated cluster could be seen.

Feature extraction methods are more effective than the feature selection methods. The reason is feature selection methods consume more time due to the comparison on entire search space .The most used methods that were found for genomic data was PCA [4] , tSNE [5] , combination of both PCA and t-SNE and deep learning based models. The PCA was applied as in section 5.2 and there were no well separated clusters could be seen. No well separated clusters were given for t-SNE as well. There was an approach with the combination of both PCA and t-SNE which have given well separated clusters for the MNIST dataset. That method also applied to the EC dataset and no well separated cluster could be found. There is a comparison for that method with both MNIST and EC datasets in section 5.5 and 5.6.

When considering deep learning approaches, there was an approach which consists of deep autoencoders method which was designed to identify the subtypes of Glioblastoma [25]. Since there were no non-cancer organ-specific tissue controls for the endometrial that method could not be applied to the EC dataset. Then a deep learning approach was considered which was applied to METABRIC breast cancer dataset [35] .It consists of a semi supervised approach with a prior biological knowledge (previously identified subtypes) integration. The supervised model uses those previously identified classes to train its weights and it projects the row genomic data into a low dimensional representation space. Since there are previously identified 4 types for endometrial cancer

with clinical studies, this method could be applied to the endometrial cancer dataset as a start for answering the Second research question. It is assumed that the lack of data and lack of computational power, previously identified clusters can be ambiguous and there can be new subtypes due to the genetic alterations with the time. But those classes can be used to train a supervised model and the final layer of the representation layer (Which forms a low dimensional space for the input genes) of the model can be used to clustering and identify novel subtypes. When the method applied to the endometrial cancer, well separated **6** clusters could be obtained from the deep learning approach as shown in section 5.10. When the highly varied 14000 genes were used with this deep learning approach the dimensionality could be reduced further as discussed in section 5.11.

As a solution to the third research question, a set of internal and external validations were considered. The DUNN Index [32] and Silhouette Score were considered as internal validation methods while Jaccard Coefficient/Score and NMI Score were considered as external validations. When comparing internal validation scores as shown in table 5.2, deep learning based 2 approaches have shown higher values indicating that those 2 methods are better than the existing dimensionality reduction methods. When comparing those two methods, when the deep learning method applied to the highly varied genes shows better values rather than applying it to the entire dataset.

When considering the external validations, the values has shown some ambiguous pattern as shown in table 5.3 and deep learning based methods have some low values as well. We do this external validations by considering k = 4 due to the availability of clinically identified 4 subtypes for EC. But with the silhouette method, it is identified that the optimal number of clusters was **6**. This can be assumed as the reason of the ambiguity of the pattern. But with the internal validations which consider the low intra cluster similarity (closeness of the members of the same cluster) and high inter and cluster similarity (away from the data points of the different clusters) it is possible to say that the deep learning approach with highly varied data can be considered as the best method among all the methods considered for the endometrial cancer dataset. The highly varied genes and the deep learning approach together forms a better method for dimensionality reduction than the other state of the art methods.

6.3. Limitations

Since we do not have non-cancer organ-specific tissue controls endometrial cancer dataset, some methods could not be applied.

The validations could be done using only statistical methods and computational models. The results cannot be validate clinically due to the unavailability of medical laboratories.

The datasets can be not from a certain group of people. There can be different patients from different ethnicity groups. So there may be some outliers which will affect the result.

6.4. Implications for further research

This research was done only to the expression data. This can be extended to copy number variation data and mutation data as well. This method can be applied to several datasets and can be developed as a generalized method for all subtypes.

References

[1] “What Is Cancer?,” National Cancer Institute. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,” Nature, vol. 403, no. 6769, pp. 503–511, 2000.

[3] N. Toomula and H. Bindu, “Pharmacogenomics- Personalized Treatment of Cancer, Diabetesand Cardiovascular Diseases,” Journal of Pharmacogenomics & Pharmacoproteomics, vol. 03, no. 01, 2012.

[4] T. M. Josserand, “Classification of gene expression data using PCA-based fault detection and identification,” 2008 IEEE International Workshop on Genomic Signal Processing and Statistics, 2008.

[5] Maaten, Laurens Van Der , Hinton and Geoffrey, “Visualizing Data using t-SNE”,Journal of Machine Learning Research , 2018 .

[6] “THE MNIST DATABASE,” MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[7] “cBioPortal for Cancer Genomics,” cBioPortal for Cancer Genomics. [Online]. Available: https://www.cbioportal.org/ .

[8] Mahajan, S Abhishek and Shailendra Singh, “Review On Feature Selection Approaches Using Gene Expression Data”,Imperial Journal of Interdisciplinary Research,vol. 02, issue. 03, 2016.

[9] Chin, Ang Jun Mirzal, Andri Haron, Habibollah Member, Senior Nuzly, Haza Hamed and Abdull ,”Supervised , Unsupervised and Semi - supervised Feature Selection : A Review on Gene Selection” ,IEEE/ACM Transactions on Computational Biology and Bioinformatics ,2015.

