

# **Gene Selection and Cancer Classification Using a Multidimensional Fuzzy Deep Learning Approach for Gene Expression Data**



**Mahmood Khalsan**

Faculty of Arts, Science and Technology  
University of Northampton

This dissertation is submitted for the degree of  
*Doctor of Philosophy*

August 2023

I dedicate this work to all people who have shown their contributions during my PhD journey, family, friends, and supervisors. To my incredible parents for their endless support and blessings. To the one person who has offered her great efforts in everything and pushed me to do my best, my beloved wife. In addition, I would like to dedicate this work to the little angels who brought more joy to our life, my lovely son and daughter.

## **Declaration**

I hereby declare that the work contained in this thesis is entirely original to me, that it was completed after I registered for the Ph.D. degree at The University of Northampton, and that it has never been previously included in a thesis submitted to this or any other institution. I assume responsibility for carrying out the procedures following the University's most recent research ethical standards after reading them. I also have made an effort to list any potential dangers associated with this research that might develop during its execution. I have also acknowledged my duties and the rights of the participants, as well as received the necessary ethical and/or safety permission (where applicable).

Mahmood Khalsan

August 2023

## **Acknowledgements**

I want to express my gratitude to the members of my committee who gave their knowledge and priceless time. A special thanks to Prof. Dr. Mu Mu, Prof. Dr. Michael Opoku Agyeman, Prof. Dr Eman Salih Al-Shamery, and Dr. Suraj Ajit for their valuable direction, oversight, and tireless efforts throughout the entire PhD journey. Many thanks and appreciation to Prof. Dr. Lee Machado and Dr. Karen Anthony for their volunteering work of countless hours of reflecting, reading, encouraging, and contributing to this knowledge. Moreover, I am most grateful for Prof. Dr. Scott for his insightful counsel and patience during the initial stage of PhD adventure.

## Abstract

Deep Learning approaches are powerful techniques commonly employed for developing cancer prediction models using associated gene expression and mutation data. This thesis provides a comprehensive review of recent cancer studies that have employed gene expression data from several cancer types (i.e Breast, Lung, Kidney, Liver, Gallbladder, Gastric, and Thyroid) for survival prediction, tumour identification, and stratification as well as providing an overview of biomarker studies that are associated with these cancer types. The thesis captures multiple aspects of machine learning-associated cancer studies, including cancer classification, cancer prediction, identification of biomarker genes, microarray, and RNA-Seq data. The thesis discussed the technical issues with current cancer classification models and the corresponding measurement tools for determining the activity levels of gene expression between cancerous tissues and noncancerous tissues. The work has not only highlighted the issues but also attempted to address the issues that arise in the previous studies. One of the notable issues that employ gene expression data for cancer classification is the high dimensionality of the available datasets.

As a result, the work developed a fusion three feature selection initial approach to reduce the number of genes and increase the accuracy of cancer classification by using the concept of intersection. fusion three feature selection methods was developed to select an optimal subset of genes that would be used as identifiers for classification and reduce the dimensionality of the available data in gene expression. The results ratings for employed cancer datasets ranged from (95.5% to 98%) accuracy, (94.4% to 100%) precision, (94% to 100%) recall, and (95.7% to 98%) f1-score.

Therefore a new fuzzy gene selection approach was developed to identify significant genes to facilitate cancer classification and reduce the dimensionality of the available gene expression data. This method demonstrated that has the ability for classifying cancer accurately in evaluating most of the datasets that were employed. The results indicate that FGS enhanced the performance of the five common classifiers with the majority of cancer expression datasets employed and particularly when the FGS and MLP were applied together. The average results were 97%,97.3%, 96.5%, and 96.77% for accuracy, precision, recall, and f1-score respectively.

Next, Fuzzy gene selection-wrapper plus is offered as the second significant addition to this thesis. Fuzzy gene selection-wrapper plus attempts to reduce the number of genes selected by Fuzzy gene selection while maintaining the accuracy attained. The results showed that Fuzzy gene selection-wrapper plus was able to reduce the number of genes selected by Fuzzy gene selection method up to 82% without sacrificing accuracy and other evaluation metrics.

Finally a novel fuzzy classifier method developed to enhance the accuracy and increased the generalisation of the proposed algorithm in most of employing datasets. Fuzzy classifier demonstrated that it continuously achieved the highest results in all employed datasets compared to classical classifiers. The average results were 98%, 98.2%, 96.5%, and 97% for accuracy, precision, recall, and f1-score respectively for all employed datasets.

The thesis integrated the developed approaches (Fuzzy gene selection-wrapper plus, and Fuzzy classifier) into a single, fully automated end-to-end model called multidimensional fuzzy deep learning.

# Table of contents

<b>List of figures</b>	<b>xi</b>
<b>List of tables</b>	<b>xiii</b>
<b>Abbreviations</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Research Problem . . . . .	4
1.3 Aim and Objectives . . . . .	5
1.3.1 Thesis Aim . . . . .	5
1.3.2 Thesis Objectives . . . . .	5
1.4 Contributions of the thesis. . . . .	6
1.5 List of Publications . . . . .	7
1.6 Thesis Structured . . . . .	7
<b>2 Background and Literature Review</b>	<b>9</b>
2.1 Microarray and RNA-seq . . . . .	9
2.2 Feature selection . . . . .	10
2.2.1 Mutual information . . . . .	11
2.2.2 F-Classif . . . . .	11
2.2.3 Chi-squared . . . . .	12
2.2.4 Minimum redundancy maximum relevance . . . . .	12
2.3 Classifier approaches . . . . .	12
2.3.1 Support vector machine . . . . .	13
2.3.2 Decision Tree . . . . .	13
2.3.3 Random Forest . . . . .	15
2.3.4 Naïve Bayes . . . . .	16
2.3.5 K-nearest Neighbors . . . . .	17

---

2.3.6	Multilayer Perceptron . . . . .	18
2.3.7	Logistic regression . . . . .	19
2.4	A Cross-validation . . . . .	22
2.5	Evaluation metrics . . . . .	22
2.5.1	Accuracy . . . . .	22
2.5.2	Precision . . . . .	23
2.5.3	Recall . . . . .	23
2.5.4	F1-score . . . . .	23
2.6	Data repositories . . . . .	24
2.6.1	Gene Expression Omnibus . . . . .	24
2.6.2	The Cancer Genome Atlas . . . . .	24
2.7	Related Work . . . . .	24
2.7.1	Classical Machine learning studies . . . . .	24
2.7.2	Deep learning studies . . . . .	28
2.8	Discussion . . . . .	33
2.9	Summary . . . . .	36
<b>3</b>	<b>Methodology</b>	<b>38</b>
3.1	Fusion three feature selection . . . . .	38
3.2	Multidimensional fuzzy deep learning model . . . . .	39
3.2.1	Pre-processing stage . . . . .	42
3.2.2	Fuzzy gene selection . . . . .	42
3.2.3	Fuzzy gene selection wrapper plus . . . . .	47
3.2.4	Fuzzy classifier method . . . . .	49
3.3	Mitigate Overfitting . . . . .	49
<b>4</b>	<b>Experimentation Framework</b>	<b>56</b>
4.1	Introduction . . . . .	56
4.2	Datasets employed for all developed approaches . . . . .	57
4.3	Comparing stage . . . . .	57
4.4	Evaluation stage . . . . .	59
4.4.1	A cross-validation . . . . .	59
4.4.2	Evaluation Performance . . . . .	60
<b>5</b>	<b>Discussing Experimental Results</b>	<b>63</b>
5.1	Experimentation of applying FTFS . . . . .	63
5.1.1	Datasets Employed to evaluate the FTFS Method. . . . .	63

---

5.1.2	The results achieved with FTFS . . . . .	63
5.1.3	Comparing FTFS to prior studies . . . . .	64
5.2	Experimentation of applying FGS . . . . .	67
5.2.1	The datasets used to examine FGS efficiency . . . . .	67
5.2.2	The results achieved with FGS . . . . .	68
5.2.3	Discuss FGS results . . . . .	69
5.3	Experimentation of applying FC . . . . .	84
5.3.1	The datasets used to examine FC efficiency. . . . .	85
5.3.2	The results achieved with FC. . . . .	85
5.3.3	Discuss FC results . . . . .	89
5.4	Experimentation of applying FGSWP . . . . .	94
5.4.1	The datasets used to examine FGSWP efficiency . . . . .	94
5.4.2	FGSWP's results with Classical Classifiers . . . . .	94
5.4.3	Results of FGSWP with FC . . . . .	101
5.4.4	Comparison of Findings . . . . .	102
5.4.5	Synthesis of Findings . . . . .	103
<b>6</b>	<b>General Discussion and Future Directions</b>	<b>106</b>
6.1	Summary of Findings . . . . .	108
6.1.1	Implications of FTFS Findings . . . . .	108
6.1.2	Implications of FGS Findings . . . . .	108
6.1.3	Implications of FC Findings . . . . .	109
6.1.4	Implications of FGSWP Findings . . . . .	109
6.1.5	Implications of MDFDL Findings . . . . .	109
6.2	Future Directions . . . . .	110
6.2.1	Integrating different inputs . . . . .	110
6.2.2	Advancements in cancer biomarkers . . . . .	110
6.2.3	Interpretable ML . . . . .	110
	<b>References</b>	<b>111</b>

# List of Algorithms

1	Data Pre-processing . . . . .	43
2	Vote step . . . . .	45
3	Fuzzy Gene Selection Process . . . . .	48

# List of figures

1.1	Deep Learning Structure . . . . .	3
1.2	Deep learning vs classical machine learning . . . . .	4
1.3	Deep learning performance vs human performance . . . . .	4
2.1	SVM and its Hyperplane Selection . . . . .	14
2.2	Decision Tree Process . . . . .	15
2.3	Random Forest Process . . . . .	16
2.4	KNN and its Hyperplane Selection . . . . .	18
2.5	Multilayer Perceptron (MLP) Structure . . . . .	20
2.6	K-Fold Cross Validation Process with K=5 . . . . .	22
2.7	Percentage of achieved accuracy in previous studies . . . . .	36
3.1	Block diagram illustrates the FTFS process . . . . .	41
3.2	The architecture of the developed model . . . . .	51
3.3	The developed Topology . . . . .	52
3.4	Block Diagram of Developed Fuzzy Gene Selection Process . . . . .	53
3.5	An overview of the developed FGSWP . . . . .	54
3.6	An overview of developed FC method. . . . .	55
4.1	Experimentation framework process of FTFS . . . . .	61
4.2	Experimentation framework process of MDFDL . . . . .	62
5.1	Comparing accuracy score when using and omitting FGS (GSE45827) . . . . .	69
5.2	Comparing accuracy score when using and omitting FGS (GSE19804). . . . .	74
5.3	Comparing accuracy score when using and omitting FGS (GSE14520) . . . . .	75
5.4	Comparing accuracy score when using and omitting FGS (GSE77314) . . . . .	75
5.5	Comparing accuracy score when using and omitting FGS (TCGA1) . . . . .	76
5.6	Comparing accuracy score when using and omitting FGS (GSE33630) . . . . .	77
5.7	Comparing accuracy score when using and omitting FGS (TCGA4) . . . . .	78

---

5.8	Comparing accuracy score when using and omitting FGS (TCGA2) . . . . .	78
5.9	Comparing accuracy score when using and omitting FGS (GSE53757) . . . . .	79
5.10	Comparing accuracy score when using and omitting FGS (TCGA7) . . . . .	80
5.11	Comparing accuracy score when using and omitting FGS (GSE43580) . . . . .	81
5.12	Comparing accuracy score when using and omitting FGS (TCGA6) . . . . .	81
5.13	Comparing accuracy score when using and omitting FGS (GSE10072) . . . . .	82
5.14	Comparing accuracy score when using and omitting FGS (GSE75037) . . . . .	83
5.15	Comparing accuracy score when using and omitting FGS (GSE66499) . . . . .	84
5.16	Comparing accuracy score when using and omitting FGS (GSE84437) . . . . .	84
5.17	FC vs classical classifiers (GSE33630) employing FGS . . . . .	89
5.18	FC vs classical classifiers (GSE45827) employing FGS . . . . .	89
5.19	FC vs classical classifiers (GSE84437) employing FGS . . . . .	90
5.20	FC vs classical classifiers (GSE66499) employing FGS . . . . .	90
5.21	FC vs classical classifiers (GSE19804) employing FGS . . . . .	91
5.22	FC vs classical classifiers (TCGA1) employing FGS . . . . .	91
5.23	FC vs classical classifiers (GSE43580) employing FGS . . . . .	92
5.24	FC vs classical classifiers (GSE14520) employing FGS . . . . .	92
5.25	FC vs classical classifiers (GSE53757) employing FGS . . . . .	93
5.26	FC vs classical classifiers (GSE10072) employing FGS . . . . .	93
5.27	FC vs classical classifiers (TCGA2) employing FGS . . . . .	93
5.28	FC vs classical classifiers (TCGA6) employing FGS . . . . .	93
5.29	A comparison of FGS vs FGSWP in five classifiers for (GSE43580) . . . . .	95
5.30	A comparison of FGS vs FGSWP in five classifiers for (GSE53757) . . . . .	97
5.31	A comparison of FGS vs FGSWP in five classifiers for (GSE33630) . . . . .	98
5.32	A comparison of FGS vs FGSWP in five classifiers for (GSE45827) . . . . .	99
5.33	A comparison of FGS vs FGSWP in five classifiers for (TCGA1) . . . . .	100
5.34	A comparison of FGS vs FGSWP in five classifiers for (GSE10072) . . . . .	101

# List of tables

2.1	Comparison of Classifiers for Cancer Classification using Gene Expression Data . . . . .	20
2.2	A summary of previous studies of applying classical ML to analyse cancer gene expression datasets . . . . .	29
2.3	A summary of previous studies applied DL to analyse cancer gene expression datasets . . . . .	34
3.1	Evaluate the effectiveness of using the FTFS with top 50,100,150 . . . . .	40
4.1	The full details of the datasets used to train and test the developed approaches.	58
5.1	Evaluate the effectiveness of using the FTFS method with six classifiers on six cancer datasets. . . . .	65
5.2	Comparing the FTFS method to prior studies . . . . .	67
5.3	Evaluate the effectiveness of using FGS on multiple classifier approaches across sixteen cancer expression datasets. . . . .	70
5.4	A comparison of five classical classifiers vs Fuzzy classifier method . . . . .	86
5.5	A comparison of FGS vs FGSWP across five classical classifiers. . . . .	96
5.6	A comparison of FGS vs FGSWP using FC method. . . . .	102
5.7	Comparing the developed model against prior studies . . . . .	103

# Abbreviations

## Roman Symbols

*ABC* Artificial Bee Colony

*AC* Accuracy

*ACC* Adrenocortical carcinoma

*ANN* Artificial Neural Network

*ATC* Anaplastic Thyroid Carcinomas

*BASIC* Backward Elimination Hilbert- Schmidt Independence Criterion

*BCDForest* Boosting Cascade Deep Forest

*BLCA* Bladder Urothelial Carcinoma

*BPSO – DT* Binary Particle Swarm optimization with Decision Tree

*BRCA* Breast invasive carcinoma

*BRCA* BReast CAncer gene

*CCA*) well- known canonical correlation analysis

*CESC* Cervical squamous cell carcinoma and endocervical adenocarcinoma

*CHOL* Cholangiocarcinoma

*CNN* Convolutional Neural Networks

*CNS*) Central Nervous System

*COAD* Colon adenocarcinoma

- 
- CSFS* Consistency Subset Feature Selection
- DGS* Deep gene selection
- DL* Deep Learning
- DLBC* Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
- DNN*) Deep Neural Network
- DRE – DNN* A differential regulatory network embedded deep neural network
- DT* Decision Tree
- ELM*) Extreme Learning Machine
- ELM* Extreme Learning Machines
- ESCA* Esophageal carcinoma
- EVD* Extreme Value Distribution
- F1* F1-score
- FC* Fuzzy Classifier
- FGS* Fuzzy Gene Selection
- FGSWP* Fuzzy Gene Selection-Wrapper Plus
- FS* Feature Selection
- FTFS* Fusion Three Feature selection
- GA* Genetic Algorithm
- GBM* Gradient Boosting Machines
- GBM* Glioblastoma Multiforme
- gcFores* Multi-Grained Cascade Forest
- GEO* Gene Expression Omnibus
- GEP* Expression Programming
- GGA*) Grouping Genetic Algorithm

- 
- GNB* Gaussian Naive Bayes
- HNSC* Head and Neck squamous cell carcinoma
- ICA* Independent Component Analysis
- IDL* Integrative deep learning
- IFS* Incremental Feature Selection
- IG* Information Gain
- IWSS – MB*) incremental wrapper-based subset selection with Markov Blanket
- KICH* Kidney Chromophobe
- KIRC* Kidney renal clear cell carcinoma
- KIRP* Kidney renal papillary cell carcinoma
- KNN* K-Nearest Neighbors
- LAML* Acute Myeloid Leukemia
- LGG* Brain Lower Grade Glioma
- LR* Logistic Regression
- LSTM* Long Short Term Memory
- LUAD* Lung adenocarcinoma
- LUSC* Lung squamous cell carcinoma
- MCSF* Monte Carlo
- MDFDL* Multidimensional Fuzzy Deep Learning
- MI* Mutual Information
- ML* Machine Learning
- MLP* Multilayer Perceptron
- mRMR* Minimum redundancy maximum relevance
- Multi – ROC* Multi-category Receiver Operating Characteristic

---

<i>NB</i> )	Naive Bayes
<i>NCA</i>	Neighbourhood Component Analysis
<i>PCA</i>	Principal component analysis
<i>PRAD</i>	Prostate adenocarcinoma
<i>Pre</i>	Precision
<i>PTC</i>	Papillary Thyroid Carcinomas
<i>Rec</i>	Recall
<i>RF</i>	Random Forest
<i>RFB</i>	Recursive Feature Elimination
<i>RLR</i>	Randomized Logistic Regression
<i>RNN</i>	Recurrent neural network
<i>RSS</i>	Random Subspace
<i>SBC</i>	Structural Binary Classification
<i>SDAE</i>	Stacked denoising Autoencoders
<i>SOM</i>	Sequential minimal optimization
<i>SVDE</i>	Singular Value Decomposition Entropy
<i>SVM</i>	Support Vector machine
<i>T1</i>	Stage1
<i>TCGA</i>	The Cancer Genome Atlas
<i>TL</i>	Transfer Learning
<i>UCEC</i>	Uterine Corpus Endometrial Carcinoma
<i>WS</i>	Wrapper Subset
<i>XGB</i>	eXtreme Gradient Boostin

# Chapter 1

## Introduction

### 1.1 Introduction

According to the World Health Organization, cancer is a universal danger that kills people all over the world [1] [2]. Cancer is defined as a group of cells that grow from sites of the human body and often spread to distant metastatic areas [3]. Abnormal cell development results from the intricate interplay of genes (deregulated by mutation and epigenetic alterations) and the environment (i.e. carcinogens), some patients inherit mutations from their parents (i.e. BRCA) but most cancer mutations are somatic (i.e. mutations acquired during the life course) [4]. As a result, whole-genome expression analysis has become a valuable tool for identifying significant gene pathways that are dysregulated and cause aberrant cellular proliferation and metastatic spread. Whole-genome expression (transcriptomic) analysis offers the ability to enable early cancer prediction, diagnosis, clinical outcome determination, and the possibility of spreading illness. Moreover, molecular cancer classification can improve the efficacy of customized therapies such as immune-checkpoint inhibitors including anti-PD1 and anti-CTLA-4 [4]. Using transcriptomic approaches such as microarrays and, more recently, RNA-Seq to measure gene expression differences between healthy and unhealthy tissue has required researchers to develop bioinformatic pipelines that include mathematical and statistical methods to analyse these large novel datasets. Typically, the identification of biomarker genes involves the identification of a subset of genes associated with a specific disease or subset of diseases. Identification of new gene signatures will aid in the early prediction of cancer and may help with identifying patient subgroups that may be responsible for or resistant to particular therapies [5]. In this regard, biomarkers have the potential to play a pivotal role in the determination of treatment strategies [6]. The recent availability of these datasets publicly enables researchers to apply deep learning techniques that may

accelerate data analysis and vastly improve the accuracy of cancer diagnosis, prognosis, and anticipated response to therapy.

Although resection is the one of methods for treating cancer, it is not always possible to perform surgical removal because it could severely damage surrounding tissue in some sensitive areas, such as the spinal cord. Early cancer detection contributes to cancer treatment such as resection. On this basis, attempts have been made to analyse various types of patient data to identify a method for detecting cancer at an early stage, thereby facilitating the removal of the affected area and preventing it from spreading to other parts of the body. A new method was discovered for calculating the expressed level of gene activity in cancerous and noncancerous tissues. Common measurement techniques for calculating the gene expression of thousands of genes in hundreds/thousands of samples include: (Microarray and RNA-seq methods). Those methods have an advantage over previous methods because they aid in the early detection of cancer, and they also allow for personalised treatments [4]. Microarray and RNA-seq techniques have provided the opportunity to propose different mathematical and statistical methods for analysing this massive dataset.

These Microarray and RNA-seq datasets contain noise, missing, and duplicate data, necessitating the development of new methods for resolving these issues in the gene expression data [7] [8]. Consequently, novel, and potent Feature selection (FS) and DL methods are being employed to analyse gene expression.

In general, FS is a statistical method that aims to select important genes and disregard unrelated ones [9]. The use of FS approaches to the datasets contributes to enhancing ML classifiers, resulting in reduced complexity of a classifier, higher classifier accuracy, and reduced dimensionality of the data [10]. In addition, it reduces the noise in the data that may lead to overfitting of DL approaches, thereby preventing or mitigating overfitting of DL approaches [11]. In the last decade, there has been a great deal of interest in DL in most fields, including image recognition, traffic prediction, and medical diagnosis. Feature selection methods are continuously developing to select a subset of informative genes that can be used as identifiers for training a classifier model. In addition, developing a classifier that accurately fits a small number of genes with a small number of samples. This encourages the continued development and proposal of a novel classifier algorithm to achieve accurate cancer classification or prediction commensurate with the problem size and sensitivity of the cancer subject.

Machine learning is a sub-field of artificial intelligence that allows computers to learn without being explicitly programmed [12]. Many different approaches of classical ML have been developed, including, KNN, SVM, DT, RF, NB, etc. Deep learning also called deep neural network (DNN) has shown some breakthroughs in recent years due to the increase

in computation power. DL can be defined as a sub-field of machine learning that works by generating a structure that has multiple layers in which the next layer of input is the output of the previous layer (Fig 1.1) [13]. DL structure aims to mimic the mechanisms of the human brain by interpreting the various types of data including sound, text, and images [14]. It uses principles similar to that of linear regression, where each neuron has a weighted value that is updated by applying a gradient descent algorithm through back-propagation to reduce global loss of function [15]. DL approaches contribute to deal with the difficulties of cancer prediction by speeding up analysis whilst maintaining accuracy. The most common architectures of DL are CNN, RNN, and ANN.

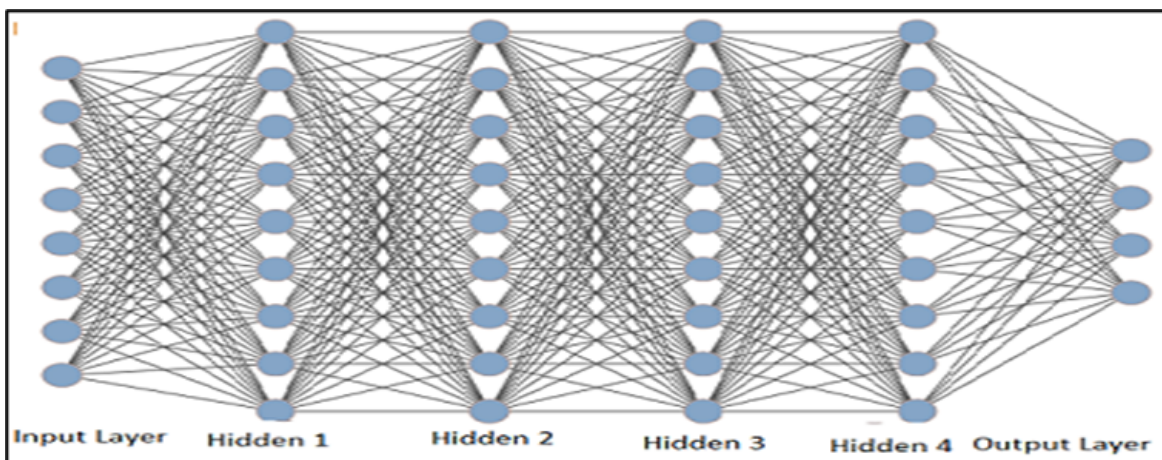


Fig. 1.1 Deep Learning Structure  
[16]

Classical ML and DL approaches both allow computers to learn from input data without the requirement of explicit programming. However, DL does not require human effort to generate feature extraction in the way that classical ML does (Fig 1.2) [17] [18]. It is also more efficient than ML when dealing with very large datasets [19]. Recently DL showed higher performance than human performance on tasks that involves image classifications (Fig 1.3) [19].

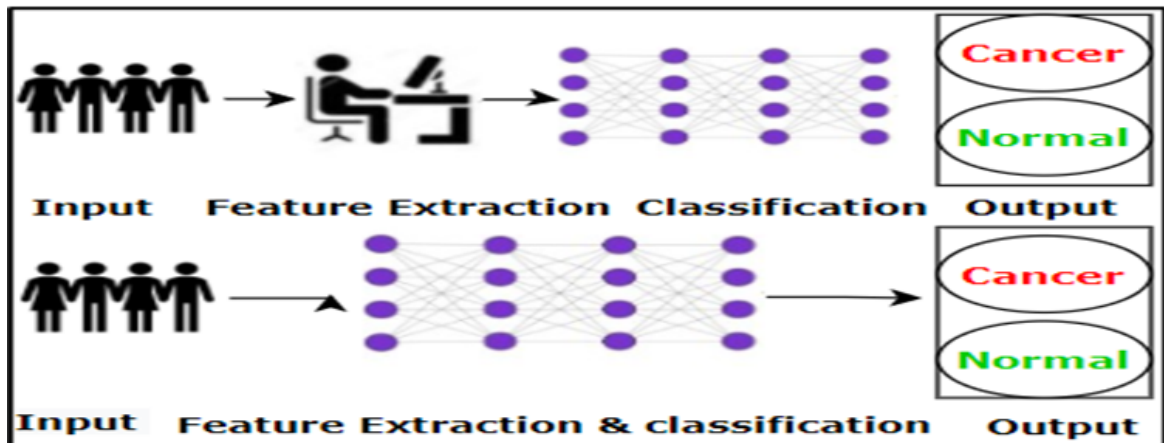


Fig. 1.2 Deep learning vs classical machine learning

[17]

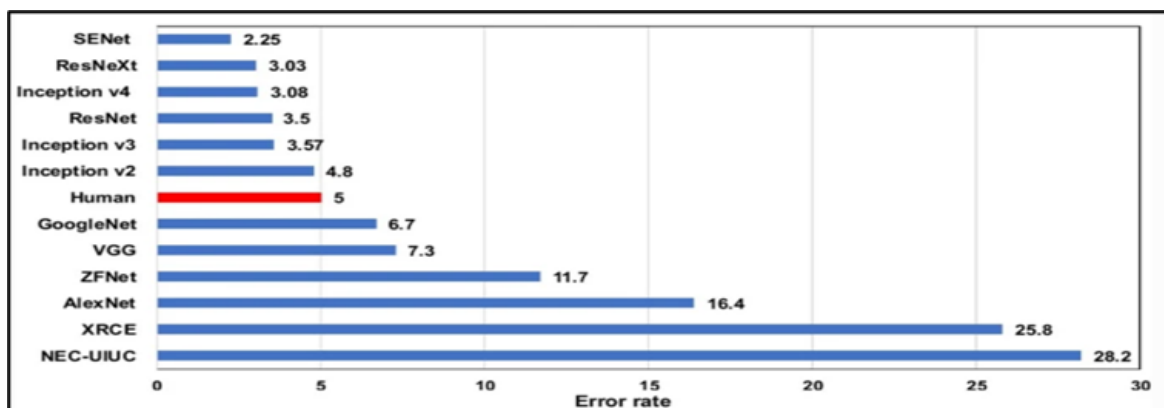


Fig. 1.3 Deep learning performance vs human performance

[19]

## 1.2 Research Problem

In the preceding decade, numerous studies have been proposed to address classification issues by providing different machine learning algorithms with different data formats (i.e. computed tomography CT scan, Magnetic resonance imaging (MRI), and gene expression). Recently, gene expression data have been used to classify cancer, which presents new challenges. The use of gene expression data and machine learning techniques, in the classification of cancer presents two major challenges.

The first challenge is the special nature of provided gene expression data. Though measurement tools (microarray and RNA-seq) continuously developed for measuring the

expressed level of thousands of genes and hundred number of samples [20] [21]. Most of the current machine learning techniques were not developed with this type of dataset properties in mind. The available gene expression dataset is distinguished by high dimensionality which is a challenge in classification [22]. Overfitting is the main issue due to the high dimensionality, while the small volume of the dataset has a negative influence on the performance of classifier models [22] [23]. A large number of genes necessitates more time spent training a model, increasing the complexity of classification, erroneous classification, and overfitting [24].

The second challenge is the massive number of irrelevant genes. Most gene expression datasets provide only a small number of genes that are related to the target (disease) [25]. This is not only increasing the time of training or testing classifier approaches but also increases the complexity of classification [26]. The most effective method to overcome this problem is to use gene selection methods that select the best subset of significant genes that aid cancer classification [27]. Therefore, using these genes as predictors (inputs) for training classifier methods.

In conclusion, the high dimensionality of gene expression data is the most significant factor that negatively impacts the performance of a classifier model by increasing the classification's complexity, prolonging the training phase, and boosting the probability of overfitting. In addition, with so many genes, early cancer diagnosis is difficult.

## **1.3 Aim and Objectives**

### **1.3.1 Thesis Aim**

The thesis aims to develop a multidimensional fuzzy deep learning approach for improved cancer classification using gene expression data.

### **1.3.2 Thesis Objectives**

- Develop a fuzzy gene selection method to reduce the dimensionality by selecting informative genes that would be used as identifiers for training the classifiers to demonstrate FGS effectiveness with these classifier algorithms.
- Develop a novel fuzzy classifier method for improving cancer classification accuracy.
- Develop a fuzzy gene selection wrapper plus method to reduce the number of genes that have been selected by FGS without sacrificing classification accuracy and other evaluation metrics.

- Based on the outcomes, a multidimensional fuzzy deep learning approach developed as a novel system intending to select a limited number of significant genes to be used as a marker. Furthermore, It gives accurate cancer classification and straightforward classification.

## 1.4 Contributions of the thesis.

To address some deficiencies in the field of cancer classification, the thesis has made these contributions described as follows.

- Develop fusion-based three-feature selection methods (FTFS) to reduce the number of genes and increase the performance of the classifiers.
- A novel fuzzy gene selection method (FGS) has been developed to reduce the dimensionality of gene expression data by selecting a subset of significant genes to train classifier algorithms. Selecting a small number of useful genes improves the performance of classifier algorithms by increasing accuracy, reducing complexity, reducing training time, and preventing or at least mitigating overfitting.
- A novel fuzzy classifier (FC) approach has been developed to enhance the generalisation and accuracy of the classifier. When applying classical classifier algorithms to multiple datasets, each classifier technique frequently yields the highest accuracy for a single dataset, as opposed to the FC classifier consistently yielding the highest accuracy for many datasets.
- Develop a fuzzy gene selection wrapper plus (FGSWP) to reduce the number of genes selected by the fuzzy gene selection method while preserving its accuracy. To validate the FGS-selected genes and their effect on the achievement of accuracy.
- The system is a fully automated deep neural network that takes the entire dataset as input and produces the classification result in a process that goes from beginning to end automatically. This is the most straightforward method to use the system, preventing the loss of information, and reducing the number of errors that may occur when feature selection is done separately. In addition to this, it is faster than the manual method. To achieve that the developed approaches ( FGSWP, and FC) are combined into a single model known as multidimensional fuzzy deep learning (MDFDL).

## 1.5 List of Publications

1. M. Khalsan et al., "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," in *IEEE Access*, vol. 10, pp. 27522-27534, 2022, doi: 10.1109/ACCESS.2022.3146312. **Published**
2. M. Khalsan, M. Mu, E. S. Al-Shamery, L. Machado, M. O. Agyeman and S. Ajit, "Intersection Three Feature Selection and Machine Learning Approaches for Cancer Classification," in *IEEE, International Conference on System Science and Engineering (ICSSE)*, Ho Chi Minh, Vietnam, 2023, pp. 427-433, doi: 10.1109/ICSSE58758.2023.10227163. **Published**
3. M. Khalsan, M. Mu, E. S. Al-shamery, L. Machado, M. O. Agyeman and S. Ajit, "Enhancing Cancer Classification Through the Development of a Fuzzy Gene Selection-Wrapper Plus Method," 2023 *IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, Kota Kinabalu, Malaysia, 2023, pp. 39-44, doi:10.1109/IICAIET59451.2023.10291820. **Published**
4. M. Khalsan, M. Mu, E. S. Al-Shamery, S. Ajit, L. R. Machado and M. Opoku Agyeman, "Developing a Multidimensional Fuzzy Deep Learning for Cancer Classification Using Gene Expression Data", 9th *International Conference on Computer Science, Engineering and Applications (CSEA 2023)*, 2023. **Published**
5. M. Khalsan, M. Mu, E. S. Al-Shamery, S. Ajit, L. R. Machado and M. Opoku Agyeman, "A Novel Fuzzy Classifier Model for Cancer Classification Using Gene Expression Data," in *IEEE Access*, vol. 11, pp. 115161-115178, 2023, doi: 10.1109/ACCESS.2023.3325381. **Published**
6. M. Khalsan, M. Mu, E. S. Al-Shamery, S. Ajit, L. R. Machado and M. Opoku Agyeman, "Fuzzy Gene Selection and Cancer Classification Based on Deep Learning Model," *The Journal of Supercomputing*, 2023. **Accepted**.

## 1.6 Thesis Structured

The thesis comprises six chapters, with the first chapter serving as the introduction. In this introductory chapter, a holistic overview of the thesis is presented, encompassing topics such as the definition of cancer, machine learning, deep learning, and gene expression. Moreover, it delves into the challenges encountered within this field,

accompanied by a visual representation of the contributions made to surmount these challenges. Additionally, the chapter articulates the thesis's aim and objectives, supplemented by a list of publications relevant to the subject matter. Chapter two offers a thorough examination of various critical aspects within the field, including methods for measuring the expression level of each gene (such as Microarray and RNA-seq), classifier approaches, gene selection methods, data repositories, cross-validation procedures, evaluation metrics, and a comprehensive review of related works. Towards the conclusion of the chapter, a summarisation of issues identified through the discussion of studies in this field is provided in the 'Related Work' section. Chapter three, dedicated to methodology, offers a detailed account of the developed approaches devised to address the challenges identified in this field. This chapter comprises graphical representations, algorithmic frameworks, step-by-step procedures for each of the developed approaches, and an elucidation of the rationale behind their developments.

Chapter four is dedicated to elucidating the experimental frameworks that underpin the implementation of the developed approaches. This chapter meticulously breaks down the experimental procedures into a series of well-defined steps, accompanied by illustrative figures to enhance clarity, and provides comprehensive descriptions for each of these crucial steps. Chapter five is focused on presenting and analysing the experimental results obtained from the developed approaches. It encompasses a thorough discussion of these findings, drawing comparisons with existing methodologies and published works. This chapter used tables to succinctly convey the results of each experiment and employs boxplots to provide a visual representation of the outcomes. Moreover, it concludes with a comprehensive examination of the results within the context of the methodological discussions. Towards the chapter's conclusion, it conducts a comprehensive evaluation of the entire system in comparison to other published works, offering a concise summary of the key findings from all the developed approaches.

# Chapter 2

## Background and Literature Review

### 2.1 Microarray and RNA-seq

Gene expression is the complex mechanism by which the genetic instructions stored in DNA are harnessed to generate vital proteins and biochemical compounds necessary for the growth, development, and operation of living organisms [28]. Based on that there are different tools developed to measure the gene expression data such as Microarray and RNA-seq. Microarray aimed to allow biologists to monitor the level of gene activity in an organism [29]. This is accomplished by gauging the expressed levels of each gene between cancerous and non-cancerous tissues. This method assists investigators in comprehending tumour classification and progression by monitoring gene expression changes with cancer [30]. This allows the identification of the top differential expressed genes that might be associated with a particular disease. Perhaps more importantly pathway and gene ontology enrichment analysis can be used to identify putative biological pathways involved in the disease based on prior experimental work that has already been done with these genes. Technically, Microarrays measure the intensity of fluorescence (fluorescently labeled cDNA molecules), where the intensity of fluorescence reflects the corresponding gene expression levels. The ability of microarrays to measure the expression of thousands of genes concurrently relies on slides (known as DNA Microarrays) pre-spotted with thousands of probes complementary in sequence to the fluorescently labeled cDNA molecules that are added to the array (ordinarily pointed to as DNA/Gene chips). The known position of the probes on the chip allows the assignation of gene expression patterns to single genes [31]. To implement a microarray experiment, both a reference sample (e.g., from normal tissue) and an experimental sample (i.e., cancer tissue) are collected, the mRNA is extracted and converted to fluorescently labeled cDNA typically with one sample labeled with a green, fluorescent marker and the other a red fluorescent marker. Then the two samples are combined and hybridized into the

microarray slide. RNA-seq is a measurement tool employing next-generation sequencing technology (e.g., Illumina HiSeq) which can be used to determine gene expression and sequence differences between different types of biological samples and has largely superseded microarray technology.

RNA-Seq also provides single base pair resolution, distinguishing allelic expression of genes, identification of novel genes and altered splice forms and has a larger dynamic range, a good signal-to-noise ratio, and is more accurate in measuring gene expression levels [32] [33]. RNA-Seq requires mapping of processed sequence reads to a reference genome (or transcriptome) and therefore is dependent on the accuracy of these reference assemblies rather than being limited by a predetermined choice of probes (on an array). This allows an increased ability to identify new gene-disease associations that may have been missed with microarray approaches [32]. RNA-Seq can detect expression at the gene, exon, transcript, and coding DNA sequence (CDS) levels, whereas Microarray can detect expression in a gene, exon-level only [34].

## 2.2 Feature selection

Feature selection is a high level of pre-processing that uses to decrease the number of features for a given dataset [34]. FS has a positive impact not only on reducing the number of features but also assists in enhancing the performance of classifier models by mitigating the time taken during the training of a classifier model, reducing overfitting [7]. It is also improving the classification accuracy because FS aims to select the features which are highly correlated with the target (class). FS methods are widely employed in several different fields including biology, engineering, computer science, and others [35]. The abovementioned microarray and RNA-seq technologies characterised that provide high-dimensionality datasets of gene expression, so FS has become an elementary method for processing high-dimensional datasets.

Generally, FS techniques are divided into four main categories filter methods, wrapper methods, embedded methods, and hybrid methods [36] [37]. All these methods aim to reduce the number of features but are technically different [38]. Filter methods attempt to select a subset of informative features independent of the classifier algorithms [39]. Filter methods involve MI, mRMR, Relief, and etc. By contrast, wrapper methods aim to select important features by interacting with classifier approaches including forward selection, backward elimination, Bi-directional elimination (Stepwise Selection), etc. While embedded methods strive to obtain the advantage from both methods (Filter and wrapper method) which means they interact with classifier methods but have less computational cost than the

wrapper method by avoiding the recurrent execution of the classifier and study of each feature group. The common embedded techniques are LASSO and RIDGE regression. Hybrid methods are a combination of filter and wrapper methods or embedded and wrapper methods [36] [40]. It aims to take advantage of both by reducing the computational time taken and better results [41]. The employing of feature selection methods is summarised in three prime benefits as follows: 1) To prevent or mitigate overfitting and enhance the performance of the classifier approach. 2) To provide faster and more accurate classification or prediction. 3) To get a deeper insight into the underlying processes that generated the data [42]. The recent advancements in feature selection methods widely used in the gene expression field can be deeply illustrated below.

### 2.2.1 Mutual information

MI is a filter feature selection technique that attempts to select the best subset of features from an original dataset that would be used as predictors (inputs) for training a classifier method. MI is also known as Communications and Information Theory. At first, it is originally developed by Shannon [43] in a seminal paper, to discover the optimal coding of a source on one hand and a noisy channel on the other hand. MI works by measuring the amount of information shared between two random variables [44] [45]. The ability to be a solution for the selection of the informative features for different classifier models and less time taken for selecting these features are the major advantages of using mutual information algorithm. MI is mathematically presented in the equation below.

$$I(X, Y) = \sum \sum p(X, Y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2.1)$$

$$= H(Y) - H(Y/X) \quad (2.2)$$

Where  $H(Y|X)$  is the conditional entropy of Y in the case of X is known.

### 2.2.2 F-Classif

F-Classif is a statistical feature selection algorithm that has been proposed to choose the best subset features of given an original dataset by calculating the ratio between different attributes. Indeed, it computes the variation between features/class labels with samples. It is also a well-known ANOVA f-test [46]. It works by measuring how far a feature is from other features. It is employed to check the means of two or more groups that are significantly different from each other. It scores the features (attributes) and ranks the features rely on

their scores. It allows to selection top feature by either threshold (i.e. top 20 features) or by threshold based on the score of features(i.e. 0.5) both cases are allowable when F-Classif algorithm is employed.

### 2.2.3 Chi-squared

Chi-squared is a statistical technique applied to test the independence of two events. At first, compute the chi-squared between each feature and the target (class label) [47]. Then, select the number of features based on the highest chi-squared scores. The formula for chi-squared is shown below [48].

$$X_c^2 = \Sigma(O_i - E_i)^2 / E_i \quad (2.3)$$

Where: C = degrees of freedom, O = observed value(s), and E = expected value (s).

### 2.2.4 Minimum redundancy maximum relevance

mRMR is a feature selection method that aims to select discriminative features that would be used as identifiers for training a classifier model. It was primarily proposed by [49] in 2005, and it is one of the robustness filter methods. Initially, this technique was typically working for classifying the DNA microarray data [50]. It works by selecting the features that have a high correlation with the target (relevance) and they have the smallest correlation between themselves (redundancy) [51]. mRMR has ranked the features based on minimal-redundancy-maximal-relevance criteria. To calculate the relevance of the mRMR, F-statistic is used for continuous attributes (features) or mutual information for discrete features while the redundancy is calculated by employing the Pearson correlation coefficient (for continuous features) or mutual information for discrete features. Although, it works well with microarray datasets and is applied to most fields. However, it has some limitations, it is usually utilised to involve an extremely sensitive standard relevance and redundancy gauges to the presence of outliers in the data. Moreover, it is computationally expensive.

## 2.3 Classifier approaches

Classifier algorithms aim to accelerate the classification process and improve classification effectiveness. The classification approaches are a supervised learning technique used to determine the category (class, label, or target) of new observations based on training data. In classification, an application or program that is capable of learning from historical datasets and classifying new datasets based on what it learned in the training phase classifies the

new datasets. Consequently, this section of the thesis aims to illustrate the procedure of the proposed common classifier algorithms. This section intends to illustrate the process of developing classifier algorithms and highlight the strengths and weaknesses of each classifier.

### 2.3.1 Support vector machine

Support vector method is initially developed by V.Vapnik in 1965 when he was attempting to solve problems in pattern recognition and continuously developed until formally proposed by V.Vapnik in the 1990s as a support vector machine [52]. SVM is a supervised learning method have been designed for solving both classification and regression challenges. However, primarily it is developed as a solution for classification problems because it achieved outstanding performance in this area. SVM is classified as one of the powerful classifier algorithms because it accomplished good results in most fields (i.e. images, text, biological sequence analysis, and biological data mining) [52]. Originally, SVM is supporting only binary classification, but it can be used for multi classes classification by employing the same principle by dividing the multi classes problem into multiple binary classifications [53]. SVM is working by generating the best line (Hyperplane) to segregate the input into different spaces as shown in Fig 2.1, then it strives to find the hyperplane in n-dimensional space that separates different data points [54] [55]. SVM works relatively well when there is a clear margin of separation between classes, SVM is relatively memory efficient [52].

Though, the SVM algorithm accomplished good results even if the given datasets are providing insufficient information about the data, and it works well with unstructured datasets. However, the SVM model has some limitations, it is not working efficiently when large or noisy (i.e., target classes are overlapping) datasets are employed [52]. Moreover, it presents inaccurate classification accuracy once the high dimensional datasets are used (huge features with a few numbers of samples). Other weaknesses are discovered in the SVM algorithm which are the difficulties of choosing the convenient kernel solution function and takes longer time for the training process especially when a large dataset is used compared to other classifier methods [56] [57].

### 2.3.2 Decision Tree

DT is a supervised machine learning method that is designed as a solution for classification and regression tasks, but it is frequently applied for solving classification purposes. It has been widely used for classification in several areas (i.e. finance, marketing, engineering, and medicine) [59]. DT is a “Tree-shaped diagram representing a sequential decision process in which attribute values are successively tested to infer an unknown state” [60]. The structure

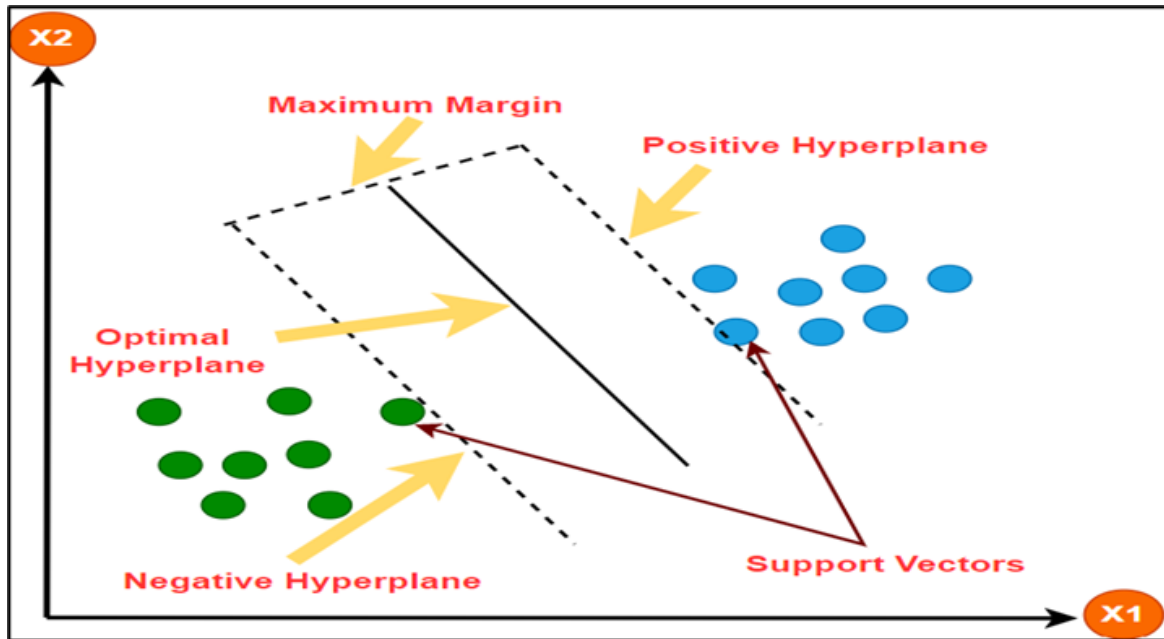


Fig. 2.1 SVM and its Hyperplane Selection  
[58]

of DT is quite analogous to a flow chart which works like a top-down recursive method. Generally, DT consists of three major nodes (root node, internal/test nodes, and leaf nodes). A root node is an initial node of the DT algorithm which represents the entire samples of the given datasets and then splits these samples into sub-groups according to different features, those sub-groups are divided into decision nodes under the root node. These are nodes where variables are evaluated but which are not the final nodes where predictions are made. Leaf nodes are the final nodes that are indivisible where the prediction of class labels is made as shown in Fig 2.2. DT is created by recursively dividing the training samples employing the features from the data that work well for a certain task to achieve this using Gini index or the Entropy for decisive decision trees.

The work of DT can be summarised in five fundamental sequential steps as follows [61]: The entire given dataset is loaded to the first node of the DT which is called the root node to suppose the dataset is  $S$ . Then, find the best features for the given datasets by employing an attribute selection measure (ASM). Divide the dataset into sub-groups that involve the possible values to discover the best features for the given dataset. Therefore, generate decision nodes that include the best attributes that have been selected by ASM in step two. Finally, continuously repeating the third step for the division of the given dataset into sub-groups to make a new decision tree. This process stops only when the possibility of classifying nodes is unable, these nodes are called leaf nodes where each leaf node represents

one class or its probability. Although, DT is an easy and interpretable algorithm and is not required for normalization with less effort for data preparation during pre-processing compared with other classifier methods. However, one of the notable limitations of DT is a small change in the data that may result in a seminal change in the structure of DT which leads to instability. Training time, complexity, and overfitting are other weaknesses shown in the DT when more class data labels are employed.

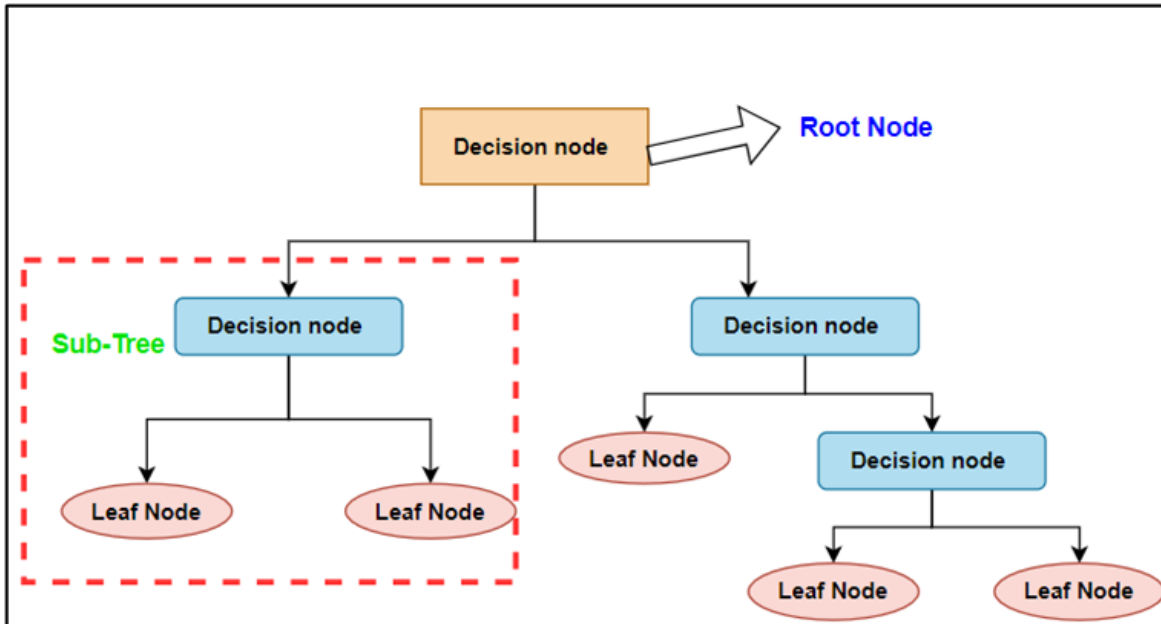


Fig. 2.2 Decision Tree Process  
[62]

### 2.3.3 Random Forest

RF is one of the common machine learning approaches that belongs to the supervised learning method, and it is also called feature bagging or “the random subspace method”. It was first proposed by Tin Kam Ho in 1995 [63]. It can be applied to both regression and classification challenges in machine learning. However, it is not working well with regression problems. RF relies on the concept of ensemble learning which means, it is a combination of multiple decision trees to overcome a difficult issue and enhance the performance of the model [64]. RF is a classifier method that includes a set of DTs on different subsets of the given dataset and takes the average to enhance the predictive accuracy of the dataset. RF works by using multiple trees where it takes the predictions from each tree and depends on the majority votes of predictions and predicts the outcomes as shown in Fig 2.3 [65]. On the other hand, for classification, the predictions are combined using a majority voting scheme while for

regression, the output is obtained relying on the average prediction of these trees. Generally, RF has three major hyperparameters which are required to be assigned before the training stage. These involve node size, tree numbers, and the number of features sampled. Although, RF assists in enhancing the accuracy and avoiding or mitigating the overfitting issue in a decision tree.

Besides, it works well with large datasets, and it does not require data normalisation because it applies a rule-based method. However, it has some weaknesses, it takes more time for training compared to a single decision tree because it is a collection of various decision trees, and it also requires more computational power because it builds several trees to combine their outputs. Moreover, RF is like a black box method, where a very little control over what the model does. Randomisation in both bagging samples and feature selection, the trees in the forest tend to select uninformative features for node splitting. This makes RFs have low accuracy when applied to datasets that characterise high-dimensional data [66]. Large datasets are analysed by random forest which leads to more resources being required to store the data.