[10] Sreepada, Rama Syamala, Swati Vipsita and Puspanjali Mohapatra. "An efficient approach for microarray data classification using filter wrapper hybrid approach." Advance Computing Conference (IACC), 2015 IEEE International. IEEE, 2015.
[11] S. Solorio-Fernandez, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, “Ranking Based Unsupervised Feature Selection Methods: An Empirical Comparative Study in High Dimensional Datasets,” Advances in Soft Computing Lecture Notes in Computer Science, pp. 205–218, 2018.
[12] G. Li, X. Hu, X. Shen, X. Chen, and Z. Li, “A novel unsupervised feature selection method for bioinformatics data sets through feature clustering,” 2008 IEEE International Conference on Granular Computing, 2008.
[13] S. Solorio-Fernandez, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, “A new Unsupervised Spectral Feature Selection Method for mixed data: A filter approach,” Pattern Recognition, vol. 72, pp. 314–326, 2017.
[14] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang & Xiaofang Zhou,”2,1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning” ,IJCAI International Joint Conference on Artificial Intelligence,, pp. 1589 - 1594, 2011.
[15] Mitra, C. A. Murthy, and Sankar K. Pal,“Unsupervised Feature Selection Using Feature Similarity”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 301-312, 2002.
[16] S. Kar, K. D. Sharma, and M. Maitra, “Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique,” Expert Systems with Applications, vol. 42, no. 1, pp. 612–627, 2015.
[17] P. S. Deepthi and S. M. Thampi, “PSO based feature selection for clustering gene expression data,” 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2015.
[18] C. V. Meegen, S. Schnackenberg, and U. Ligges, “Unequal Priors in Linear Discriminant Analysis,” Journal of Classification, 2019.
[19] W. Li, J. E. Cerise, Y. Yang, and H. Han, “Application of t-SNE to human genetic data,” Journal of Bioinformatics and Computational Biology, vol. 15, no. 04, p. 1750017, 2017.
[20] A.Lengyel and Z. Botta-Dukát, “Silhouette width using generalized mean – a flexible method for assessing clustering efficiency,” 2018.

[21] S. Zhang and J. Yu, “A New Connectivity-Based Cluster Validity Index,” 2010 Chinese Conference on Pattern Recognition (CCPR), 2010.

[22] M. Misuraca, M. Spano, and S. Balbi, “BMS: An improved Dunn index for Document Clustering validation,” Communications in Statistics - Theory and Methods, pp. 1–14, 2018.

[23] Google. [Online]. Available: https://colab.research.google.com/

[24] “learn,” scikit. [Online]. Available: https://scikit-learn.org/stable/

[25] J. D. Young, C. Cai, and X. Lu, “Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma,” BMC Bioinformatics, vol. 18, no. S11, 2017.

[26] L. Fayed, “Differences Between a Malignant and Benign Tumor,” Verywell Health, 27-Jan-2020.[Online].Available:https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240.

[27]L. Fayed, “Differences Between a Malignant and Benign Tumor,” Verywell Health, 27-Jan-2020.[Online].Available:https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240.

[28] A. Kamal, N. Tempest, C. Parkes, R. Alnafakh, S. Makrydima, M. Adishesh, and D. K. Hapangama, “Hormones and endometrial carcinogenesis,” Hormone Molecular Biology and Clinical Investigation, vol. 25, no. 2, 2016.

[29] M. K. Mcconechy, J. Ding, M. C. Cheang, K. C. Wiegand, J. Senz, A. A. Tone, W. Yang, L. M. Prentice, K. Tse, T. Zeng, H. Mcdonald, A. P. Schmidt, D. G. Mutch, J. N. Mcalpine, M. Hirst, S. P. Shah, C.-H. Lee, P. J. Goodfellow, C. B. Gilks, and D. G. Huntsman, “Use of mutation profiles to refine the classification of endometrial carcinomas,” The Journal of Pathology, 2012.

[30] D. A. Levine, “Integrated genomic characterization of endometrial carcinoma,” Nature, vol. 497, no. 7447, pp. 67–73, 2013.

[31] L. Li, “K-Means Clustering with scikit-learn,” *Medium*, 04-Jun-2019. [Online]. Available:https://towardsdatascience.com/k-means-clustering-with-scikit-learn-6b47a369a83c.

[32] J. Bezdek and N. Pal, “Cluster validation with generalized Dunn's indices,” Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems.

[33]“sklearn.metrics.silhouette_score,”*scikit*.[Online].Available: https://scikitlearn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[34]“Jaccard index (R),” *ClustEval*.

[Online]. Available: https://clusteval.sdu.dk/1/clustering_quality_measures/7.

[35] R. Chen, L. Yang, S. Goodison, and Y. Sun, “Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data,” Bioinformatics, 2019.

[36] M. C. P. de Souto, I. G. Costa, D. S. A. de Araujo, T. B. Ludermir, and A. Schliep, “Clustering cancer gene expression data: A comparative study,” *BMC Bioinformatics*, vol. 9, pp. 1–14, 2008