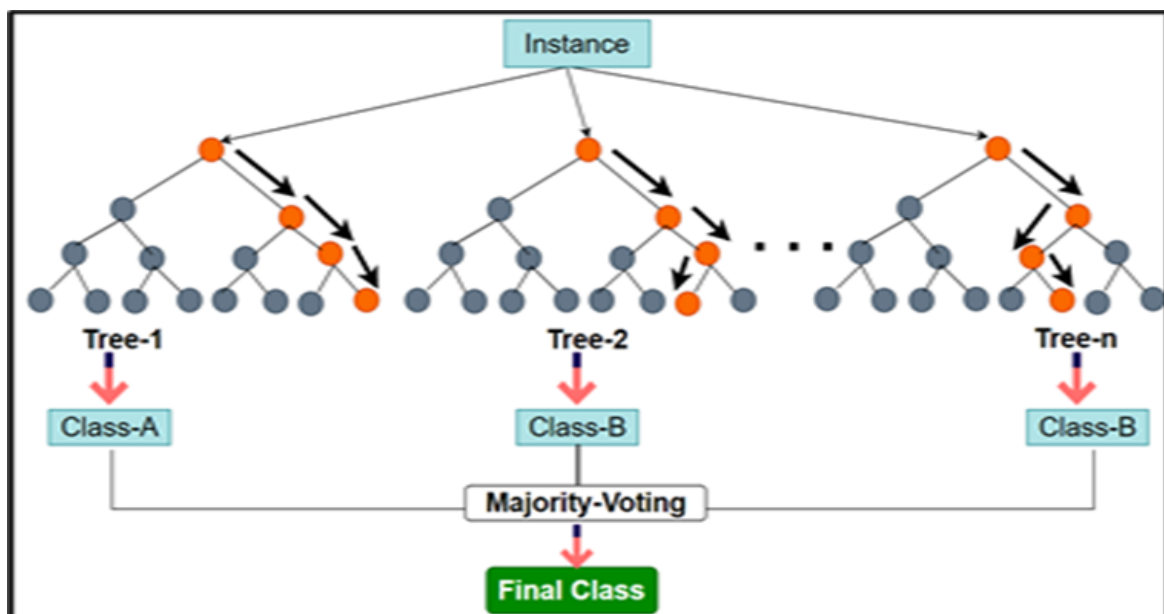


Fig. 2.3 Random Forest Process  
[67]

### 2.3.4 Naïve Bayes

The supervised learning algorithm NB is based on the Bayes theorem and assumes that the presence or absence of one feature is not as significant as the presence or absence of another

feature. Due to its success with dimensional training datasets, it is frequently used to solve classification issues, particularly text classification [68]. As one of the top 10 classifier algorithms in data mining, NB is extensively used in a variety of applications, including spam filtering, text analysis, and recommendation systems [69]. There are generally three different forms of NBs: Gaussian Naive Bayes Classifiers (GNB), Multinomial Naive Bayes Classifiers (MNB), and Bernoulli Naive Bayes Classifiers (BNB). When features are not discrete, GNB is frequently used, MNB is used when the features follow a multinomial distribution, and BNB is employed when the features are of the Boolean type. The main benefits of using Naive Bayes over other classifier methods are its ease of implementation, the accuracy of findings, and the speed of prediction. The NB method, however, presupposes that all of the features are independent, making it impossible to understand the link between the individual variables [70]. Due to comparisons with the suggested model, this thesis is solely concerned with the GNB.

GNB is a supervised learning algorithm that originated from Bayes' Theorem, and it discusses that all characteristics are independent features and that the changes in these features do not influence each other. It is a more effective algorithm for classifying large datasets compared to other classifier algorithms. It works under the principle that independent features which means one feature is independent of the other features [71]. GNB is demonstrated efficacious in classification problems where it is achieved with a minimum error rate [72] [73]. Faster, flexible, working well with large and less time taken during the training stage are the most notable benefits that arise when GNB is applied [74]. However, it has some weaknesses, as it assumes that all features are independent which may lead to missing the algorithm the chance to learn the relationship between these features [68] [75].

### 2.3.5 K-nearest Neighbors

KNN is a supervised machine learning algorithm that can be applied to both classification and regression problems, but it predominantly employs classification purposes [76]. It works under the principle that similar things are near to each other which leads to KNN being mostly used for recommended system applications. In other words, KNN calculates the distance between the tested class and the trained class, the predicted new class relies on the closest distance to the trained class in feature space  $k$  [77]. For instance, in Fig 2.4 the star introduces the new class that is required to classify, is it A or B, in the case of the  $K$  equal to 3 it belongs to class B while it belongs to A when  $k$  is equal to 6. From this example, it can find out that the  $K$  value plays an important role in determining the prediction of the new class. Consequently, finding the best  $K$  value is the most important thing in the KNN

algorithm for getting accurate classification. Generally, there is no standard way of obtaining the best K value.

The most effective way often used for discovering the perfect K value is to adopt a list of different (i.e.  $k=1, k=5, \dots$ ) numbers and check the accomplished accuracy with each used number and select the K value that achieved the best accuracy [78]. This is the only method that has been used to reach the best K value, this method is computationally expensive, and more time is taken [76]. KNN is easy to implement where only one thing which is the calculation of the distance between different points. However, KNN has some cons, it is not suited when large datasets are employed because the calculation of the distance between the points is very costly, and it is also not working well with high dimensional datasets because the calculation of the distance for each dimension is very complicated [76]. Additionally, KNN is sensitive to noise, and missing data and feature scaling is another limitation of KNN where it requires normalisation and standardisation properly [79] [80].

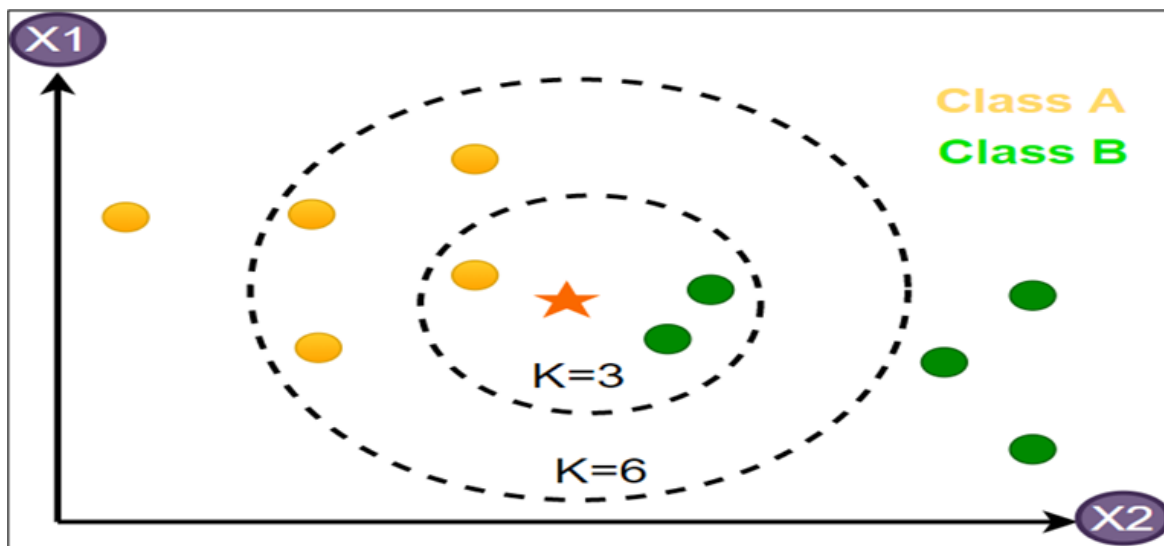


Fig. 2.4 KNN and its Hyperplane Selection [77]

### 2.3.6 Multilayer Perceptron

MLP is a type of feed-forward neural network (ANN) that is widely employed in pattern recognition, classification problems, and prediction [54]. It is fully connected to the dense layers, which convert any input dimension to the required dimension. Generally speaking, an MLP structure must have three major types of layers (input layer, hidden layer, and output layer). The input layer is the first layer of the neural network composed of input neurons that

receive the entire data that is used for further processing by subsequent layers of artificial neurons. The input layer is different from other layers that it does not take any information from the previous layer because it is the first layer of the neural network. The output Layer is the last layer of the neural network that is responsible for producing the output results for a given dataset. The hidden layer is a set of mathematical functions that aim to process the data to get the output. The hidden layer is usually more than one because each one can present a specific task for instance, classifying images, one hidden layer for eyes and another hidden layer for ears, and so on. The hidden layer is located between the input layer and the output layer as shown the Fig 2.5. MLP was initially proposed in 1950 and didn't take its intention until 1986 when the backpropagation algorithm was developed to train the Multilayer Perceptron method. MLP is a deep learning algorithm.

It can summarise the work of the MLP algorithm as follows. Initially, transforming the input data forwarding from the input layer to the output layer. MLP is learned by updating the connection weights between the neurons to ensure a backpropagation method is used after the input data of each node in MLP is processed [81]. Therefore, calculating the errors by identifying the difference between the predicted classes by MLP and the trained label classes and applying supervised learning to learn MLP to minimise the calculated errors. Finally, repeat the three previous steps over multiple iterations to learn perfect weights. The possibility of applying the MLP algorithm to non-linear problems and working well with large and small datasets at the same ratio accuracy achieved and a faster prediction provided after the training are the significant advantages of using the MLP model [71]. However, it has some limitations such as the appropriate functioning of the algorithm based on the quality of the training data in case the method works incorrectly, generalisation issue arises.

### 2.3.7 Logistic regression

LR is a statistical approach for dealing with both classification and regression issues. It is classified as supervised learning. LR is based on the probability idea, which is determined using a sigmoid function [82]. It is often divided into three types: Binomial is only used when there are two alternative classifications in the provided datasets, such as cancer or normal. A multinomial is used when there are three or more unordered categories of the dependent variable, such as "dogs," "cats," and "sheep," whereas an ordinal is used when there are three or more ordered sorts of dependent variables, such as "High," "Medium," or "Low".

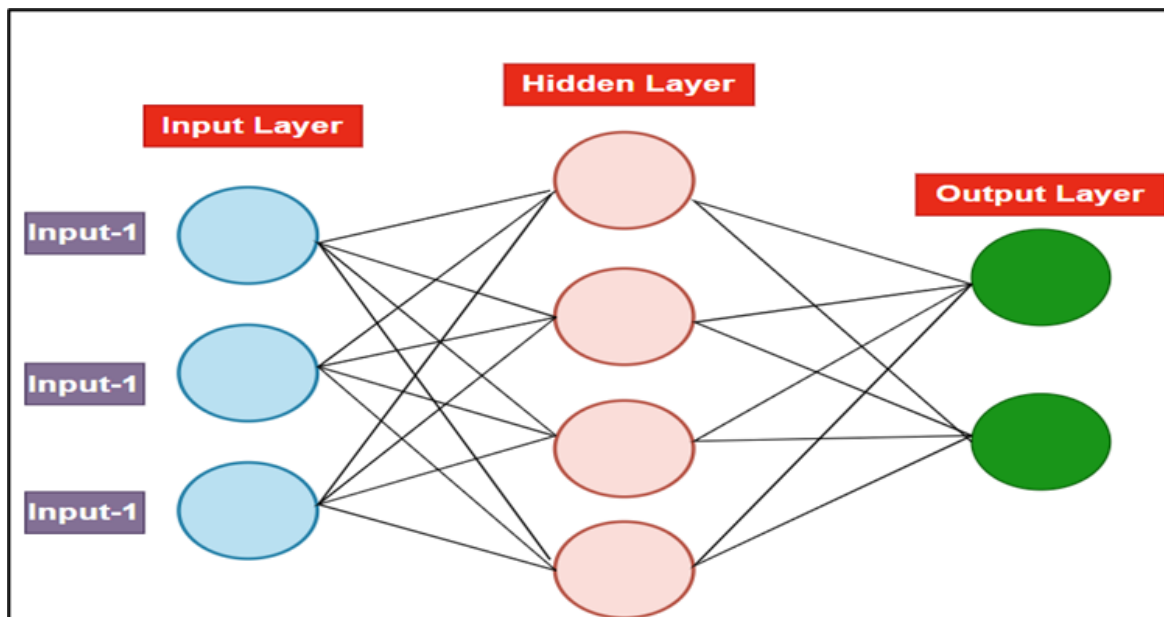


Fig. 2.5 Multilayer Perceptron (MLP) Structure  
[16]

Table 2.1 Comparison of Classifiers for Cancer Classification using Gene Expression Data

Approach	Strengths	Weaknesses	Reference
DT	<ul style="list-style-type: none"> <li>- Intuitive interpretation and visualization capabilities.</li> <li>- Versatile handling of diverse data formats Resilience in the face of missing data entries</li> </ul>	<ul style="list-style-type: none"> <li>- Susceptible to overfitting, particularly noticeable with extensive datasets.</li> <li>- Constrained accuracy and generalization capabilities.</li> <li>- Challenges in capturing intricate relationships within the data.</li> </ul>	[83]
SVM	<ul style="list-style-type: none"> <li>- Demonstrates superior accuracy and generalization, particularly in high-dimensional contexts.</li> <li>- Exceptional performance with non-linear datasets.</li> </ul>	<ul style="list-style-type: none"> <li>- Necessitates meticulous parameter adjustment for optimal performance.</li> <li>- Characterized by a black-box nature, making decision interpretation challenging.</li> <li>- Prone to being influenced by outliers.</li> </ul>	[84]

LR	<ul style="list-style-type: none"> <li>- Requires minimal effort for interpretation and implementation.</li> <li>- Well-suited for processing large datasets efficiently.</li> <li>- Offers probabilities for different classes, aiding in decision-making</li> </ul>	<ul style="list-style-type: none"> <li>- Constrained by its focus on linear relationships, potentially overlooking complex patterns.</li> <li>- Vulnerable to issues related to feature scaling sensitivity.</li> <li>- Performance may degrade when handling highly correlated features.</li> </ul>	[85]
GNB	<ul style="list-style-type: none"> <li>- Streamlined and efficient training process.</li> <li>- Resilient against irrelevant features, enhancing model robustness.</li> <li>- Minimal need for parameter adjustments, streamlining the tuning process.</li> </ul>	<ul style="list-style-type: none"> <li>- Operates under the assumption of feature independence, which may not align with the complexities of gene expression data.</li> <li>- Susceptible to underperformance when confronted with intricate relationships within the data.</li> </ul>	[86]
MLP	<ul style="list-style-type: none"> <li>- Capable of capturing intricate non-linear relationships within the data.</li> <li>- Possesses a versatile architecture adaptable to various tasks.</li> <li>- Well-suited for handling large-scale datasets efficiently.</li> </ul>	<ul style="list-style-type: none"> <li>- Susceptible to overfitting without appropriate regularization techniques.</li> <li>- Demands extensive training duration owing to its complexity.</li> <li>- Presents challenges in interpreting decisions due to its opaque, black-box nature.</li> </ul>	[87]
RF	<ul style="list-style-type: none"> <li>- Integrates multiple decision trees to enhance accuracy and robustness.</li> <li>- Demonstrates resilience against overfitting and noise in the data.</li> <li>- Versatile in its capability to handle diverse data types.</li> </ul>	<ul style="list-style-type: none"> <li>- Presents challenges in interpreting individual features due to its black-box nature May incur high.</li> <li>- computational costs when dealing with large datasets.</li> <li>- Exhibits decreased interpretability compared to single decision trees.</li> </ul>	[88]
KNN	<ul style="list-style-type: none"> <li>- Straightforward and simple to implement.</li> <li>- No need for explicit training procedures.</li> <li>- Capable of accommodating diverse data types.</li> </ul>	<ul style="list-style-type: none"> <li>- Prone to being influenced by outliers.</li> <li>- Accuracy may diminish notably in high-dimensional settings.</li> <li>- Exhibits subpar performance when dealing with imbalanced datasets</li> </ul>	[89]

## 2.4 A Cross-validation

Cross-validation is a statistical technique used in machine learning called cross-validation that seeks to reduce or eliminate overfitting problems in various classifier paradigms. With the use of the cross-validation approach, a model may be trained on several training datasets as opposed to only one. By folding the dataset into many configurations and training the model on each configuration [90]. The model can generalise as a consequence, which is an indication of a robust model. It also helps to show a better indication of the performance of the algorithmic prediction. As illustrated in Fig 2.6, the datasets are divided into k-folds, such as  $k=5$ .



Fig. 2.6 K-Fold Cross Validation Process with  $K=5$

## 2.5 Evaluation metrics

In general, four evaluation parameters (accuracy, precision, recall, and f1-score) are used to evaluate the performance of the proposed classifier model and the five classical classifier approaches that were used for comparison. These measurement parameters seek to determine how well a classifier is doing. These evaluation parameters began as follows:

### 2.5.1 Accuracy

AC is an assessment metric used to identify which classifier is best for a certain dataset. In ML, AC is defined as the ratio of successfully predicted observations to total observations. It

is computed mathematically as follows [91].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2.4)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. A TP is the correctly predicted positive value which means that the value of the actual class is cancer and the value of the predicted class is also cancer. A TN is an outcome where the model correctly predicts the negative class. An FP is an outcome where the model incorrectly predicts the positive class. FN is an outcome where the model incorrectly predicts the negative class.

### 2.5.2 Precision

Pre is defined as the percentage of successfully predicted positive findings to total predicted positive observations [91]. Mathematically, it is illustrated in the formal below.

$$Precision = \frac{TP}{TP + FP} \quad (2.5)$$

### 2.5.3 Recall

Rec is defined as the percentage of retrieved instances out of all relevant instances. It is sometimes referred to as sensitivity [91]. The recall equation is shown as.

$$Recall = \frac{TP}{TP + FN} \quad (2.6)$$

### 2.5.4 F1-score

F1 is defined as the weighted average of precision and recall, with a perfect F1 score of 1 and the poorest score of 0 [91]. In short, it's combined the precision and recall of a classifier algorithm into a single metric. Mathematically, it is described as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (2.7)$$

## 2.6 Data repositories

### 2.6.1 Gene Expression Omnibus

GEO is a global data repository for functional genomics that supports Minimum Information About a Microarray Experiment compliant data submissions [92]. The repository supports both RNA-seq and Microarray data, whereas GEO mostly provides Microarray data. There are a total of 3635328 disease-specific samples provided by GEO. GEO is freely available for experimental use with curated gene expression profiles.

### 2.6.2 The Cancer Genome Atlas

TCGA is a ground-breaking cancer genomics program that provides 84,031 samples of 33 distinct forms of cancer [93]. TCGA provides datasets that are measured by both microarray and RNA-seq instruments. However, the majority of these datasets measure the level of gene expression in normal and cancerous tissues using RNA-seq.

## 2.7 Related Work

### 2.7.1 Classical Machine learning studies

Aydadenta et al. [94] proposed k-means algorithm for the clustering method for selecting a subset of genes, then the Relief method was applied for ranking the result of clustering. The selected genes were used for training the random forest classifier model to evaluate the effectiveness of the proposed feature selection method. The proposed model was trained and tested using three gene expression cancer datasets (colon, lung, and prostate cancer). The proposed model achieved accuracy at 85.87% for colon, 98.9% for lung cancer, and 89% for prostate cancer. The study performed well in lung cancer but achieved low accuracy with the other two cancer datasets and that may achieve lower accuracy with larger datasets or with multi-class datasets.

Rostami et al. [95] developed a novel social network analysis-based gene selection method for selecting a limited number of informative genes that would be used for training different classifier algorithms. This method was developed by integrating node centrality and community detection concepts. The study employed five microarray datasets and four classifier techniques. The highest average accuracy for the five datasets was 87.7% when the Extreme Learning Machine classifier was used. Although, the proposed model achieved better accuracy when compared to other gene selection methods. However, the accomplished

results were not high when compared to other studies in this field. Another disadvantage, the system was tested only using microarray datasets. Additionally, it used a multi-phase approach which is computationally expensive in high-dimensional datasets.

Four classifier approaches were used for distinguishing between two breast cancer subtypes (triple negative and non-triple negative) using RNA-seq gene expression datasets [96]. Additionally, different feature selection methods at different thresholds were employed to select crucial genes that were used for training the classifier models. The datasets include 110 triple negative and 992 non-triple negative samples that were downloaded from TCGA. The highest achieved accuracy was with applying SVM compared to other classifier methods that were employed in this study (KNN, DT, and NB). mRMR method accomplished the highest classification when 32 genes were selected, which was 83%. Although, this study used different ML methods and different feature techniques but was not achieved a good results compared to other published work . Another limitation was the use of unbalanced datasets and that impact on the performance of classifier models. Furthermore, the study used only one dataset, and that is not enough to evaluate a proposed model according to several research in this field. More importantly, this study used conventional classifier approaches which are not working well with large datasets, high dimensionality datasets (number of features larger than samples), or noise data as provided gene expression [55].

Wang et al. [97] developed a novel feature selection method (IWSS-MB) for selecting the best subset of genes that was employed for classifying six gene expression datasets (microarray datasets) that were obtained from GEO. This study used two classifier algorithms (KNN, and NB) for evaluating the effectiveness of applying IWSS-MB. 87% as the average accuracy over entire datasets that were employed in this study by using KNN and IWSS-MB methods together. While 90.6% was the average accuracy when NB and IWSS-MB methods were applied together. Despite, the proposed novel feature selection (IWSS-MB) working well in terms of reducing the dimensionality of employing datasets (reducing the number of genes) and reducing the time taken during the training stage, the achievement accuracy was good compared to other studies. This is primarily due to the limitations of classical classifier models when applied to high-dimensional and large datasets, such as gene expression data. These models are not inherently designed to handle such complex datasets, as calculating distances between each data instance becomes computationally expensive and can be sensitive to noise and missing values.

Extreme Learning Machines (ELM) have been developed as classifier techniques with Correlation Coefficient as a feature selection method to reduce the number of genes and increase the performance of this kind of sophisticated model [98]. The study used 60 central nervous system tumours (in addition to other tumours). The study accomplished a 79%

accuracy score. A caveat to this study is that it used a small dataset. In addition, it also used only one gene expression dataset for evaluating the performance of the proposed model, and that may only show good results for this specific dataset. It may not be broadly applicable.

Five machine learning techniques, namely RF, SVM, NB, C4.5, and KNN were used for classifying breast cancer [99]. In this study, somatic mutation data, from 358 patients, was obtained from TCGA and was used to predict breast cancer. The highest accuracy accomplished with this study was 70% using RF, while the other machine learning techniques used in the study were less accurate ranging between (49 to 69). Gene expression (TCGA and GEO) and DNA methylation data (TCGA) were used to develop a classifier that effectively discriminated lung adenocarcinoma from lung squamous cell carcinoma cases [100]. The dataset for both lung cancer subtypes was collected from [101]. They applied a feature selector (Relief/Limma) to select 30 top relevant scoring variables associated with these lung cancer subtypes. This was reduced from 27,578 DNA methylation variables and 17,814 genes from microarray gene expression. The study achieved an AUC classification performance of 89% by applying a Naïve Bayes classifier and gene functional analysis using the Ingenuity Pathway Analysis tool (IPA) and identified 19 genes, four of which were specifically associated with lung cancer subtypes (AKR1B10, AQP10, CXCR2, TP73).

Yuan et al. [102] used RF and SVM to classify two subtypes of lung cancer: (Adenocarcinomas (AC) and Squamous Cell Carcinomas (SCC)). The study also applied MCSF and IFS methods to identify informative genes. Affymetrix U133 arrays (probing 20,502 genes) were used to generate data from 77 lung AC and 73 lung SCC samples from Gene Expression Omnibus (GEO GSE43580). The study showed that when 1100 optimal genes were selected for classification using an SVM classifier, higher accuracy was achieved compared with the use of 43 informative genes obtained using an MCSF method. Accuracy decreased from 96% to 86% using SVM and 93% to 88% with RF when 43 genes chosen. Other study used different machine learning techniques with multiple GEO ovarian cancer datasets (GSE12172, GSE14407, GSE9899, GSE37648, GSE18521, GSE38666, and GSE10971) used for predicting ovarian cancer (n=530 cases) [103]. They used a 26-gene set panel for training different machine learning predictive models. The study achieved the highest accuracy of 89% using a Random Forest pipeline. The drawbacks of the study were the use of an imbalanced dataset, and the achievement results require improvements.

Tarek et al. [104] used KNN and three feature selection methods to accurately classify three cancer types leukemia, colon, and breast cancer. The three feature techniques are used to select significant genes to enhance cancer classification. The dataset used for testing the system again was obtained from TCGA. The dataset had 6,500 colons, 24,481 breast, and 3,571 leukemia samples. KNN was applied with three feature selections SVDE, EVD, and

BASIC). The accuracy achieved with the system was 80% for colon cancer, 91% for breast cancer, and 92% for leukemia.

In a study of microarray gene expression data from Lung adenocarcinoma samples (86 tumour samples and 10 non-tumour samples collected from Kent Ridge Bio-Medical Dataset Repository available from [105] with 7129 genes), an Info gain feature selection technique was applied to identify genes strongly associated with cancer samples using 70% of samples as a training set and 30% of samples as a test set. This study applied three classifier techniques to discriminate tumour and non-tumour samples after choosing the candidate genes that had known relevance to lung cancer [4]. Several selected genes were evaluated for biological relevance in lung cancer pathology. The system tested on the dataset provided an output of six genes with high Info Gain scores that might be linked with lung cancer (FABP4, FHL1, CLEC3B, Monoamine Oxidase-A, Platelet endothelial cell adhesion molecule-1, and Selenoprotein P). Additionally, the system employed these as biomarker genes to classify lung cancer by applying three classifier algorithms MLP, RSS, and SMO. The accuracy achieved was 86.6%, 68%, and 91% MLP, RSS, and SMO, respectively.

El-Manzalawy et al. [106] developed a novel multi-view feature selection method to analyse gene expression (RNA-Seq) data in combination with copy number alteration and protein array data to predict renal clear cell carcinoma (KIRC) survival. XGB was applied for training and testing the genes selected by the multi-view feature algorithm and relied on canonical correlation analysis (CCA). The study achieved 76% accuracy. One of the notable limitations of the proposed feature selection methods was that it was performed using unsupervised CCA which may lead to reduced accuracy. Additionally, the study had a quite low accuracy score so it requires some enhancement which might be achieved using supervised variants of the CCA method [107] [108]. The study again did not use additional sources of data for evaluating the proposed model's efficiency.

Xu et al. [109] used a Multi-Grained Cascade Forest (gcForest) and dependent feature selection method for predicting four subtypes of breast cancer. Again, TCGA RNA-Seq data was used, and feature selection was developed for selecting 30 informative genes used for improving classification accuracy and reducing training time. The study compared the gcForest classifier with three different machine-learning approaches (KNN, SVM, and MLP). gcForest showed higher accuracy scores compared with the other classifiers. 92% accuracy was accomplished in this study. Although the research yielded valuable results. However, it has some caveats. The gcForest classifier works under the decision tree principle so it is poorly suited for processing continuous gene expression data and must perform discretisation of the data which leads to information loss. Additionally, the study did not use external data for evaluating the proposed model. Another study used mRMR to select a

small number of informative genes for training the KNN algorithm [110]. The study was used to classify thyroid carcinoma. The dataset was retrieved from (GEO GSE33630) and contains 105 samples with 54,675 probes corresponding to 20,283 protein-coding genes. The study obtained 85.7% accuracy with the top ten genes. Even though the number of genes was reduced however the accuracy was substandard.

Hybrid L1/2+2 Regularization as a feature selection method that aims to select the important genes was proposed [111]. This study used the SVM classifier approach for classifying lung cancer (DI GSE19804). A balanced dataset was employed that includes 60 normal samples of lung and 60 samples of lung cancer. The study achieved 94.17% accuracy with only 10 important genes. Although, the study showed significant improvements in terms of reducing the number of selected genes and showed satisfactory accuracy in somewhat. However, there are some limitations of the study: only one dataset was used for training and testing the SVM classifier rather than using different datasets for different cancer types to evaluate the performance of the proposed model. Results obtained with a single dataset are not practical. Another issue is that additional metrics were not considered in this study, such as precision, recall, and f1-scores. Accuracy alone is not necessarily a suitable method to evaluate a system.

### 2.7.2 Deep learning studies

Cahyaningrum et al. [126] proposed Artificial neural networks (ANN)-genetic algorithm (GA) methods for classifying three microarray gene expression datasets of cancer (colon, lung, and prostate cancers). These three datasets were downloaded from the Kent Ridge Biomedical Data Set Repository which was divided into 62,102,181 samples of binary label classes for colon, prostate, and lung cancer respectively. The outcomes accuracy was 83.33%, 76.47%, and 89.93% for colon, prostate, and lung Cancer respectively. As long as the main focus of the study is improving the performance of the models, the accomplished results were not shown high enough accuracy specifically with prostate cancer. Moreover, the proposed model was not evaluated using big datasets or multi-class labels as shown by several existing research in this field. The use of big data or multi-class labels may not give the same efficiency for the proposed model. More importantly, most of the studies that used different measurement tools for evaluating their suggested models such as recall, precision, and f-score were not considered in this study.

Sun et al. [127] developed a novel multimodal deep neural network (MDNN) algorithm for breast cancer prediction. 24368 genes across 2509 breast cancer and 548 normal samples were used for evaluating the proposed model. The proposed algorithm achieved higher

Table 2.2 A summary of previous studies of applying classical ML to analyse cancer gene expression datasets

Datasets	Gene Selection	classifier	Accuracy	Reference
Kent Ridge Bio-Medical Dataset Repository	IG	SMO, RSS, and MLP	86.6%, 68.3%, 91% for MLP, RSS, and SMO respectively	[4]
TCGA- BRCA	SDAE	SVM, SVM-RBF, and ANN	91.74%, 91.74%, and 94.78%, for ANN, SVM, and SVM-RBF, respectively	[112]
TCGA-BRCA	None	RF, SVM, NB,C4.5, KNN	RF=70%, SVM=69% NB=57%, C4.5=60% KNN=49%	[99]
Leukaemia, colon and breast cancer	BAHSIC, EVD, SVD	KNN	92%,80% ,91% leukaemia, colon, breast cancer respectively	[104]
Leukaemia cancer, prostate, and colon cancer data	signal-to-noise ratio, Fisher, ReliefF-statistics	SVM, KNN	Between 85% and 100%	[113]
TCGA-LUAD	ReliefF	RF	83%	[114]
GSE4922, GSE2034, GSE6532, GSE7390 GSE11121	PCA, Autoencoder neural network	AdaBoost	GSE4922 (75%), GSE2034 (72%), GSE6532 (77%), GSE7390 (75%), GSE11121 (85%)	[115]
Breast cancer	RFB, RLR	SVM, KNN, MLP, DT, RF, LR, Ada (Adaboost), GBM	88.8% was the highest accuracy when RFE and SVM applied	[116]
Harvard Medical School, The University of Michigan, The University of Toronto, Women's Hospital Breast cancer	None	SVM, and C4.5	88%-94% when the Brigham and Women's Hospital dataset was used 83% with C4.5	[117]
GDS3257, DNA methylation from TCGA	ReliefF, and Limma	NB	95%	[100]
GEO,TCGA (lung cancer subtypes)	mRMR, Multi-ROC , IFS	RF, multiclass support vector	86.5%	[118]
GSE43580	MCSF, IFS	RF, SVM	86%-96% with SVM 88%-93% with RF	[102]
GSE6044, GSE2109, and Harvard	None	SBC	93%	[119]

Datasets	Gene Selection	Classifier	Accuracy %	Reference
TCGA (LUAD and LUSC)	DGE, PCA, mRMR Lasso	Xgboost, RF	92.9%	[120]
GEO (Lung, Ovarian, and Colon)	MI, GA	SVM	80% -98%	[121]
GEO (GSE12172, GSE14407, GSE9899, GSE37648, GSE18521, GSE38666, and GSE10971)	None	RF	89%	[103]
TCGA (clear cell renal cell carcinomas )	None	SVM	71.15%	[122]
TCGA (Renal clear cell carcinoma)	multi-view feature selection	XGB	76%	[106]
Central Nervous System tumours	Correlation Coefficient	ELM	79%	[98]
TCGA-BRCA	None	gcForest	92%	[109]
GEO and TCGA (Lung cancer)	None	NB	89%	[100]
GEO Colon cancer, Acute leukemia, Prostate tumor, High-grade Glioma Lung cancer II, Leukemia2 data	ICA, ABC	NB, and SVM	Between 92% and 98% with NB and 93% and 97% with SVM	[123]
TCGA-BRCA	None	Ensemble of LASSO	68%	[124]
GSE68465 and GSE8894	CSFS, WS, GEP, DGS	SVM	85% respectively	[125]
GSE33630	mRMR	KNN	85.7%	[110]

accuracy when compared with SVM, RF, and LR. mRMR was also applied as a feature selection method to reduce the number of genes to enhance accuracy. The accuracy achieved was 82%, 80%, 79%, and 76% for MDNN, SVM, RF, and LR respectively. However, recall values were low in this study (45%, 36%, 22%, and 18% for MDNN, SVM, RF, and LR respectively) and precision was 95% for all algorithms. Although this study achieved satisfactory accuracy, further enhancements are required. Additionally, recall values were very low negatively impacting the proposed model performance. The study was tested only on a breast cancer dataset whereas, most studies have used multiple cancer datasets to prove the validity of the results that have been obtained with their models.

Hila et al. [128] developed a novel feature subset selection with an optimal adaptive neuro-fuzzy inference system for gene expression classification. The study used four microarray gene expression datasets for four different types of cancer (Leukemia, Prostate, DLBC Stanford, and Colon Cancer). The system has been compared against existing classifier models to illustrate the effectiveness of the proposed model. The highest classification accuracy was accomplished at 89.47%, 83.33%, 80.65%, and 73.33% for colon cancer, leukemia, prostate cancer, and DLBC Stanford datasets, respectively. Achievement outcomes have not been high compared to previous studies that have been developed in this field. The datasets that were employed for training and testing the proposed model were very small, all the datasets were less than 100 samples (normal and cancer samples) except the prostate was 102 samples. Another limitation, only binary class datasets were employed in this study the proposed model may not work with the same efficiency with multi classes datasets.

Danaee et al. [112] developed a deep learning approach and an SDAE algorithm as a feature selection method to select informative genes distinguishing breast cancer samples from normal breast tissue samples using RNA-Seq gene expression data. The approach was applied to 1097 cancer samples and 113 healthy controls downloaded from TCGA. Three classifier techniques (ANN, SVM, and SVM-RBF) were used to evaluate the performance of the algorithm, achieving accuracy of 91%, 91%, and 94% respectively.

CNN and combining spectral clustering information processing proposed to classify lung cancer using both protein interaction network data and gene expression data from 639 samples (152 benign and 487 malignant) [129]. The dataset is available from NCBI GEO datasets (ID GSE66499). This study achieved 81%, 88%, 78%, and 74% accuracy, recall, precision, and specificity, respectively. This study, as for the others described did not employ validation data to check the efficiency of the model. Moreover, 190 out of 487 cancer samples were randomly chosen and this explains the results obtained. (190) Malignant samples were randomly selected, and that would not be efficient.

CNN deep learning algorithm with microarray gene expression data across eight cancer types [118]. The largest tumour cohort used in this study was 286 samples for breast cancer probing the expression of 13321 genes, while the smallest cohort size was for brain cancer (42 samples probing 5597 genes). The overall sample size for most tumour types was small compared with prior studies. The study had the lowest accuracy of 41% for one of the two breast cancer datasets using CNN. However, in comparison to alternative approaches (mSVMRFE-IRF and varSelRF), CNN is typically superior in terms of accuracy and minimising the number of genes used for classification.

Long Short-Term Memory (LSTM) as a classifier technique and the Matthews Correlation Coefficient as a feature selection method used for classifying five subtypes of kidney cancer using microRNA (miRNA) data [130]. The dataset for the five subtypes of kidney cancer was obtained from the TCGA dataset. The study achieved an overall accuracy between 88%-92% identifying 35 miRNAs with a strong discriminative ability for renal cancer subtypes. Limitations of this study were the lack of replication in a new dataset to test the efficiency of the suggested model and the models were used on datasets that were not balanced in terms of equal sample sizes between the different cancer subtypes.

Cheerla et al. [131] developed a CNN-based model that uses gene expression data from TCGA for predicting 20 types of cancer. The dataset used 1,881 samples across these 20 cancers profiling the expression of 60,383 genes. The model achieved accuracy between 52% to 78% when applied to single cancer from the 20 types of cancer, whereas the accuracy was between 66% to 93% when applied across the pan-cancer dataset.

Xu et al. [132] developed a novel Deep Flexible Neural Forest (DFNForest) model to classify subtypes of three different tumours (Lung, Breast, and Glioblastoma multiforme) using TCGA RNA-Seq data. It is tested as an alternative to deep neural networks. The study combined two feature reduction methods (fisher ratio and neighborhood rough set) to reduce the dimensionality of the data, avoid overfitting and select informative genes [133]. The Novel DFNForest model achieved 93%, 88%, and 84% accuracy in classifying subtypes of breast, lung, and GBM cancers, respectively. The study highlights the variability in the effectiveness of subtype classification across different tumour types.

Junyi et al. [13] developed a differential regulatory network embedded deep neural network (DRE-DNN) approach from a canonical DNN. DRE-DNN is applied to predict liver cancer (hepatocellular carcinoma) using three datasets (GEO: GSE10143 and GSE14520 and TCGA). The proposed model achieved 86%, 74%, and 72% average AUC values for GSE10143, GSE14520, and TCGA datasets respectively which was improved on conventional DNN AUC values. The study used different sources of data for validation and measuring the performance of the proposed method. It used sufficient datasets to train the

DRE-DNN model, and whilst it was useful as a tool for prognosis it did not accomplish good results for classification purposes. However, it goes some way to addressing the overfitting problem of the model.

Serhat et al. [134] proposed a novel hybrid filter wrapper feature selection method for selecting a subset of informative genes for diagnosis and classification. CNN and ReliefF were applied for classifying different cancer microarray datasets (Ovarian, Leukemia, and CNS). For the CNS data, 60 samples probed for 7129 genes were used for testing the feature selection and classifier technique. ReliefF was applied to select a subset of informative genes to increase the performance of the CNN and reduce the time for training the model. The ReliefF-CNN method achieved 83% (increased from 65%) accuracy with CNS data. Comparatively small sample cohort sizes were used from CNS patients as this was freely available. Therefore, multicentre studies are required that provide larger datasets for analysis. The accuracy scores obtained are suboptimal and may need improvement to inform clinical decision-making. Techniques are required that have high efficiency with small datasets.

CNN and transfer learning (TL) were employed for lung cancer prediction [135]. CNN was used to extract features from high-dimensional datasets. The dataset TCGA for 33 different types of cancer (10535 samples and top 20K most variably expressed genes) was used but the study focused on the lung cancer dataset to test the proposed model. The highest accuracy was 72% densely connected multi-layer feed-forward neural network(MLNN). The investigation had quite low accuracy scores and was limited to one type of cancer. Examination of other cancer types may not achieve the same accuracy.

Table 2.3 gives a brief detailed explanation of the deep learning approaches that were developed to analyse gene expression datasets. The datasets used, approaches used, and accuracy attained are discussed in detail. The table summarises the studies that have been conducted in this field.

## 2.8 Discussion

Numerous research endeavors have showcased the effectiveness of diverse machine learning algorithms within this domain. For instance, SVN and KNN have garnered considerable attention for their adeptness in managing high-dimensional datasets and capturing nonlinear correlations. Notable investigations, such as those [99, 112, 113, 123] have documented encouraging outcomes in discerning between various cancer types utilizing gene expression data.

Table 2.3 A summary of previous studies applied DL to analyse cancer gene expression datasets

Datasets	Gene selection	Classifier	Accuracy	Reference
TCGA-BRCA	None	MDNN, SVM, RF, and LR	82%, 80%, 79% and 76% for MDNN, SVM, RF, and LR respectively	[127]
TCGA (20 types of cancers)	None	CNN	Between 52% and 78% when applied to a single cancer 66% to 93% when applied across the pan-cancer dataset	[131]
TCGA (BRCA, COAD and KIRP)	None	standard Lasso, DeepNeti and DeepNetii	65%,62% and 65% standard Lasso,DeepNeti and DeepNetii respectively for BRCA data. 77%,72%and 75% standard Lasso,DeepNeti and DeepNetii respectively for KIPAN data. 57%, 58% and 57% standard Lasso,DeepNeti and DeepNetii respectively with COAD data	[136]
TCGA (Liver), GSE10143, and GSE14520)	None	DRE-DNN	GSE10143 (86%), GSE14520 (74%), and TCGA (72%)	[137]
TCGA (kidney cancer subtypes)	NCA	LSTM	95 %	[130]
GEO (CNS )	ReliefF	CNN	83%	[134]
GEO (breast cancer subtypes)	None	Deep CNN (DCNN)	96%	[138]
TCGA (Pancancer, BRCA, GBM, LUNG) GEO (Adenocarcinoma, Colon, Brain)	None	BCDForest	, TCGA ( between 41% and 97%) and GEO ( between 92% and 96%)	[139]
TCGA- BRCA	None	CNN	88%	[140]
TCGA-LUAD	ANOVA, TL	CNN,MLNN, and SVM	CNN+TL(68%), MLNN+TL (72%), and ANOVA+SVM (69%)	[135]
GEO (different cancer types )	IFS, mRMR	RNN	73.9% for normal tissues, and 63.9% for cancers	[141]

Datasets	Gene Selection	Classifiers	Accuracy	Reference
Breast cancer	None	DNN	71%	[142]
Cancer subtypes	None	Deep cancer subtype classification (DeepCC)	90%	[143]
Liver Cancer	Deep TL	CNN	1D CNN model = 80.36%, 2D CNN model = 98.86%	[144]
8 Cancer datasets	CSSMO	CNN	99%	[145]
Brain tumor	PSCS	CNN	98.7%	[146]
Prostate Cancer	PCD	AIFSDL	96%	[147]

Similarly, ensemble methodologies such as random forests and gradient boosting have demonstrated significant efficacy in cancer classification endeavors. These approaches amalgamate multiple models to enhance predictive precision and resilience, rendering them apt for scrutinizing intricate gene expression datasets. Investigations [103, 109, 115, 116, 120, 121] underscore the prowess of ensemble methods in delineating informative gene expression patterns correlated with diverse cancer types.

Moreover, in recent years, deep learning methodologies, notably CNN, DNN, and RNN, have garnered substantial attention for their capacity to autonomously extract hierarchical representations from raw gene expression data. Research, as referenced in citations [131, 137, 134, 140, 141] have showcased the potential of deep learning architectures in capturing nuanced features and interactions within gene expression profiles, thereby enhancing classification performance.

Additionally, feature selection techniques have been extensively investigated to augment the interpretability and generalizability of machine learning models in cancer classification tasks. Techniques such as PCA, Relief, and mRMR [100, 118, 114, 141] have been employed to reduce the dimensionality of gene expression data while preserving pertinent information.

Analysis of overall results in the related work section reveals varied performance: 26% of the studies achieved very poor accuracy (0-80%), contrasting with only 7% reaching ideal accuracy (98-100%). Additionally, 12% were deemed poor (80-85%), 12% very good (95-98%), 22% achieved acceptable results (85-90%), and 21% were classified as good (90-95%) as illustrated in Fig 2.7. These findings highlight the need for further studies to enhance cancer classification. Specifically, with 26% of prior research achieving very poor accuracy and only 7% reaching ideal levels, there is clear room for improvement in the field. Additionally, the high dimensionality, while the small volume of the dataset has a negative influence on the performance of classifier models [22] [23].

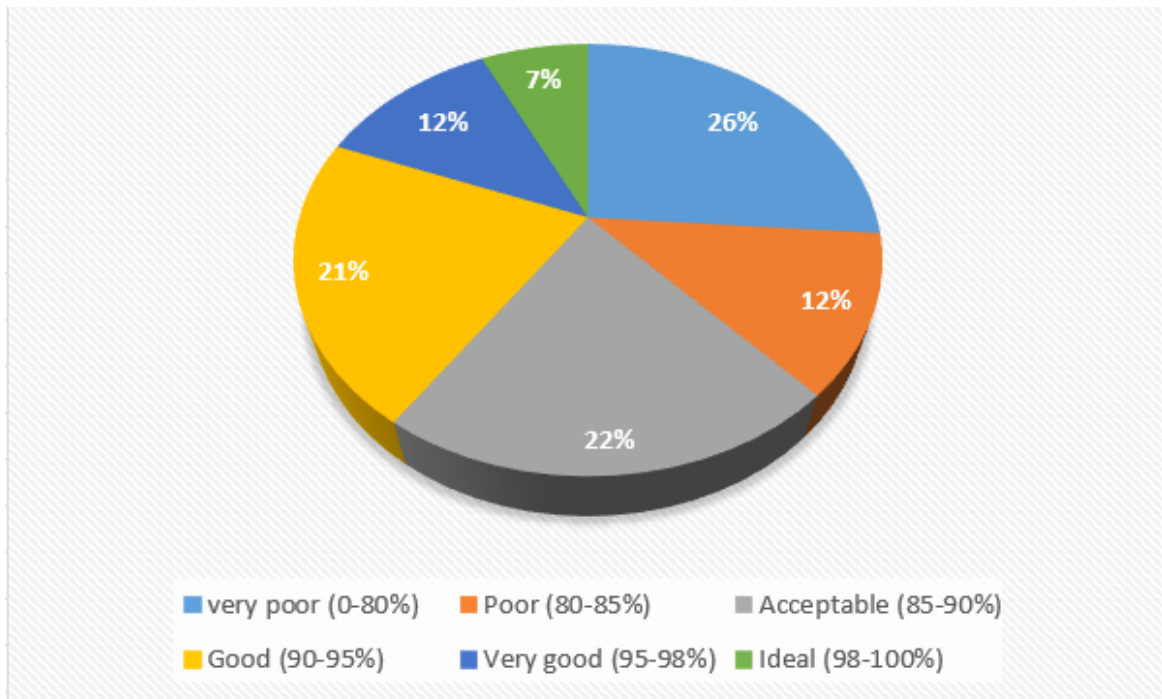


Fig. 2.7 Percentage of achieved accuracy in previous studies

## 2.9 Summary

Several mathematical and statistical approaches have been proposed for different purposes, including cancer classification, biomarker discovery, and cancer prediction using gene expression data. Although gene expression datasets are large in terms of volume, they typically contain relatively small cohort sizes and many variables (e.g., gene expression values). This is a recognised issue for both DL and classical ML algorithms (though to a lesser extent) [148]. In cancer genomics, there are several repositories providing access to high-quality, curated public data allowing training of DL models. However, pre-processing and harmonization are required across these newly developed datasets. Large public data resources for gene expression in cancer are limited to a few key sources (i.e. TCGA and GEO). Most Deep learning techniques require big data to develop accurate models that can be applied to new cancer datasets. To mitigate these existing issues different techniques such as regularization methods (ridge and lasso or L1 and L2), dropout, data augmentation, and reduction of NN complexity have been employed to enhance the performance of the model. However, the issue was not completely solved.

Early cancer prediction and classification enhancement with very high accuracy is necessary and could be achieved by developing mathematical methods with less complexity and less computationally time-consuming. For some tumour types (i.e. gallbladder cancer)

---

classification, prediction, or gene biomarker identification studies using ML or DL is limited. Identifying individual gene signatures for each cancer type is important because it may contribute to early diagnosis of the disease. In addition, knowledge of these gene pathways could have a significant impact in determining the underlying pathology of these tumours and potentially druggable pathways. Some gene pathways may be implicated across multiple cancer types and identifying these may aid risk prediction for individual patients. Cancer subtype classification algorithms need improving and extending across different tumour types to facilitate optimal patient treatment. A notable limitation of AI in analysing gene expression data is known as “The curse of dimensionality” in which higher dimensional data may reveal random effects that do not replicate in related patient cohorts [149] [23].

# Chapter 3

## Methodology

### 3.1 Fusion three feature selection

FTFS aims to identify informative genes with a positive influence on cancer classification, these genes will be used as identifiers for training a classifier model. FTFS used three feature selection methods (MI, F-ClassIf, and mRMR). Then, the intersection concept is employed to select only the genes that have been chosen by the three feature selection methods, while disregarding the others. The development of the FTFS method involves three fundamental stages, as depicted in Fig 3.1. Firstly, three feature selection techniques, namely MI, F-classif, and mRMR, were employed to assign a score and rank to each gene for each technique. Next using threshold to select the top (50, 100, 150) genes with the highest scores for each feature selection technique. In the experiment showed that using top 150 genes better than others As described in Table 3.1. Based on that FTFS was used top 150 genes for each. In total, 450 genes were selected across the three methods. Finally, an additional filtration step was applied using the concept of intersection. This step involved selecting only those genes that were chosen by all three feature selection techniques. Consequently, the final outcome of the developed method was a subset of important genes, which were identified based on their selection by all three methods.

These feature selection techniques, including MI, F-classif, and mRMR are widely favored in gene expression analysis due to their ability to provide diverse perspectives on feature relevance [150–155]. MI captures nonlinear associations between features and target variables [156], F-classif identifies features with substantial discriminative power [46], while mRMR strikes a balance between relevance and redundancy [51]. By integrating these methods, a more comprehensive understanding of feature importance can be achieved. Moreover, the combination of techniques mitigates the susceptibility to various data noise and biases, thereby enhancing the robustness of the feature selection process. This amalgamation

also fosters the generation of more generalizable results, as features identified by one method can complement those selected by another, resulting in a more balanced and informative feature set that is conducive to better generalization to unseen data. Given the typically high-dimensional nature of gene expression data, characterized by the simultaneous measurement of thousands of genes, MI, F-classif, and mRMR are particularly well-suited for navigating this complexity and identifying relevant features amidst noise and redundancy.

Although, FTFS has some advantages such as reducing the number of genes, speeding up the training stage, and enhancing classifier performance. However, it has some limitations such as it used a constant threshold that might lead to ignoring some genes even though have the same score when the feature selection method was used. For example, the threshold for the top 10 genes, which means that gene 11 even if has the same score, as gene 10, however, will be ignored in this case. Another limitation, FTFS used the intersection concept which means only the genes that have been selected by all three feature selection methods will be selected, and others will be ignored. For instance, if a gene scored with mRMR 0.2, with MI 0.1, and F-ClassIf 0.1, FTFS will select this gene even if obtain low scores in all methods. While the FTFS method will not select a gene score of 0.8 for both (mRMR and MI) even has high scores for two of the methods. In summary, FTFS based on the genes selected by the three features and ignores others. It does not take into consideration the gene score for selecting genes. Based on these limitations, the thesis developed the fuzzy gene selection method as described in the next development.

## 3.2 Multidimensional fuzzy deep learning model

This section is used to describe the approaches that have been successfully developed to meet the goal of the thesis. The developed model is divided into four stages. First, pre-processing attempts to clean the data such as removing duplicates, missing data, and normalise the data. Then, a fuzzy gene selection method is developed to identify the significant genes that highly influence cancer classification. Moreover, developing a fuzzy gene selection wrapper plus to reduce the number of genes that have been chosen by the FGS method with keeping the same or improving the accuracy. Finally, a fuzzy classifier method is developed to enhance cancer classification and increase the generalisation of the model to accurately classify cancer in cancer types. In summary, the approaches (FGS, FGSWP, and FC) are integrated into a single powerful framework well-known as the Multidimensional fuzzy deep learning (MDFDL) model. Fig 3.2 depicts the developed model's construction structure in detail.

A novel and efficient deep learning topology is presented, designed for the MDFDL algorithm, that facilitates an end-to-end process for precise gene selection and accurate

Table 3.1 Evaluate the effectiveness of using the FTFS with top 50,100,150

Dataset	Genes Number	Classifier Approach	Ac %	Pre %	Rec %	F1 %
GSE19804	50	kNN	96	95	97	96
	100		100	98	97	97.4
	150		100	100	100	100
GSE19804	50	MLP	98	93	98	95.4
	100		96	95	98	96.4
	150		97	94	100	97
GSE77314	50	kNN	97	96	100	97.9
	100		97	95	100	97.4
	150		96.6	94.4	100	97
GSE77314	50	MLP	95	93	100	96.3
	100		96	94	100	96.9
	150		96.6	94.4	100	97
GSE45827	50	kNN	80	82	100	92
	100		90	100	95	100
	150		97	96	100	98
GSE45827	50	MLP	80	80	90	85
	100		96	94	100	96.9
	150		100	100	100	100
GSE14520	50	KNN	87	89	90	89.5
	100		95	94	95	94.5
	150		95.5	96	95	95.9
GSE14520	50	MLP	90	80	90	85
	100		93	94	100	96.5
	150		95.5	97	97	97
GSE13355	50	KNN	96	90	100	95
	100		95	94	95	94.5
	150		96	96	96	96
GSE13355	50	MLP	90	80	90	85
	100		93	94	100	96.5
	150		98	98	98	98
TCGA1	50	KNN	96	95	95	95
	100		95	93	93	93
	150		95	92.5	93	92.6
TCGA1	50	MLP	90	80	90	85
	100		97	97	95	95.9
	150		97.9	96.7	96.4	96.6

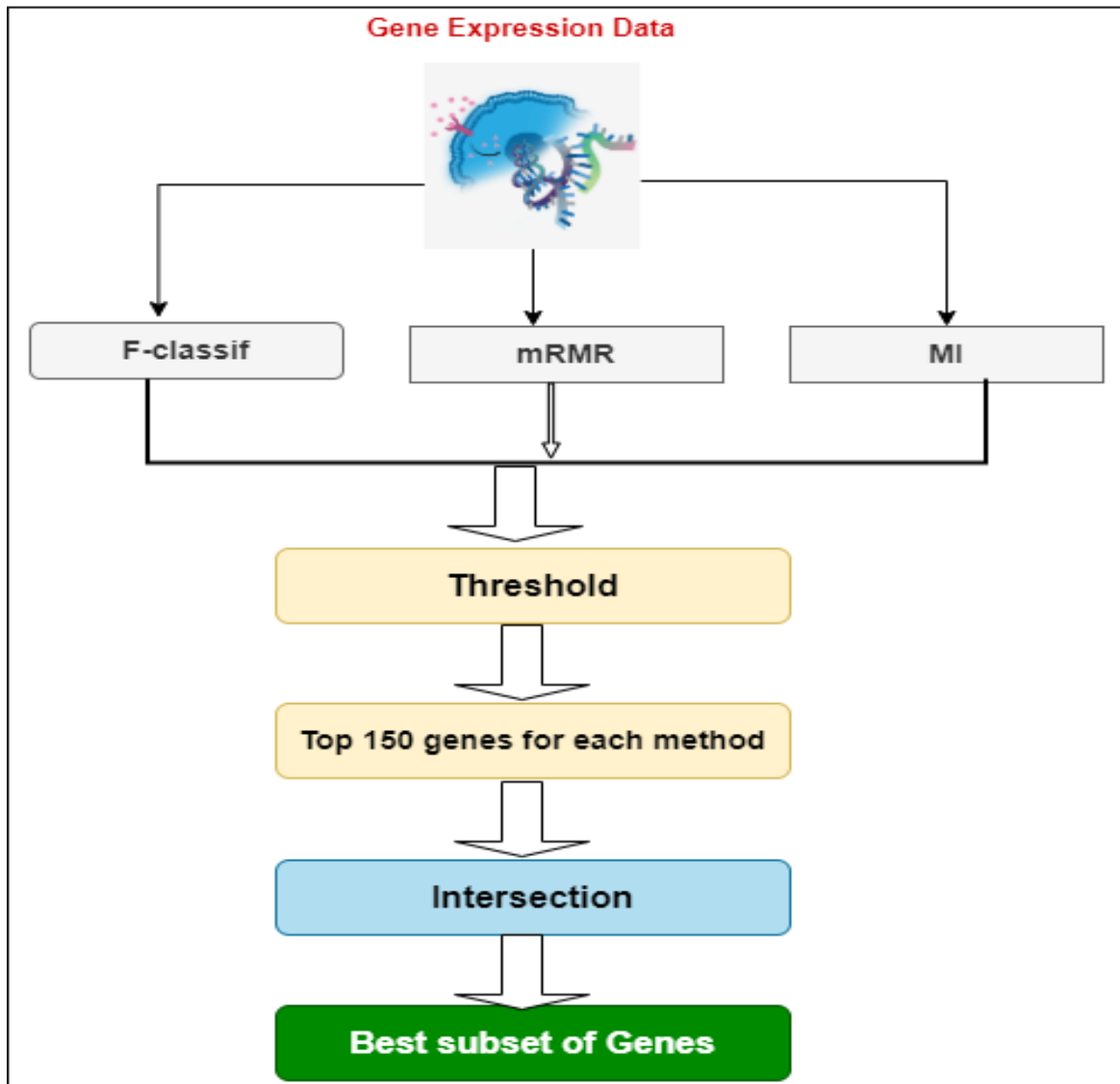


Fig. 3.1 Block diagram illustrates the FTFS process

classification tasks, explicitly classifying (normal vs cancer samples, cancer types, and cancer subtypes). This advanced architecture is illustrated in Fig 3.3. The newly devised topology showcases a unique arrangement of layers, carefully designed to streamline the gene selection process and enhance classification performance. To achieve this, 24 hidden layers were incorporated, sandwiched between one input layer and one output layer.

The key highlight of this architecture lies in its innovative composition, featuring 9 vertical layers strategically positioned to expedite the computation and minimise processing time. These vertical layers synergistically collaborate with an additional 15 sequential (horizontal) layers, optimising the selection of informative genes and facilitating accurate

classification. The new topology demonstrates superior capabilities in terms of computational efficiency, gene selection accuracy, and classification precision.

Overall, the integration of 9 vertical layers and 15 sequential layers makes this novel topology a powerful tool in the domain of gene selection and cancer classification, paving the way for more accurate and timely diagnosis.

### 3.2.1 Pre-processing stage

In this stage three challenges were achieved, described as follows.

- Addressing Missing Values: Missing values can significantly impact the performance of a classifier, so it is crucial to handle them appropriately. In the case of gene expression data, missing values do not exist for a gene's expressed level. However, there might be missing gene symbols. To mitigate this issue, the raw data excluding the genes with missing symbols was eliminated during this step.
- Handle the duplication: removing only the duplicated genes symbol.
- Normalisation is a method mostly used as a section of data preparation for ML and especially, within neural network classifier methods. The major aim of normalisation is to change the values of numeric columns in the dataset to use a common scale, without distorting variations in the ranges of values or losing information. The most important type of normalisation is min-max normalisation which has been used in this work. The below equation is used for calculating the value.

$$V = \frac{v - \min_A}{\max_A - \min_A} \quad (3.1)$$

Where  $\max_A$  is the maximum value of original values for a feature.  $\min_A$  is the minimum value of original values for a feature. and  $N_{\max_A}, N_{\min_A}$  are the maximum and minimum intervals of value.

The Pre-processing is a prior stage to machine learning is included the removal of the raw data that had missing or duplicate gene symbols. A min-max normalisation is a re-scaling of the raw data to be between 0-1 as shown in Algorithm 1.

### 3.2.2 Fuzzy gene selection

FGS was developed to reduce the number of genes in gene expression data to simplify a classifier technique, shorten the training time, improve accuracy, and alleviate overfitting

**Algorithm 1** Data Pre-processing

- 
- 1: Input: Two-dimensional array DS(n,m) where n is the number of samples and m is the number of genes
  - 2: Output: Two-dimensional array DP (n,m) after pre-processing

**Handling Missing Values**

- 3:
- 4: Begin
- 5:  $i = m$
- 6: **while**  $i \neq 0$  **do**
- 7:     **if**  $v$  not in  $i$  **then** ▷ If  $v$  gene symbol is missing
- 8:         ignore  $i$
- 9:          $i = i - 1$
- 10:     **end if**
- 11: **end while**

**Normalization**

- 12:
- 13: Min[j] and Max [j] two arrays holding (minimum and maximum)
- 14:  $j = m$
- 15:  $i = n$
- 16: **while**  $j \neq 0$  **do** Set max and min to the first value of feature j
- 17:
- 18:     **while**  $i \neq 0$  **do**
- 19:         **if**  $v_{ji} < \min$  **then**
- 20:              $\min = v_{ji}$
- 21:              $i = i - 1$
- 22:         **else if**  $v_{ji} > \max$  **then**
- 23:              $\max = v_{ji}$
- 24:              $i = i - 1$
- 25:         **end if**
- 26:     **end while**
- 27:     Update each value  $v$  feature  $j$  as follows:

$$V = \frac{v - \min_A}{\max_A - \min_A} \quad (3.2)$$

- 28:      $j = j - 1$
  - 29: **end while**
  - 30: End
-

issues. It has been built by integrating three feature selection methods. Then, a step function is used to obtain preliminary three subsets of genes with their scores based on the three techniques used. In this sense, each gene may have three different scores according to the three techniques used. To obtain the best single score of a gene, fuzzification, and defuzzification methods were used. Lastly, the "step function" is used as the final stage to select the best subset of significant genes which have a high influence on classification as shown in Fig 3.4.

FGS has been divided into three key technical stages: voting, fuzzification, and defuzzification. The voting stage involves the selection of three preliminary subsets of genes. The subsequent stages, namely fuzzification, and defuzzification, aim to filter genes further and identify a subset of significant genes. The overall procedure for the FGS method, detailing the progression through these stages, is visually represented in Fig 3.4, providing a clear illustration of the method's workflow. These stages are described as follows:

### **Voting stage**

Three feature selection methods have been employed for selecting informative genes (MI, F-classify, and chi-squared). Each feature selection method selects a different number of genes based on the step function (SF). SF has been calculated in the equation 3.3. This formula is designed to avoid a limited number of selected genes which may be leading to ignoring some genes that have the same score when using the constant number of genes such as the top 10 genes. This formula also provides more flexibility to the SF value compared with constant score values such as 0.3 may non or small-selected features by a feature selection method have scored equal to 0.3 in this case, we lose some important genes that could be selected by other feature selection methods. The outcome of this method is three different lists for the three feature selection techniques ( the size of lists is different ) which means may first list come up with 10 genes and the second with 30 while the third comes with 23. Because of the flexibility of the SF used, there is an unknown number of genes are selected. The vote step is illustrated in algorithm 2.

$$SF = \max(FSS) * 0.3 \quad (3.3)$$

Where SF is the step function, FSS is the feature selection method's score for all genes. While max is the highest possible score for all genes assessed by each feature selection technique.

**Algorithm 2** Vote step

---

```

1: Input: Two-dimensional array DS(n,m) where n is the number of samples and m is the
   number of genes, the Step function is a  $\max(\text{score}) \cdot 0.3$ 
2: Output: Three vectors from the three filter methods
           Employing mutual information method to rank genes
3: Begin
4:  $f \leftarrow m$ 
5: while  $f \neq 0$  do
6:   Computing score of gene f using MI method
7:   if  $\text{score} < SF$  then
8:     Ignore this gene
9:      $f = f - 1$ 
10:  else if  $\text{score} \geq SF$  then
11:     $\text{SMI}[f] = \text{score}$ 
12:     $f = f - 1$ 
13:  end if
14: end while
           Employing F-classif method to rank genes
15:  $f \leftarrow m$ 
16: while  $f \neq 0$  do
17:   Computing score of gene f using F-classif method
18:   if  $\text{score} < SF$  then
19:     Ignore this gene
20:      $f = f - 1$ 
21:   else if  $\text{score} \geq SF$  then
22:      $\text{F-Classif}[f] = \text{score}$ 
23:      $f = f - 1$ 
24:   end if
25: end while
           Employing chi-squared method to rank genes
26:  $f \leftarrow m$ 
27: while  $f \neq 0$  do
28:   Computing score of gene f using the chi-squared method
29:   if  $\text{score} < SF$  then
30:     Ignore this gene
31:      $f = f - 1$ 
32:   else if  $\text{score} \geq SF$  then
33:      $\text{chi-squared}[f] = \text{score}$ 
34:      $f = f - 1$ 
35:   end if
36: end while
37:  $\text{Ds}(n,m) = (\text{SMI}[f], \text{F-chi-Classif}[f], \text{squared}[f])$ 
38: (Return three vectors of genes with three different scores for each gene)
39: End

```

---

### Fuzzification step

This is the process of converting crisp data into fuzzy data by using membership functions that aim to transform the crisp data into data ranging between (0-1). There are different types of membership functions. In this work, the Triangular Membership Function has been employed which is calculated in formal 3.4. The outcome of this step is all the scores of all genes for the three lists are ranging between (0-1) to unify the score of each gene in all lists that use for the next step. Mathematically illustrated as follows.

$$MF = \frac{W_i - a}{b - a} \quad (3.4)$$

Where MF is the membership function. W is the crisp value (score) for a gene. a = lowest possible score (min). b= highest possible score. This membership function applied for the three feature selection methods which means, there are MF1, MF2, and MF3 in this work.

### Defuzzification step

This step is a process for converting the output data to crisp data. This step is the final stage of the gene selection method that has been used to select informative genes. The selected genes from these steps have been used as identifiers for training the classifier approaches. This step aims to get one score for each gene by averaging the scores that have been selected by the three feature selection methods and proceed to the previous step to range the score between 0 and 1. The output is the best average score for each gene from the three lists that were obtained. Mathematically calculated as follows.

$$ASG = \frac{MF_i + MF_i + MF_i}{N} \quad (3.5)$$

Where ASG is the Average Score for a gene through the three feature selection methods. MF is the membership function for each gene. N is the number of feature selection methods that have been employed. In this work (N equals 3). From the two processes above, it can be inferred that fuzzification and defuzzification have been used to achieve the goal of having the best single score for each gene while filter feature selection approaches offer diverse scores for the same gene. As a result, employing the SF for deciding which genes are significant to use as markers for cancer classification, is illustrated in the equation below.

$$SF = \max(FSS) * 0.5 \quad (3.6)$$

Again, SF developed to be more flexible by getting the max score for the selected genes, therefore, multiply by 0.5. This calculation provides three major benefits: First, avoid getting null genes selected in all cases, while if constant SF is used may happen, for example, if the SF is 0.5 and all the genes score with 0.49. Second, avoid ignoring genes with the same score if it is compared to select top 10 genes may the eleventh gene has the same score as the tenth has while not be selected if the SF is with the top 10 genes.

Secondly, Fuzzy logic has been used a further analysing to select less number and more important genes by the developed fuzzy gene selection method. The developed FGS used Triangular Membership Function as fuzzification and centre of gravity as defuzzification with SF (illustrated in the defuzzification step ) to select informative genes that have a high impact on cancer classification as illustrated in algorithm 3.

In conclusion, developing fuzzy gene selection has provided some advantages over the fusion of three feature selection methods even though both techniques used a combination of feature selection approaches. FGS addressed the limitations that were identified when FTFS used. FTFS used a constant threshold to identify the top 150 genes for each feature selection method. As a result, some genes with the same score are ignored. Another disadvantage is that FTFS exclude some genes even if they received good scores in two feature selection methods since FTFS works by selecting only the genes that have been chosen in the three feature selection methods and eliminating others. In other words, it was based on genes chosen using the three feature selection approaches rather than gene scores. The FGS approach was used to overcome these two deficiencies. The first issue has been addressed by employing a step function rather than using a fixed threshold. The second constraint has been solved by employing a fuzzy approach to take into consideration the gene's score as well as the genes chosen by different feature selection approaches.

### 3.2.3 Fuzzy gene selection wrapper plus

Wrapper approaches are an inadequate option for analysing high-dimensional datasets since they are computationally expensive when applied to this type of data [21] [157]. It is mostly used when the number of genes is limited. Gene expression data is characterised by high dimensionality. To overcome these constraints, it integrated the principles of fuzzy gene selection and wrapper approaches in new methods well-known as FGSWP. FGSWP assumes that the genes selected by FGS are used as input for wrapper methods that employ backward elimination, which assumes that all the genes are used as input and are removed one by one while checking the accuracy, if removing a gene decreases the accuracy, it is kept, otherwise, the genes are ignored. This approach produces a list of genes that influence accuracy when they are eliminated. This method tries to reduce the number of genes selected

**Algorithm 3** Fuzzy Gene Selection Process

---

```

1: Input: Three vector of genes (chi-squared, F-Classif, and MI)
2: Output: Significant genes (SG) vector
3: Begin
    //Computing MF for each input(score) for MI method
4:  $i \leftarrow m$ 
5:  $j \leftarrow s$ 
6: while  $i \neq 0$  do
7:   while  $j \neq 0$  do
      
$$MF1[j] = \frac{X_i - a}{b - a} \quad (3.7)$$

8:      $j = j - 1$ 
9:   end while
10:   $i = i - 1$ 
11: end while
    //Computing MF for each input(score) for F-classif method
12:  $i \leftarrow m$ 
13:  $j \leftarrow s$ 
14: while  $i \neq 0$  do
15:   while  $j \neq 0$  do
      
$$MF2[j] = \frac{X_i - a}{b - a} \quad (3.8)$$

16:      $j = j - 1$ 
17:   end while
18:    $i = i - 1$ 
19: end while
    //Computing MF for each input(score) for chi-squared method
20:  $i \leftarrow m$ 
21:  $j \leftarrow s$ 
22: while  $i \neq 0$  do
23:   while  $j \neq 0$  do
      
$$MF3[j] = \frac{X_i - a}{b - a} \quad (3.9)$$

24:      $j = j - 1$ 
25:   end while
26:    $i = i - 1$ 
27: end while
    // Unification of scores using the ASG method
28:  $i \leftarrow m$ 
29: while  $i \neq 0$  do
      
$$ASG = \frac{MF_i + MF_i + MF_i}{N} \quad (3.10)$$

30:   if  $ASG \geq SF$  then
31:     SF[i]= Gene
32:      $i = i - 1$ 
33:   end if
34: end while
35: Return SG (significant genes)
36: End

```

▷ If  $ASG \geq SF$  select the gene

---

by the FGS method while maintaining the accuracy attained. Furthermore, it confirmed the FGS-selected genes and how they affect accuracy when they are eliminated. On the other hand, it verifies the FGS-selected genes and how they impact cancer classification. Most crucially, by introducing FGSWP, the benefits of filter techniques, such as speed, and wrapper methods, such as accuracy, have been attained while avoiding the shortcomings of both. The steps of building FGSW from input to output are shown in Fig 3.5.

### 3.2.4 Fuzzy classifier method

The main aim of developing FC is to enhance the accuracy of cancer classification and increase the generalisation of an algorithm to be accurate with all given datasets. FC assumes that applying three classifier algorithms (LR, SVM, and MLP) for a dataset then obtains the probability of predicting a class label for each classifier. Therefore, getting the max of the average for each class label for the three classifier approaches. As a result, the class label that has the highest max of the average from the three classifier methods will be chosen as the predicted class that can be used to compare with the actual class label, this process is called soft. Another process is named as majority which works if there are two classifier methods to predict the class label as A and the only one to predict the class label as B, the output will be class A because two out of three classifiers are predicted as A.

Based on that, FC developed depending on these two methods (soft and majority) to get the advantage from both. FC works by combining soft and majority methods to predict class labels. For instance, if the predicted class label when applying the soft method is A and the predicted class label when applying the majority method is B, in this scenario, FC applied a member function that takes into consideration the two methods (soft and majority methods) by adding 0.6 to the max average of the class label that has been selected by majority method and divided by two. Then comparing the output of this process with the max average, if the output is greater the max average will be selected as the predicted class, otherwise, the predicted class will be the same predicted by the max average (soft) method. The steps for how implementing the FC method are described in Fig 3.6.

## 3.3 Mitigate Overfitting

In order to mitigate overfitting, this thesis employs three strong methods. First and foremost, the foundational technique of cross-validation divides the dataset into several subgroups for training and validation. This approach ensures that the model is generalizable beyond the training set by facilitating a detailed evaluation of its performance across various partitions.

Second, feature selection becomes a critical tactic, carefully selecting the genes that are most informative while eliminating those that are unnecessary or superfluous. This method reduces the likelihood of overfitting and reduces model complexity by focusing on the most important features of the data. Lastly, by combining several models to provide predictions, ensemble learning techniques—in particular, ensemble methods—play a critical role in reducing overfitting. By combining the strengths of each distinct model, the results are strengthened and made more broadly applicable. When combined, these approaches provide a strong framework that addresses overfitting issues in gene expression data analysis and guarantees that machine learning models avoid overfitting traps while capturing significant trends.

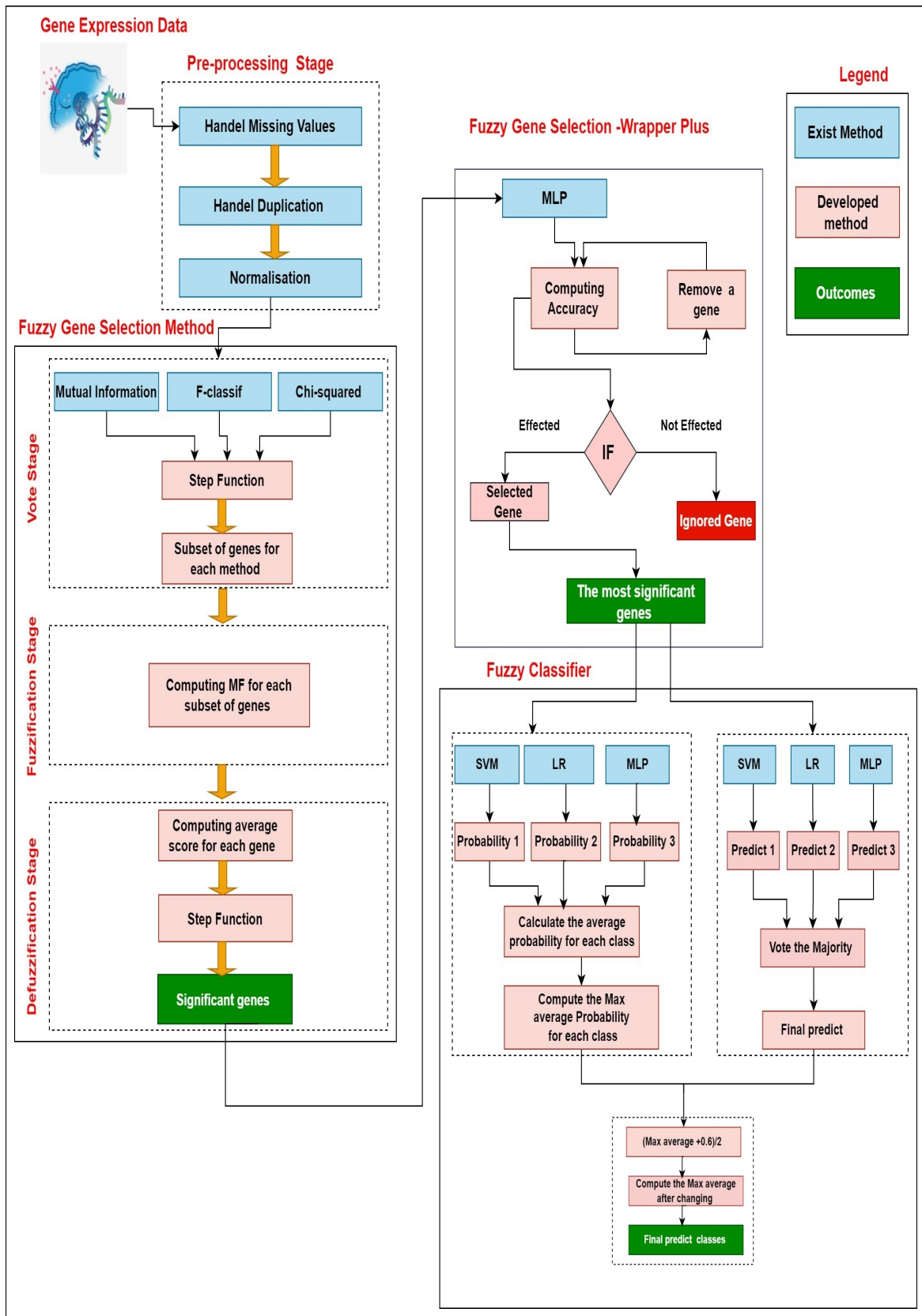


Fig. 3.2 The architecture of the developed model

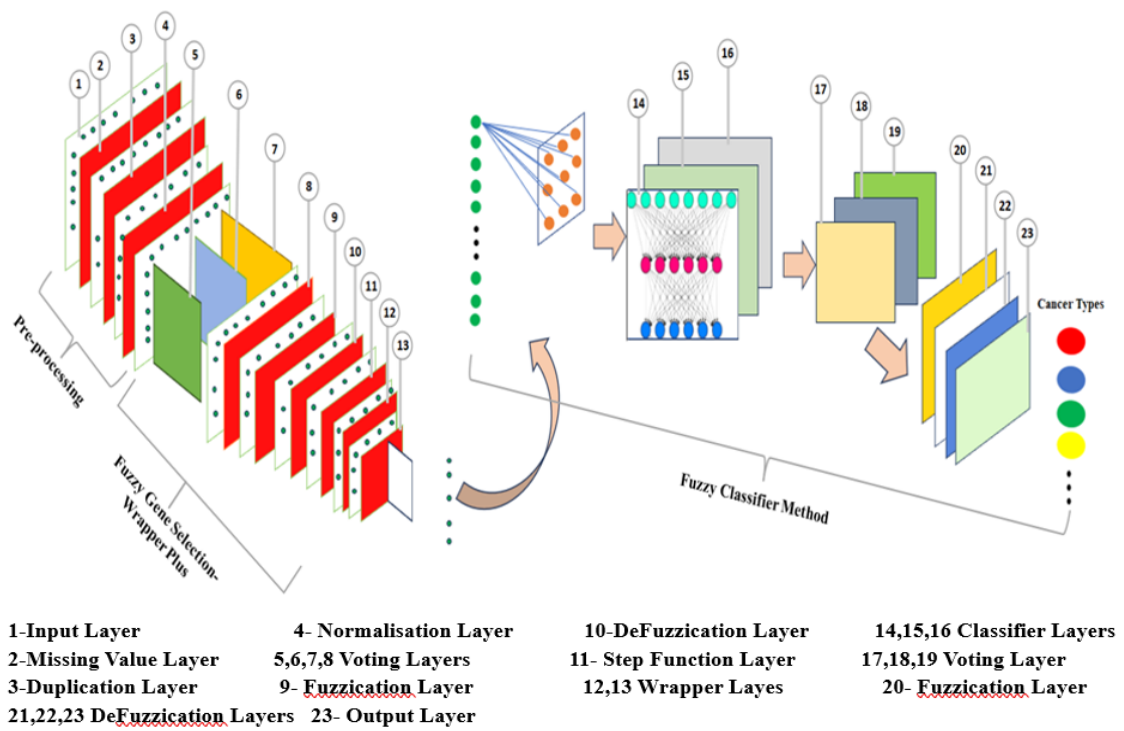


Fig. 3.3 The developed Topology

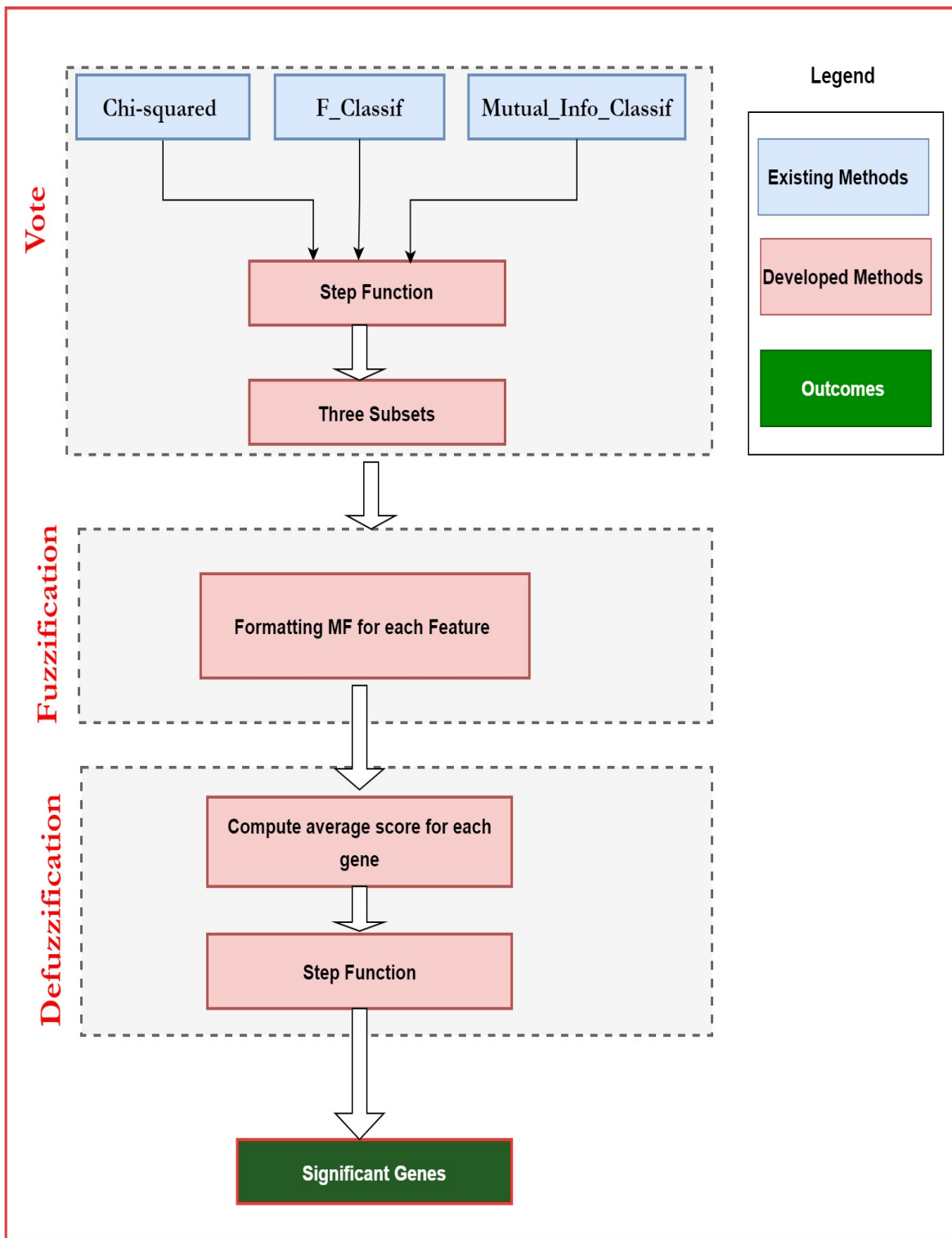


Fig. 3.4 Block Diagram of Developed Fuzzy Gene Selection Process

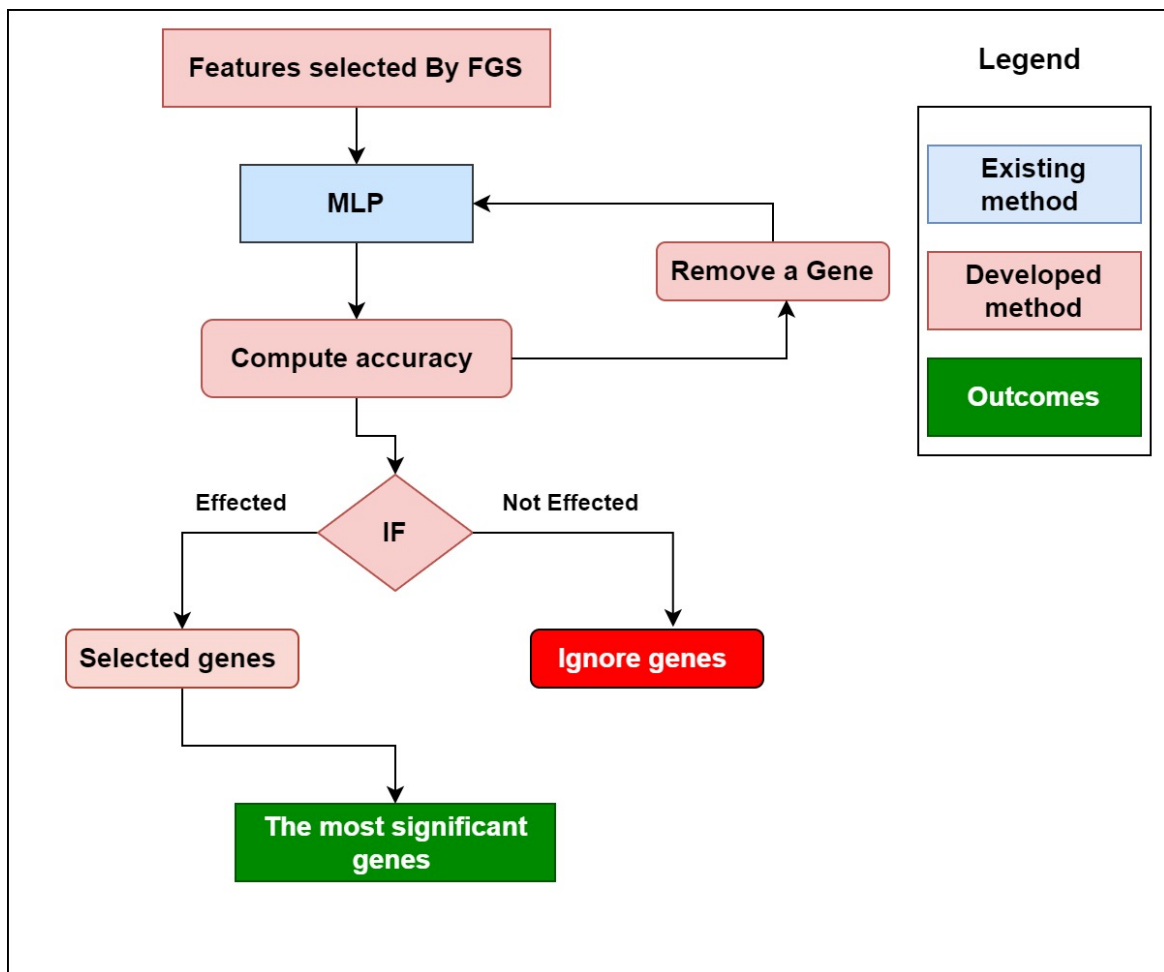


Fig. 3.5 An overview of the developed FGSWP

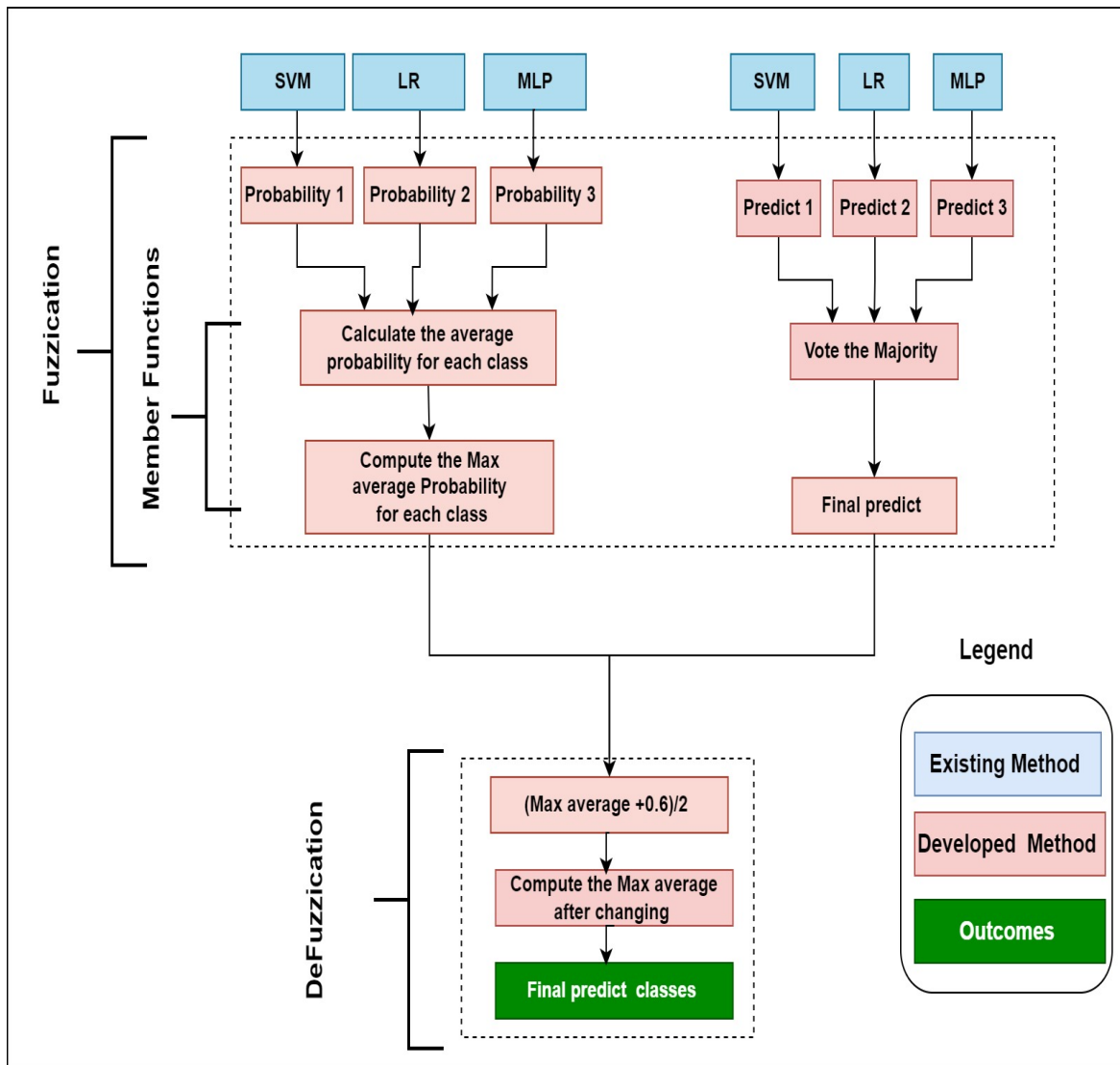


Fig. 3.6 An overview of developed FC method.

# Chapter 4

## Experimentation Framework

### 4.1 Introduction

This chapter aims to describe the experimentation framework process to evaluate the developed approaches (FTFS, FGS, FGSWP, and FC). Initially using classical classifiers without using any feature selection methods. Then using the FTFS method with classical classifiers. Also, FTFS were compared to previously published works as shown in Fig 4.1. Secondly, the developed approaches (FGS, FGSWP, and FC) compared into different categories as follows: 1) Apply classical classifiers on cancer expression data without using gene selection techniques. 2) Use classical classifiers with the FGS method. 3) Use classical classifiers with the FGSWP method. These three comparisons were used to ensure that the two developed methods (FGS and FGSWP) have an advantage in cancer classification. 4) Comparing FC against the classical classifiers when using FGS, and FGSWP. To show the effectiveness of developing FC against other classifiers when the same identifier genes are used for training the classifier approaches. Additionally, the developed model (MDFDL) was compared to prior studies to demonstrate that the developed model has superior to the previous studies that use the same datasets. The comparisons of the developed model against prior studies and classical classifier techniques in terms of the evaluation metrics and the number of genes selected to train a classifier.

The chapter also includes the details of the dataset employed to train and test the developed model. The experimentation framework process is described in Fig 4.2. The developed model has been implemented using Python programming language, using the Anaconda software distribution. The hardware setup includes an Intel Core i7-8565U processor and 32 GB of RAM.

## 4.2 Datasets employed for all developed approaches

Gene expression data from the microarray and RNA-Seq technologies was used to train and evaluate the proposed model. To evaluate the developed model's efficacy across different dataset sizes, both big and small datasets were employed. To test the developed model on both, the datasets include multi and binary class labels. Table 4.1 provides a complete breakdown of the datasets. A total of 11,687 sample data, distributed between binary and multiclass labels, were used for training, and testing the developed model in this thesis. To evaluate the correctness of the suggested approach in all forms of data, whether they are balanced or imbalanced class labels. Nineteen cancer expression data were used 11 As microarray and 8 as RNA-Seq gene expression data are described in detail in Table 4.1. The key details are shown in Table 4.1 datasets ID, measurement method (Microarray or RNA-Seq), number of samples, and gene number for each dataset. Moreover, the number of samples for each class for each dataset.

## 4.3 Comparing stage

In order to evaluate the performance of the developed model, it was compared against well-established classifier techniques commonly used in cancer classification studies. Additionally, the developed model was also benchmarked against previously published works that used the same datasets. The following comparisons were conducted to assess the effectiveness of the developed method.

- FTFS approach was employed in combination with classical classifiers. Initially, the classical classifiers were applied without any gene selection methods. Subsequently, the classical classifiers were applied in conjunction with FTFS. Finally, FTFS compared to prior studies to demonstrate the superior performance of the developed model by surpassing the results of prior studies.
- In the context of gene selection, the FGS approach was employed in combination with classical classifiers. Initially, the classical classifiers were applied without any gene selection methods to establish a baseline. Subsequently, the classical classifiers were applied in conjunction with FGS. Finally, the FGS technique was utilized with a fuzzy classifier. This different comparison to ensure the efficacy of gene selection approach in combination with classical and fuzzy classifiers.
- The fuzzy gene selection-wrapper plus technique was employed and compared to fuzzy gene selection FGS using classical classifiers (DT, SVM, KNN, GNB, and

Table 4.1 The full details of the datasets used to train and test the developed approaches.

Datasets	Method	No.samples	No.genes	No.class	No.sample for each Class
GSE14520	Microarray	445	13425	2	Liver cancer ( Cancer 227 , Normal 218)
GSE66499	Microarray	680	33298	2	Lung cancer (Cancer 490 , Normal 190)
GSE53757	Microarray	144	23516	2	Kidney Cancer (Cancer 72, Normal 72)
GSE10072	Microarray	107	13298	2	Lung cancer ( Adenocarcinoma 58 Normal 49)
GSE45827	Microarray	155	29873	6	Breast Cancer subtypes (Basal 41, Her2 30, Luminal B 30 Luminal A 29, CellLine 14 Normal 11)
GSE19804	Microarray	120	45782	2	Lung cancer (Cancer 60, Normal 60)
GSE33630	Microarray	105	23518	3	Thyroid Cancer (PTC 49, Normal 45, ATC 11)
GSE84437	Microarray	433	48710	4	Gastric cancer (T1 188,T2 132 T3 80,T4 33)
GSE13355	Microarray	180	23519	3	Skin Cancer (Normal 64, Involved skin 58, uninvolved skin 58)
GSE43580	Microarray	150	54675	2	Lung cancer ( Adenocarcinomas 77, Squamous Cell Carcinomas 73)
GSE75037	Microarray	166	16383	2	Lung cancer ( Adenocarcinomas 83, non-malignant 83)
GSE77314	RNA-seq	100	29087	2	Liver cancer (Cancer 50, Normal 50)
TCGA1	RNA-seq	2086	971	5	Five cancer types ( BRCA 878, KIRC 537, UCEC 269, LUSC 240, LUAD 162)
TCGA2	RNA-seq	964	20531	5	Breast Cancer subtypes ( LumA 431, LumB 195, Basal 143, Normal 128, Her2 67 )
TCGA3	RNA-seq	2974	56603	6	Six Cancer types (KIRC 538, Lung 533, LGG 511, HNSC 500, COAD 478, BLCA 414)
TCGA4	RNA-seq	1288	56603	9	Nine cancer types (CESC 304 KIRP 288, ESCA 161, GBM 156 LAML 151, ACC 79 ,KICH 65, DLBC 48, CHOL 36)
TCGA5	RNA-seq	592	56603	2	Lung cancer (Cancer 533, Normal 59)
TCGA6	RNA-seq	197	768	2	Liver cancer (Cancer 94, Normal 103 )
TCGA7	RNA-seq	801	20531	5	Five cancer types (BRCA 300 KIRC 146, COAD 78, LUAD 141, PRAD 136)

MLP). Initially, classical classifiers were used to compare FGSWP against FGS. This evaluation aimed to assess the performance of FGSWP in comparison to FGS when combined with classical classifiers. Additionally, a fuzzy classifier was employed to compare FGSWP against FGS.

- The fuzzy classifier was used and compared to classical classifiers employing FGS as well as FGSWP. The first comparison involved evaluating FC against classical classifiers using FGS, with the goal of evaluating the performance of FC in relation to classical approaches when combined with FGS. The second comparison concentrated on FC against classical classifiers employing FGSWP, aiming to identify the efficiency of FC compared to classical methods when using FGSWP. These comparisons provided insights into the performance and suitability of FC in the context of both FGS and FGSWP gene selection approaches.
- To ensure the superiority of the developed model (MDFDL), a thorough comparison was conducted against previously published works as presented in Table 5.7. This comparison aimed to demonstrate the superior performance of the developed model by surpassing the results achieved by prior research. By rigorously evaluating MDFDL in relation to existing methodologies, the study sought to establish its effectiveness and highlight its advancements over previous approaches.

## 4.4 Evaluation stage

### 4.4.1 A cross-validation

Cross-validation is a statistical technique that allows the classifier to train multiple times on the same dataset by dividing the dataset into multiple folds [158]. It can use a dataset for training and testing several times, increasing the classifier model's generalisation because all the cases have been trained. It is also regarded as one of the most successful methods for preventing or mitigating the overfitting issue and increasing the generalisation of a classifier [159] [160]. It also helps to gauge how well more accurately the algorithmic prediction is performed. Generally, there is no best option for the number of folds in cross-validation. It is based on different factors such as the size of the data, and the intended trade-off between bias and variance in the model evaluation. In this thesis, we used k cross-validation with  $k=5$  for both comparing classifier approaches and the developed fuzzy classifier method.  $K=5$  was used for different reasons illustrated as follows:

- **Sufficient Training and Testing Data:** Gene expression data is characterised by a small number of samples. In other words, the available gene expression dataset is small even if it has high dimensionality (Cross validation works on the number of samples).
- **Computational Efficiency:** Using 5 kfolds with gene expression dataset less computational time while still providing a reliable estimate of the model's performance.
- **Stability of Results:** Cross-validation estimates can have some variability due to the randomness in the partitioning of the data. Using a higher number of folds can increase this variability. With 5-fold cross-validation, the estimate tends to be reasonably stable while still providing a good approximation of the model's performance.

#### **4.4.2 Evaluation Performance**

Four evaluation metrics were employed to evaluate the performance of the developed model. These include (accuracy, precision, recall, and f1-score). Although achieving high accuracy for a model is crucial, it is not necessarily the best choice to evaluate how well a model performs [161]. Imagine developing a classifier algorithm to classify cancerous from normal. A model that predicts all patients are healthy will be 95% accurate if only 5% of patients have cancer. The same model that accurately predicts that all patients are healthy reaches 99% accuracy when just 1% of patients have cancer. Naturally, the "accuracy" is deceptive. Both of these models fail to detect any malignancies, making them ineffective for disease classification [162]. This is why additional metrics are required to weight different types of errors.

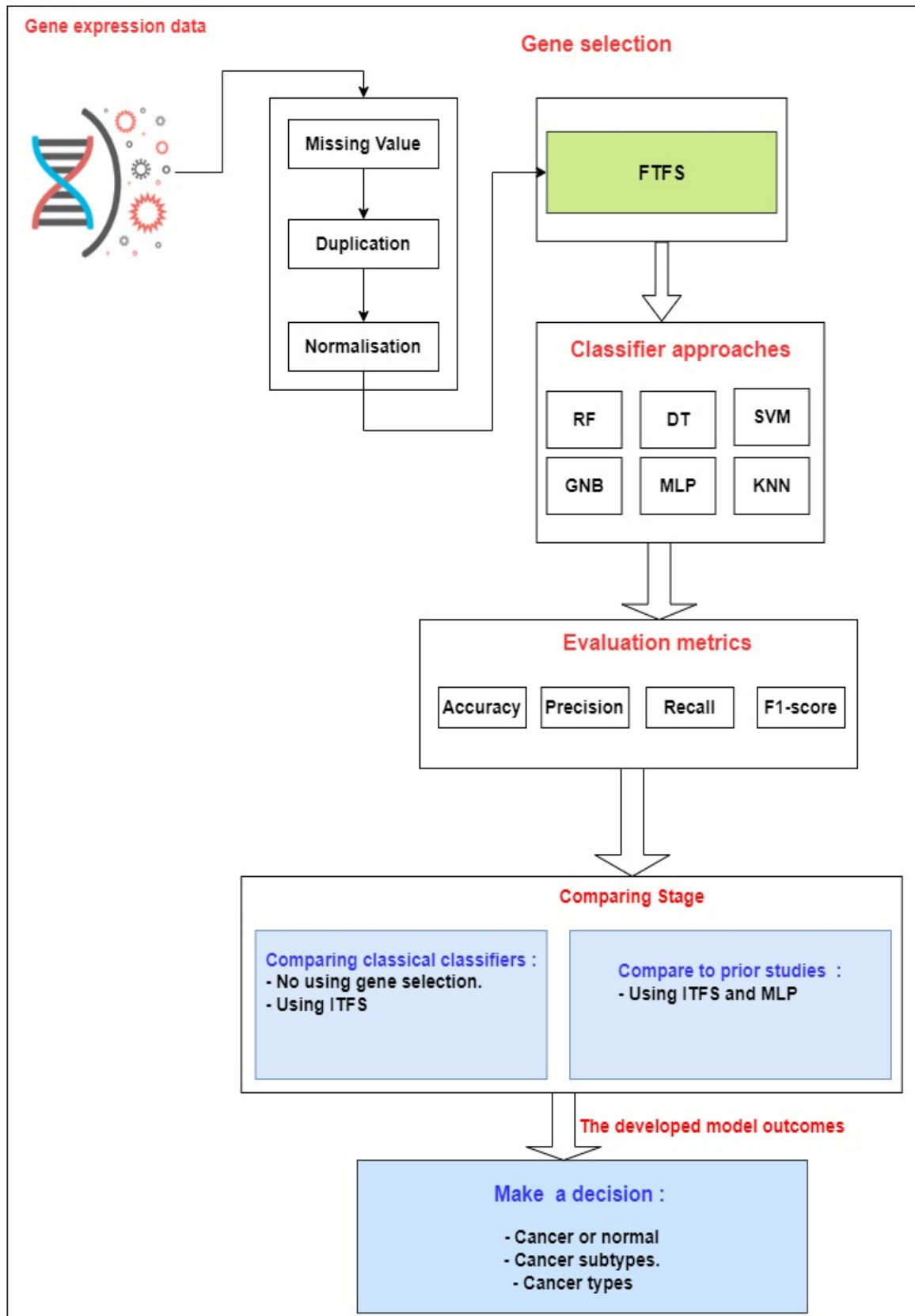


Fig. 4.1 Experimentation framework process of FTFS

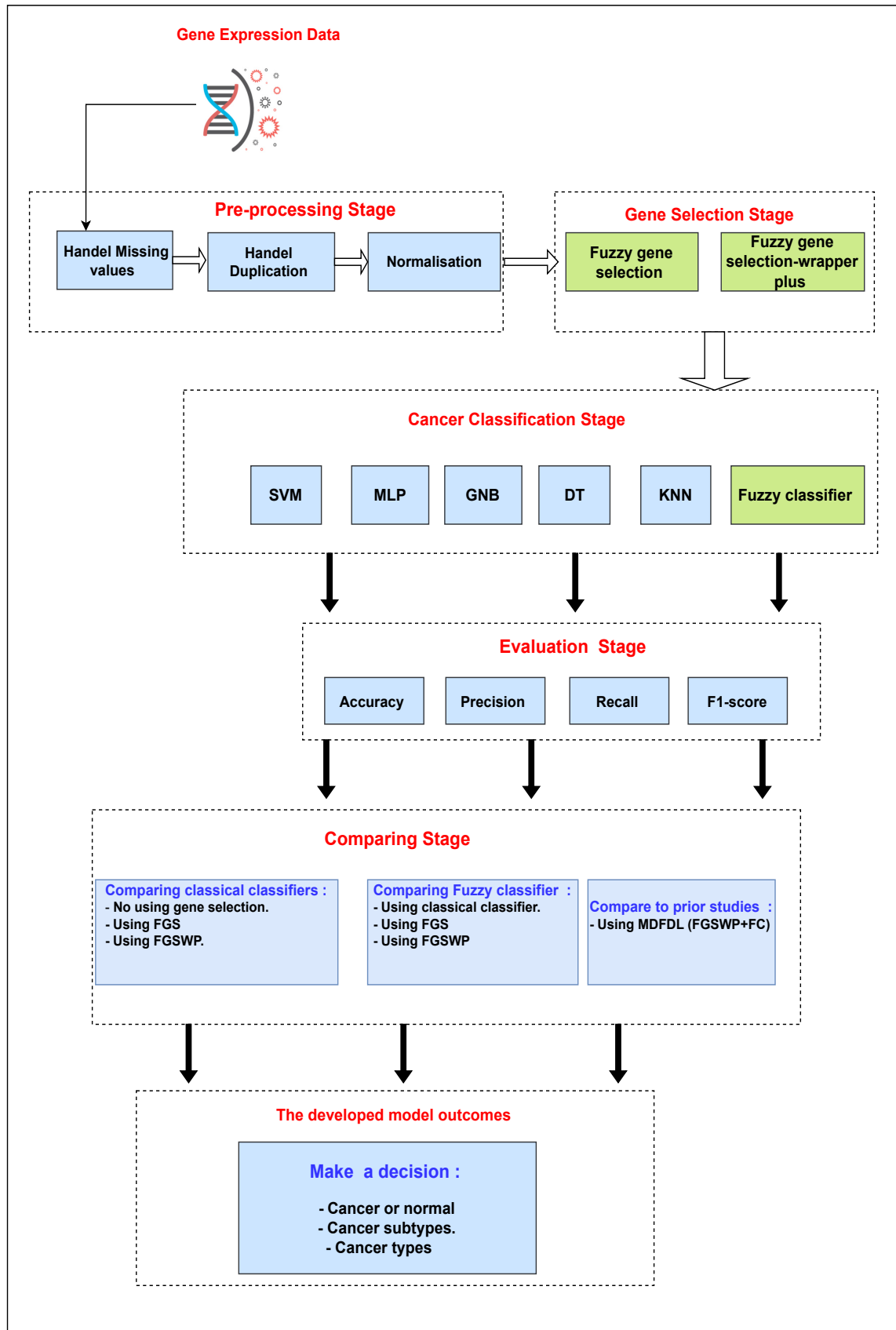


Fig. 4.2 Experimentation framework process of MDFDL

# Chapter 5

## Discussing Experimental Results

### 5.1 Experimentation of applying FTFS

#### 5.1.1 Datasets Employed to evaluate the FTFS Method.

Six gene expression datasets (n=6) were utilised to train and evaluate the proposed model, which was then compared to other classifier approaches. Microarray and RNA-Seq data were employed in the approach. The datasets (GSE45827, GSE14520, GSE77314, GSE19804, TCGA1 and GSE13355) were downloaded from TCGA and GEO. The datasets used contain both binary and multi-class labels, as well as a range of cohort sample sizes, to ensure that the proposed model can effectively classify cancer using both approaches.

#### 5.1.2 The results achieved with FTFS

To reduce the number of genes required for each classifier, FTFS was applied to six expression datasets. This method was evaluated by using a cross-validation method with k=5 folds. Table 5.1 provides a detailed description of the datasets used and evaluation metrics to evaluate the classifier's performance. The findings demonstrated that employing FTFS has a good impact on classifier approaches, notably the MLP classifier. FTFS approach enhanced the performance of all six classifiers tested across six datasets. Furthermore, the FTFS technique has reduced the number of genes required to train the classifier algorithms. As a result, this method contributed to minimise the complexity of the classifier and the time taken during training. The highest average improvement of accuracy in the six datasets was when MLP and FTFS employed together compared to employing MLP individually. The outcomes achieved accuracy between (95.5% to 98%) for the six datasets when FTFS used. The findings demonstrated that FTFS has highly improved the accuracy of four datasets

(GSE19804, GSE13355, GSE77314, and GSE77314) while slightly enhancing these datasets (TCGA and GSE14520). Although, FTFS significantly reduced the number of genes in all employed datasets from thousands to only tens. However, for some datasets, such as (GSE19804, GSE13355, and GSE14520), there is potential for improvement in lowering the number of selected genes.

### 5.1.3 Comparing FTFS to prior studies

Overall, the improvement of using FTFS with MLP outperformed other classifiers. Based on that, FTFS and MLP were compared to prior studies. The developed model ( FTFS and MLP) was compared to previous publications that used the same datasets. FTFS showcased its effectiveness by significantly reducing the number of genes while improving accuracy and other evaluation metrics such as precision, recall, and f-score. The results of this comparison are presented in Table 5.2.

The results of the developed model, which used the same datasets for liver cancer (GSE14520), demonstrated superior performance compared to the study referenced as [137] that used DRE-DNN. The developed model achieved results with 96% accuracy, across all evaluation metrics. In contrast, the prior study achieved lower scores with 82%, 83.3%, 95%, and 88.9% for accuracy, precision, recall, and f1-score respectively. Furthermore, an interesting observation is that the number of selected genes decreased significantly from 1253 in the study referenced as [137] to 97 in the developed model. This reduction in the number of selected genes indicates that the developed model was able to identify a much smaller subset of genes that are most relevant to liver cancer classification. This reduction in the number of genes can have practical implications such as reducing computational complexity.

The findings of the developed model, which used the same datasets for lung cancer (GSE19804), demonstrated superior performance compared to the study referenced as [163] that used HLR-SVM. The developed model achieved results with 97% accuracy and other evaluation metrics. In contrast, the previous study achieved lower scores with 94% accuracy with no mention of other metrics such as precision-recall or f1-score.

Although the developed model achieved comparable results to the study referenced as [164], which used the BPSO-DT-CNN approach to classify five cancer types (TCGA1), there was a significant reduction in the number of genes used by the developed model. The study [164] used 971 genes for classification, whereas the developed model effectively reduced this number to only 76 genes. This reduction in the number of genes is noteworthy as it indicates that the developed model was able to identify a smaller subset of genes that are highly relevant for accurate cancer classification. The ability to achieve similar performance

Table 5.1 Evaluate the effectiveness of using the FTFS method with six classifiers on six cancer datasets.

Dataset	Gene Selection	N-Genes	Classifier	Ac %	Pre %	Rec %	F1 %
GSE19804	No	45782	DT	91.6	90	92.9	91
			KNN	91.6	85	100	91.8
			RF	92	95	94	94.5
			SVM	94.4	94	94	94
			MLP	52	25	50	34
			GNB	97	94	100	97
GSE19804	FTFS	99	DT	95	100	96.4	97
			KNN	95	95	95	95
			RF	94	96	95	96
			SVM	96	96	96	96
			MLP	97	100	94	96.9
			GNB	98	98	98	98
GSE77314	No	29087	DT	92	95	90.5	92.6
			KNN	93.3	89.4	100	94.4
			RF	96.6	94.4	100	97
			SVM	96.6	94.4	100	97
			MLP	86.6	100	76.4	86.6
			GNB	73.3	100	52.9	69
GSE77314	FTFS	34	DT	95.3	93.5	98.8	96
			KNN	96.6	94.4	100	97
			RF	96.6	94.4	100	97
			SVM	96.6	94.4	100	97
			MLP	97.6	95.4	100	98.5
			GNB	96.6	94	100	97
GSE45827	No	29873	DT	82.5	81.5	79.5	78
			KNN	78.7	79	80.9	77.8
			RF	93.6	94.5	93.9	93.5
			SVM	93.6	93	93.9	93.5
			MLP	29.7	5	16.6	7
			GNB	87	89	86	87
GSE45827	FTFS	31	DT	91.9	91.4	92.8	91.4
			KNN	90	90	100	95
			RF	95	94.5	94.3	94.4
			SVM	95	95	95	95
			MLP	98	98	98	98
			GNB	97.8	98	98.8	98.4

Dataset	Gene Selection	N-Genes	Classifier	Ac %	Pre %	Rec %	F1 %
GSE14520	No	13425	DT	91.9	91	93.7	92.3
			KNN	92.5	89.4	97	93
			RF	95.5	95	96	95.7
			SVM	96	95.7	97	96.4
			MLP	95.5	94.4	97	95.7
			GNB	94	93	95.7	94.3
GSE14520	FTFS	97	DT	92.9	93.6	92.8	93
			KNN	95.5	96	95	96
			RF	95.5	94.4	97	95.7
			SVM	93	91.7	95.7	93.7
			MLP	95.5	97	97	95.7
			GNB	94.7	93	97	95
GSE13355	No	23518	DT	76.6	77.5	76.6	76.8
			KNN	79.6	79.6	79.6	79.6
			RF	85.5	85.6	85	85
			SVM	87	87.7	87	87
			MLP	83.3	83	83	83
			GNB	85	85	85	85
GSE13355	FTFS	91	DT	97	97	97	97
			KNN	96	96.5	96	96
			RF	96	96.4	96.4	96
			SVM	98	98	97	97.5
			MLP	98	98	98	98
			GNB	88.8	88.8	88.8	88.8
TCGA1	No	971	DT	92.5	89	87.8	88.4
			KNN	92.3	89	87.5	88
			RF	93.6	94.5	93.9	93.5
			SVM	95.6	92.6	92.4	92.6
			MLP	95.8	93.5	93	93
			GNB	94.5	91.7	93.5	92.5
TCGA1	FTFS	76	DT	92.5	90.8	87.9	89
			KNN	95	92.5	93	92.6
			RF	98	98.5	98.3	98.3
			SVM	98	96.4	96.8	96.6
			MLP	97.9	96.7	96.4	96.6
			GNB	94.5	90.5	92.3	91.3

with a significantly reduced gene set is advantageous in multiple ways. It facilitates the analysis process, mitigates overfitting, and reduces computational complexity.

The results of the developed model demonstrated superior performance compared to two published works referenced as [165] and [166], which were used to classify breast cancer subtypes (GSE45827). The developed model achieved perfect scores of 100% for all evaluation metrics, using only 31 genes. In contrast, the study referenced as [118], which employed the CFS-NB approach, accomplished scores of 89.7%, 92%, 89.7%, and 90% for accuracy, precision, recall, and f1-score, respectively, with only 20 genes. Additionally, the study referenced as [119], which utilized the Rough set-KNN method, achieved scores of 96.86%, 96.9%, 97.34%, and 97.8% for accuracy, precision, recall, and f1-score, respectively, with 38 genes. The developed model outperformed both [165] and [166] in terms of achieving perfect scores for all evaluation metrics, demonstrating its superior ability to accurately classify breast cancer subtypes. Furthermore, despite using a slightly higher number of genes compared to [165] and lower [166], the developed model still achieved exceptional performance. These results highlight the effectiveness of the developed model in achieving high accuracy and comprehensive evaluation metrics, showcasing its potential as a valuable tool for breast cancer subtypes classification.

Table 5.2 Comparing the FTFS method to prior studies

Dataset	Approaches	N-Genes	AC %	Pre %	Rec %	F1 %	Reference
GSE14520	DRE+DNN	1253	82	83.3	95	88.9	[137]
	FTFS+ MLP	97	96	96	96	96	The model
GSE19804	HLR+SVM	10	94	No	No	No	[163]
	FTFS+MLP	99	97	97	97	97	The model
TCGA1	BPSO-DT+CNN	971	96	94.96	95	95	[164]
	FTFS+MLP	76	95	95	94	95	The model
GSE45827	CFS+NB	20	89.7	92	89.7	90	[165]
	Rough set+KNN	38	96.86	96.9	97.34	97.8	[166]
	FTFS+MLP	31	98	97	98	97.6	The model

## 5.2 Experimentation of applying FGS

### 5.2.1 The datasets used to examine FGS efficiency

Sixteen cancer gene expression datasets were used for training and testing the developed FGS. The datasets comprised RNA-seq and Microarray data. The datasets were downloaded from TCGA and GEO (GSE45827, GSE14520, GSE77314, GSE19804, TCGA1, GSE33630,

TCGA2 , TCGA4, GSE53757, TCGA7, GSE10072, GSE43580, TCGA6, GSE75037, GSE66499, and GSE84437). These datasets include multi and binary classes more details were described in Table 4.1. To avoid overfitting issues of the algorithm, a cross-validation method was used with  $k=5$  to split the datasets into multiple folds and train the algorithm on these different folds.

### 5.2.2 The results achieved with FGS

To reduce the number of genes required for each classifier, FGS was employed on sixteen datasets. The comparison of the use and omitting of the FGS was presented in Table 5.3. illustrates the results obtained when using and omitting the fuzzy gene selection strategy to five commonly used classifier algorithms. The results of the study revealed that the FGS technique exhibited a significant reducing the number of genes across all datasets analysed. Moreover, the FGS approach yielded a noteworthy enhancement in cancer classification accuracy for certain datasets, while also displaying an improvement, albeit to a lesser extent, in accuracy for other datasets.

In summary, the use of FGS in cancer expression data showed that it enhanced the accuracy, precision, recall, and f1-score of well-known classifiers. Results indicated that FGS showed superior in thirteen out of sixteen employed datasets in improving the performance of classical classifiers. However, only four of these datasets were not improved even though the number of genes was highly reduced which helps to reduce the complexity of the classifiers and mitigate overfitting issues. The FGS method improved the performance of all classifiers applied, notably with MLP. The findings demonstrate that there is no one classifier algorithm that consistently achieves the highest results across all data. For example, GNB had the highest results for kidney cancer data (GSE53757) while, MLP accomplished the highest results for lung cancer (GSE19804), and five cancer types (TCGA1). Additionally, GNB and KNN had the highest results for liver cancer (GSE14520) while SVM and KNN achieved the highest results for thyroid cancer (GSE33630). The FGS method with classical classifiers achieved poor results with these datasets (GSE84437, GSE43580, GSE66499, and TCGA6) even though it reduced the number of informative genes. The key issue with these datasets is that they are not only unable to enhance accuracy when FGS used, but the achievement results were also very poor. With some of the datasets, the accuracy was 37%. Accordingly, it was necessary to develop a new approach capable of solving these two challenges, namely generalising the classifier so that it can provide better accuracy for all data. The thesis's further work was meant to address these challenges, and the Fuzzy Classifier Method was developed for this purpose. Notably, good accuracy does not always indicate that a classifier is the best one, instead, the classifier's performance must be evaluated in terms of precision,

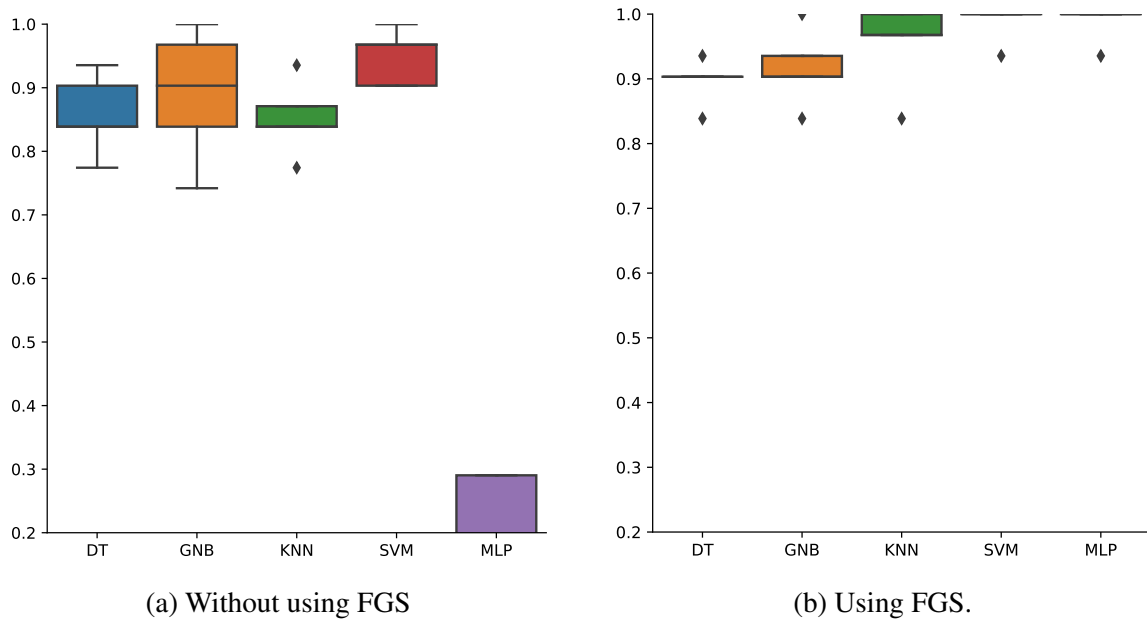


Fig. 5.1 Comparing accuracy score when using and omitting FGS (GSE45827)

recall, and f1-score. For example, KNN achieved 91% accuracy , 87.7% precision , 86.5% recall, and 86% f1-score while GNB 90% accuracy , 93.7%, 89.7%, and 90 f1-score for thyroid cancer (GSE33630). Even though, KNN had higher accuracy whereas GNB evaluated is better because the other evaluation metrics are better for GNB.

### 5.2.3 Discuss FGS results

To show the differences between the results obtained by omitting and employing the FGS technique with the five different classifier techniques, the accuracy scores with  $k=5$  have been displayed on a bar chart. Fig. 5.1 demonstrates the  $k=5$  difference in accuracy ratings between using and ignoring FGS. The results demonstrate how the usage of FGS enhanced classifier algorithms' performance, notably with the MLP classifier. The FGS method was also used to reduce the number of selected genes from 29873 to 68 genes. These results showed that the development of the FGS technique contributed to an improvement in accuracy, a reduction in the training time for models, and the provision of early cancer detection by the choice of instructive genes. Classifier models are also less complicated.

When a fuzzy gene selection method was used, as shown in Fig. 5.2 the performance of the five classifier approaches for lung cancer classification significantly improved. In comparison to other classifier models, the findings show that the MLP model offers predictions that are closer to the ideal observed value. MLP earned an average accuracy score of

Table 5.3 Evaluate the effectiveness of using FGS on multiple classifier approaches across sixteen cancer expression datasets.

Dataset	Gene Selection	N-Genes	Classifier	Ac %	Pre %	Rec %	F1 %
GSE14520	No	13425	DT	90	90.6	88.9	89.7
			KNN	94	91	97.6	94
			SVM	97	96	97.6	97
			GNB	95	95.6	94	94.8
			MLP	86.7	76.5	76.7	76.5
GSE14520	FGS	23	DT	96	95	97	96
			KNN	96.6	96	97	96.6
			SVM	96	95.6	96	96
			GNB	96.6	96	97	96.6
			MLP	97	97	97	97
GSE33630	No	23516	DT	87.6	77.6	81	79
			KNN	91	87.7	86.5	86
			SVM	93	95	92	92
			GNB	90	93.7	89.7	90
			MLP	72	55.6	64.5	58.5
GSE33630	FGS	76	DT	93	93	93.5	92.5
			KNN	94	96	92.8	93
			SVM	93	94	92.8	92
			GNB	92	88	99.8	88.8
			MLP	93	95	92	92.5
TCGA1	No	971	DT	91	87	85	85.8
			KNN	88	83	81.5	81.9
			SVM	95	91.6	91.8	91.6
			GNB	94	89.7	92	90.7
			MLP	94	90.8	89.8	90
TCGA1	FGS	25	DT	91.7	88	87	86.5
			KNN	93.6	89.8	90	89.6
			SVM	94	90.5	90.7	90.5
			GNB	92	87.7	90.8	89
			MLP	95	92	91.6	91.6
GSE19804	No	45782	DT	89	90	88	90
			KNN	90.8	88	95	91
			SVM	95.8	96.6	95	95.7
			GNB	92.5	95	90	91.9
			MLP	50	20	40	26.6
GSE19804	FGS	36	DT	92.5	93.6	91.6	92
			KNN	96.6	96.7	96.6	96.6
			SVM	96.6	97	96.6	96.6
			GNB	95.8	96.7	95	95.7
			MLP	97.5	97	98	97.5

Dataset	Gene Selection	N-Genes	Classifier	Ac %	Pre %	Rec %	F1 %
GSE43580	No	54675	DT	85.3	87.3	83.4	84.7
			GNB	84.6	91	75	82
			KNN	79.3	73	93	81.5
			SVM	83.3	84.3	80.6	82.3
			MLP	86	86.3	84.7	85
GSE43580	FGS	28	DT	80	80.4	80.8	79.8
			GNB	84.6	92.8	75	82.3
			KNN	83.3	91.9	72.5	80.4
			SVM	86.6	98	73.8	83.8
			MLP	78	77.3	78	77.6
GSE45827	No	29873	DT	85.8	83	82.6	81.5
			KNN	85	87.9	87.7	87
			SVM	94.8	96	95.8	95.8
			GNB	89	92.7	88.8	89
			MLP	20.6	6	17	7
GSE45827	FGS	68	DT	89.6	90.9	89.6	88.8
			KNN	95	96.5	96	96
			SVM	97.4	98	97.66	97.75
			GNB	91.6	94.5	92	92.8
			MLP	98	98.8	98	98.3
TCGA4	No	56603	DT	82.5	66.8	63.9	60.8
			GNB	95.4	95	94.8	94.7
			KNN	92.3	94	87.6	89.4
			SVM	98	97.8	96.7	97
			MLP	98	97.5	96.9	97
TCGA4	FGS	298	DT	88	77	73.7	70.6
			GNB	97.7	97.6	95.6	96.4
			KNN	98.5	97.7	96.7	97
			SVM	98.7	98.5	97.3	97.8
			MLP	98.6	98	97.5	97.7
TCGA2	No	20531	DT	83.7	83.4	80	81.3
			GNB	77	77.6	77.7	77
			KNN	73	75.9	64.5	66.3
			SVM	82.7	82	82.5	82
			MLP	84.4	84.4	83	82
TCGA2	FGS	116	DT	80.7	76.5	74.3	74.8
			GNB	81	76.4	79.5	77.5
			KNN	83	84	75	77.6
			SVM	86.3	87	82.6	84.3
			MLP	86.3	86	84.8	85
GSE53757	No	23516	DT	95.7	94.6	97	95.8
			GNB	95	95.6	94	94.9
			KNN	95.7	95.6	95.7	95.6
			SVM	95	94.8	95.9	95
			MLP	93.7	98.5	89	93.3

Dataset	Gene Selection	N-Genes	Classifier	Ac %	Pre %	Rec %	F1 %
GSE53757	FGS	78	DT	95	92.4	98.5	95.3
			GNB	97.8	97	98.5	97.8
			KNN	97.8	97	98.5	97.8
			SVM	97	96	98.5	97
			MLP	97	97	97	97
TCGA6	No	768	DT	58	55.8	58.4	56.8
			GNB	57.3	54.4	69	60.5
			KNN	59.8	56.8	68	61.3
			SVM	62.8	63.3	56.3	59
			MLP	62.8	63	57.5	59
TCGA6	FGS	28	DT	58.3	55.9	59.8	56
			GNB	66.5	61	84	70.5
			KNN	63.9	59.8	75.7	66.4
			SVM	67.48	64	75.6	68.6
			MLP	65.4	63.8	64	63
TCGA7	No	20531	DT	96.7	96.6	96.9	96.7
			GNB	77.5	82.6	71.5	72.5
			KNN	99.7	99.7	99.6	99.6
			SVM	99.8	99.9	99.8	99.8
			MLP	95.8	93.8	95	94.4
TCGA7	FGS	11	DT	97.5	96.3	97.4	96.7
			GNB	98.6	97.9	98.4	98
			KNN	98.8	98	98.6	98.3
			SVM	99	98.3	98.7	98.4
			MLP	99	98.7	98.8	98.7
GSE77314	No	29087	DT	95	98	91.9	94
			KNN	88.9	82	100	90
			SVM	99	98	100	99
			GNB	84	100	68	80
			MLP	93	98	88	91
GSE77314	FGS	12	DT	97	98	96	97
			KNN	98	98	100	99
			SVM	98	97	100	98
			GNB	97	98	96	96.8
			MLP	99	98	100	99
GSE10072	No	13298	DT	94.3	93.5	96	94.3
			GNB	98	100	95.7	97.7
			KNN	97	95.3	100	97.3
			SVM	99	100	98	98.9
			MLP	50.5	18	40	25
GSE10072	FGS	52	DT	93.4	93.5	93.7	93
			GNB	98	100	96	97.8
			KNN	95.3	94.8	96	95
			SVM	98	100	96	97.8
			MLP	97	100	94	96.7

Dataset	Gene Selection	N-Genes	Classifier	Ac %	Pre %	Rec %	F1 %
GSE75037	No	16383	DT	96.3	95.7	97.5	96.3
			GNB	95.7	100	91.3	95
			KNN	94	94	95	94.4
			SVM	95	95	94	94.5
			MLP	60.9	50.6	80	60.5
GSE75037	FGS	36	DT	99.4	100	98.7	99.3
			GNB	98.8	98.8	98.7	98.7
			KNN	98	98	98	98
			SVM	98.7	100	97	98.5
			MLP	99.4	100	98.8	99.4
GSE66499	No	33298	DT	66	38.5	32	34.6
			GNB	52.3	38	50	27.3
			KNN	67.7	42.4	17.8	19.7
			SVM	72.7	62	28.9	37
			MLP	68	59.3	20.5	19.3
GSE66499	FGS	150	DT	70.4	49.9	24.7	31.4
			GNB	64.7	53.4	56.8	44.6
			KNN	69	56	26.3	31.4
			SVM	75	69	36.8	44.5
			MLP	72	53.3	37.3	42.6
GSE84437	No	48710	DT	33	19.9	23	19.7
			GNB	24.7	28.5	24.4	20.6
			KNN	35.3	19.7	23.7	19
			SVM	32.3	18.3	23.9	18.6
			MLP	27	9.5	23.4	12.6
GSE84437	FGS	105	DT	35.3	22.4	24.4	21
			GNB	31	34.4	42	29.4
			KNN	36.9	28.5	28.3	25.6
			SVM	37.8	33	29.7	26
			MLP	35	32.3	32.4	29.4

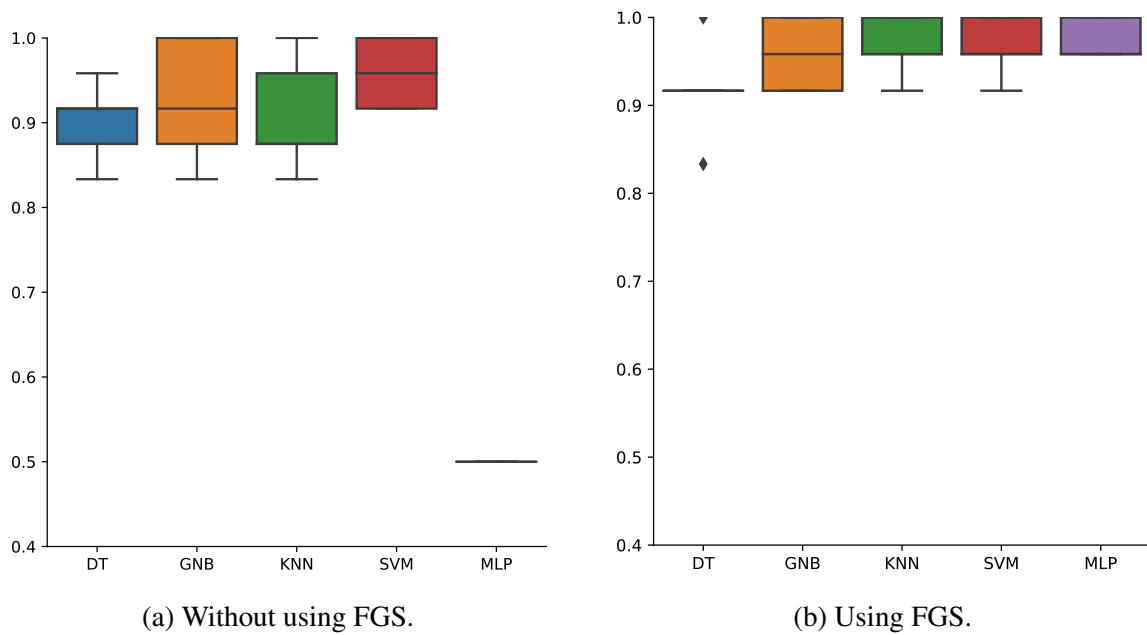


Fig. 5.2 Comparing accuracy score when using and omitting FGS (GSE19804).

97.5% with  $k=5$ . Other classifiers, achieved average scores of 96.6%, 96.6%, 95.8%, and 92.5% with  $k=5$  for SVM, KNN, GNB, and DT, respectively. Additionally, only 36 genes out of 45782 genes were employed for training the classifier models, a considerable decrease in the number of genes used. Although there is a slight improvement in the accuracy of most of the classifiers used in this investigation to classify liver cancer datasets (GSE14520). However, there is a remarkable enhancement in the MLP classifier using the FGS method, as it improved from 86.6 to 96 as an average accuracy score with  $k=5$ . More importantly, the FGS method reduced the number of genes used to train models to 23 only out of 13425. This reduces classification complexity, shortens training time, and reduces the problem of overfitting. Fig. 5.3 explains the comparison accuracy scores with  $k=5$  for the five models when FGS employed and omitted.

Most classifier approaches achieved good results. The average accuracy score with  $k=5$  was 99% for the SVM, KNN, and MLP while 97% for GNB and DT when the FGS technique was applied to the liver cancer dataset (GSE77314). These remarkable enhancements in accuracy score are shown in Fig. 5.4. Moreover, the FGS method decreased the number of genes from 29087 to only 12 genes that were used as identifiers for training classifier approaches. That leads to an increase in the model efficiency and mitigates the time taken through algorithm training, reduces the complexity of the classifiers, and provides early cancer detection.

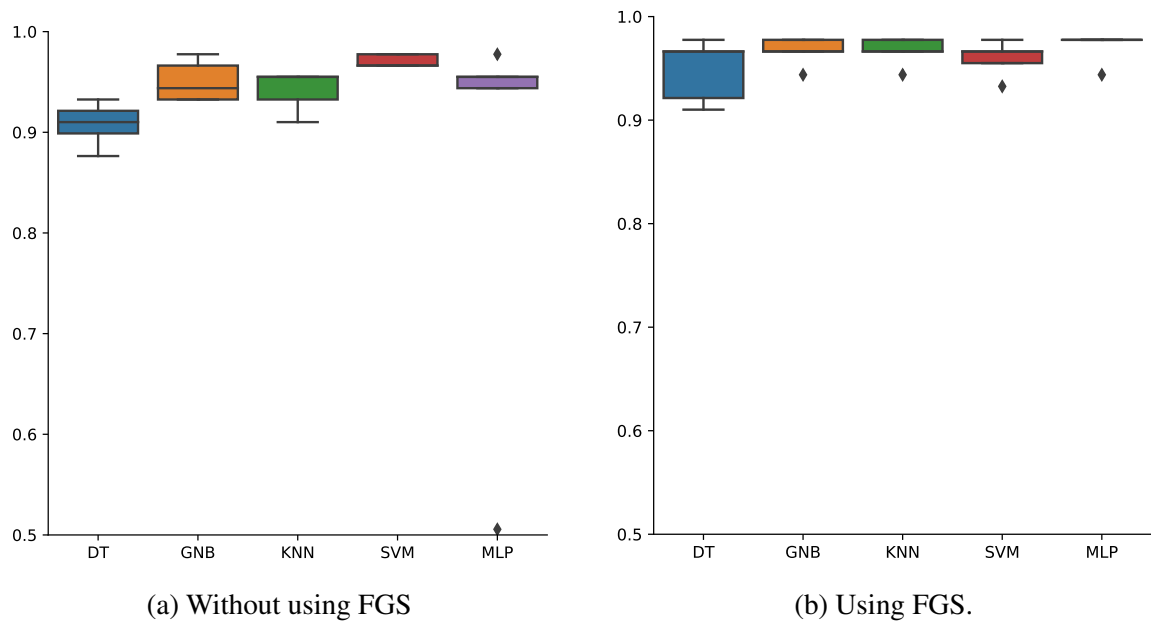


Fig. 5.3 Comparing accuracy score when using and omitting FGS (GSE14520)

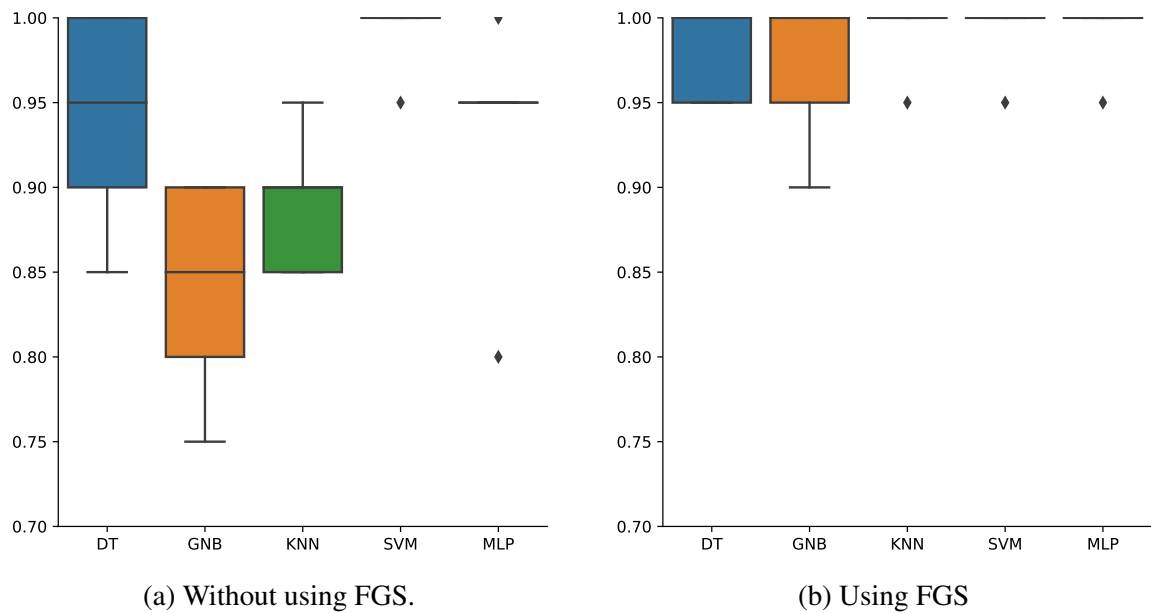


Fig. 5.4 Comparing accuracy score when using and omitting FGS (GSE77314)

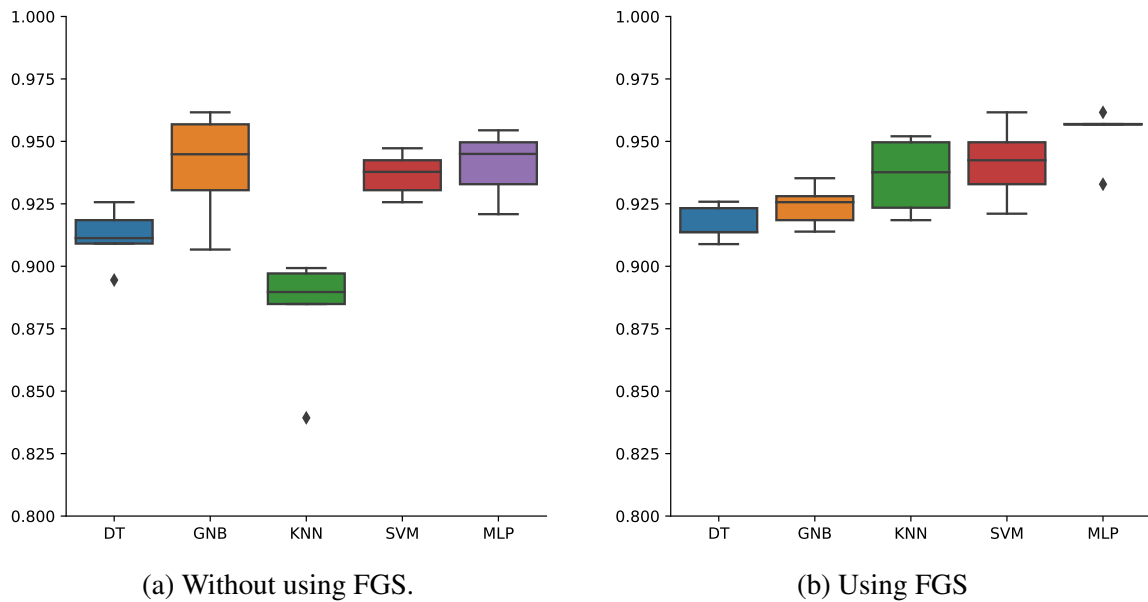


Fig. 5.5 Comparing accuracy score when using and omitting FGS (TCGA1)

There was not a significant improvement in (TCGA1) dataset because the number of genes used was not large (971), so its use did not achieve a high level of accuracy improvement. However, it improved the performance of the model by reducing the number of selected genes that were used as identifiers to train the classifier technique. As a result, the FGS method decreased the number of genes from 971 to 25 genes only. In addition, a slight improvement in the accuracy as well as the precision. It can conclude that employing FGS in the worst cases will give better accuracy and fewer genes and perform less time for training the classifier models and provides early detection of cancer. Fig. 5.5 illustrates the difference between the accuracy scores in  $k=5$  when the classifiers were applied to the datasets with and omitting FGS.

Good enhancement is obtained when the fuzzy gene selection method was applied to the thyroid cancer (GSE33630) dataset for most classifier models applied and specifically, MLP where 72% was the average accuracy score with  $k=5$  when omitting FGS while 93% when FGS employed. Besides, decreased the number of genes from 23516 to 76 genes which lead to less time consumption to train an algorithm, less complexity, interpretability, and provide early cancer detection. Fig. 5.6 clarifies the difference in the accuracy scores with  $k=5$  between applying the FGS method and omitting FGS for five different classifier models. FGS showed that it can improve classifier method performance while reducing the overfitting problem.

Fig 5.7 compares five classifier algorithms for the gene expression of nine cancer types (TCGA4) when FGS was implemented. The results demonstrate a small improvement,

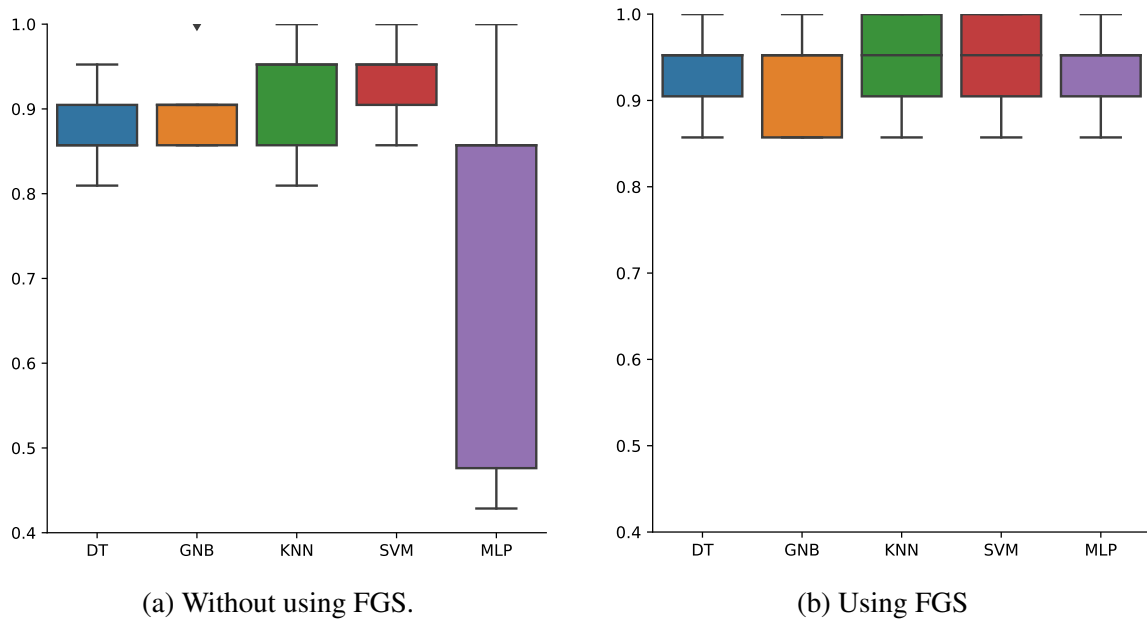


Fig. 5.6 Comparing accuracy score when using and omitting FGS (GSE33630)

particularly with MLP and DT methods. Although the improvement was not high, the number of selected genes was quite modest when compared to the original datasets, which chose 298 genes out of 56603. MLP achieved an accuracy score of 98.6% when classifying nine cancer types. Therefore, the FGS approach has the ability to identify informative genes for training a classifier with the same or improved accuracy. By selecting the strongest subset of genes, time spent for training is decreased, and the complexity of the classifier is reduced. Furthermore, the possibility of overfitting the model is reduced or even avoided. All of this contributes to improving the performance of a classifier algorithm.

Another dataset (TCGA2) was used to compare the effectiveness of the FGS approach on five classifier methods using k cross-validation with  $k=5$ . There was a minor improvement in accuracy was found with MLP as shown in Fig. 5.8. However, there was a substantial lowering in the number of genes where 116 out of 20531 genes were chosen to train and test the classifiers. The accuracy gained was modest when compared to the other data used while using and excluding FGS.

Fig 5.9 compares the five classifier algorithms for kidney cancer datasets with  $k=5$  when the FGS methodology is used and when it is not used. The results indicated that FGS has a beneficial influence on four classifiers and significantly reduces the number of genes, with 78 out of 23516 genes chosen to train the classifier algorithms. It can be concluded that when the FGS technique was combined with five classical classifier approaches, minor gains in accuracy were accomplished with a significant reduction of chosen genes. The best accuracy

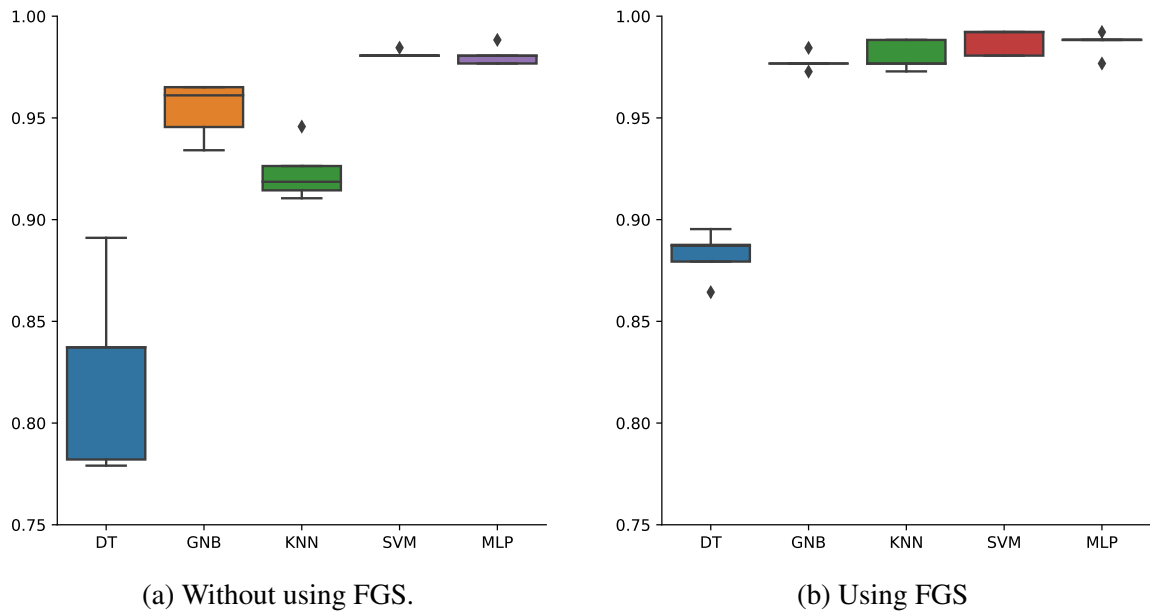


Fig. 5.7 Comparing accuracy score when using and omitting FGS (TCGA4)

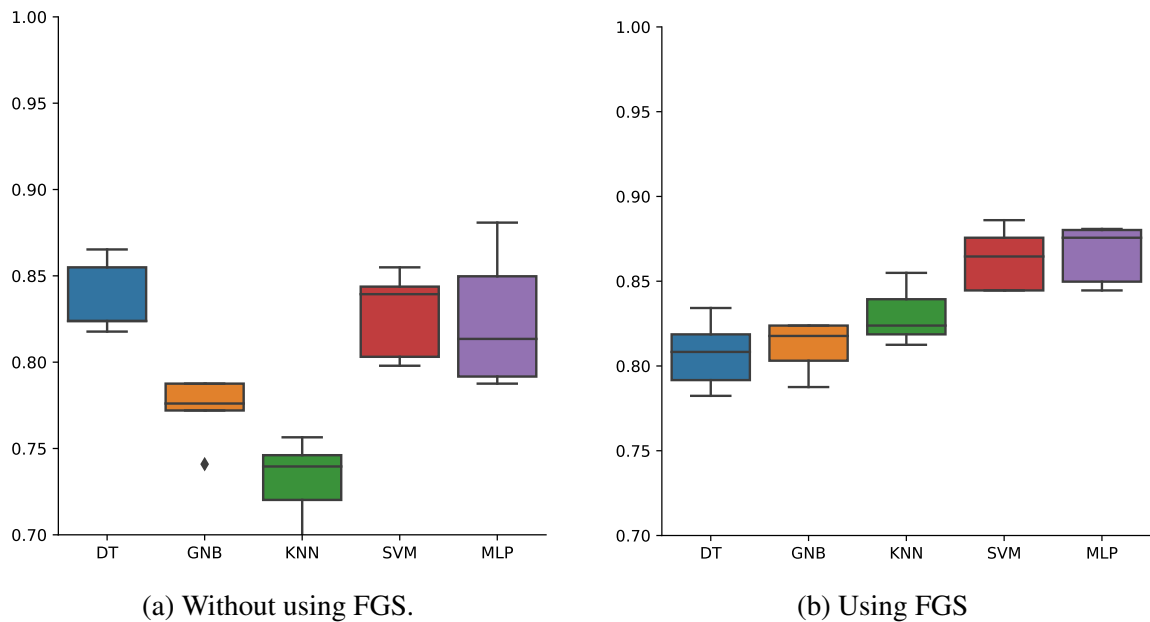


Fig. 5.8 Comparing accuracy score when using and omitting FGS (TCGA2)

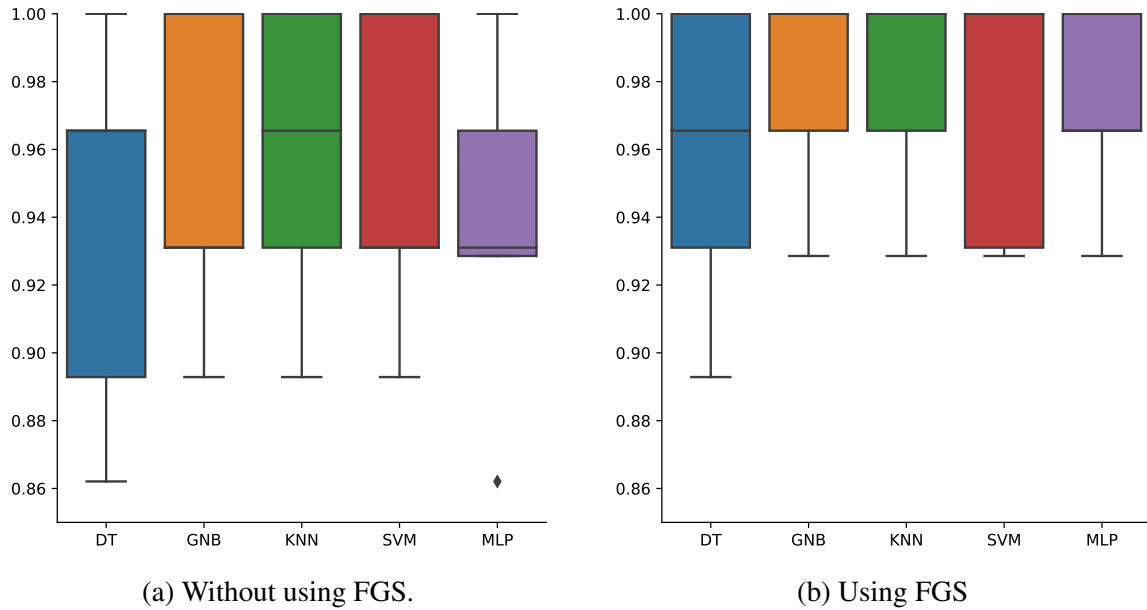


Fig. 5.9 Comparing accuracy score when using and omitting FGS (GSE53757)

achieved while omitting FGS is 95.7% when KNN is employed, whereas the highest accuracy reached when utilizing FGS and GNB combined is 97.8%. Although the enhancement was not great, the number of genes chosen was very small, which means there is less chance of overfitting, less classifier complexity, and less time required for training.

When GNB was used without the FGS technique, the accuracy was 77.5%, however, when FGS was used, the accuracy was 98.6% as shown in Fig. 5.10. Furthermore, MLP obtained 95.8% accuracy without using FGS, whereas MLP achieved 99% while using FGS. Furthermore, only 11 of the 20531 genes were chosen. FGS has once again proved its capacity to decrease the number of chosen genes while accurately classifying cancer. In conclusion, as previously stated, one of the key aims of the thesis is to reduce the number of identified genes that contribute to early cancer detection while also minimizing the overfitting issue and lowering dimensionality. FGS with dataset demonstrated its usefulness in meeting one of the thesis objectives while also improving the accuracy of some classifiers such as MLP and GNB.

The acquired accuracy in this dataset (GSE43580) was not improved while using and omitting FGS, as shown in Fig. 5.11, even though the number of chosen genes was reduced from 54675 to 28 genes. Furthermore, the accuracy attained in both situations was not good where 86% accuracy is achieved when MLP is applied to the original dataset and 86.6%

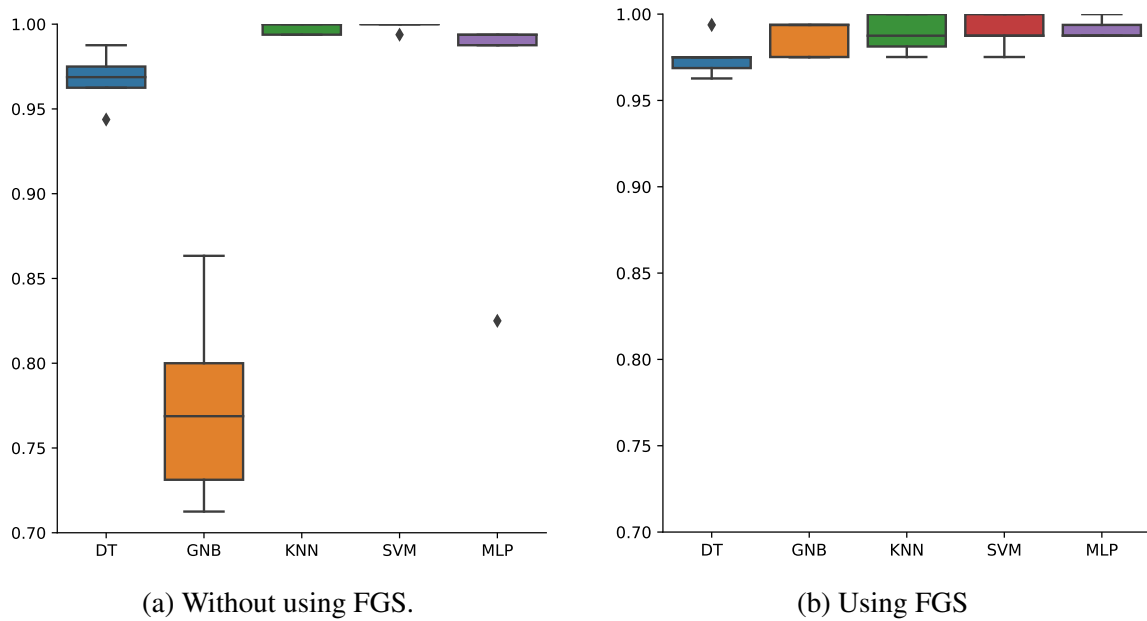


Fig. 5.10 Comparing accuracy score when using and omitting FGS (TCGA7)

accuracy is achieved when FGS and MLP are used together. This prompted the development of a novel classifier algorithm capable of reliably classifying cancer using this type of data.

Fig 5.12 indicates that using FGS slightly increases the accuracy while reducing the number of genes. The number of selected genes decreased from 768 to only 28 genes that have been identified for training the classifier algorithms. However, whether or not FGS is employed, the resulting accuracy is poor in comparison to cancer sensitivity. As a result, dealing with this type of data necessitates the development of a new classifier approach that has the potential to improve the accuracy of cancer classification. The following stage of this thesis addressed the limitations that were discovered when FGS was applied to conventional classifier techniques.

A significant improvement was shown when FGS and MLP were used simultaneously when accuracy increased from 50% to 97%. Furthermore, when FGS was used, the number of genes reduced from 13298 to 52. The majority of classifiers' performance was not improved, but the number of genes required to train them was reduced. This decreased the classifier's complexity and prevented or at least greatly reduced overfitting. To classify datasets containing data on lung cancer (GSE10072), Fig. 5.13 compares five distinct classifier methods while using and removing FGS with  $k=5$ .

Fig. 5.14 compares five typical classifier algorithms while using and excluding the FGS method with  $k=5$ . The findings indicated that the FGS technique improved the accuracy of three classifier methods (KNN, DT, and MLP) while reducing the number of chosen

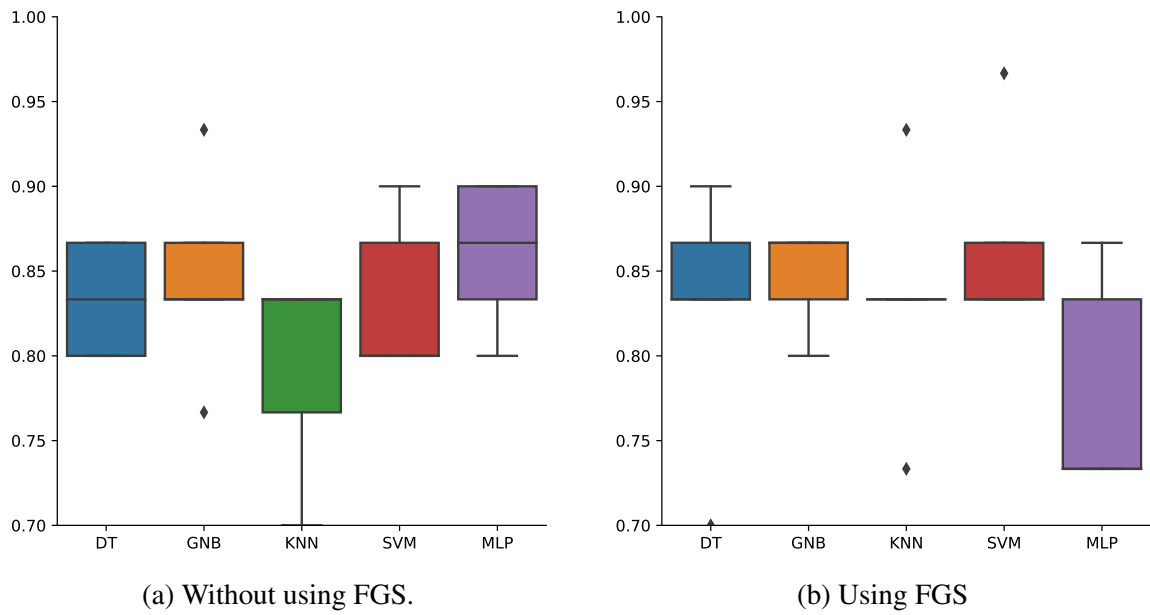


Fig. 5.11 Comparing accuracy score when using and omitting FGS (GSE43580)

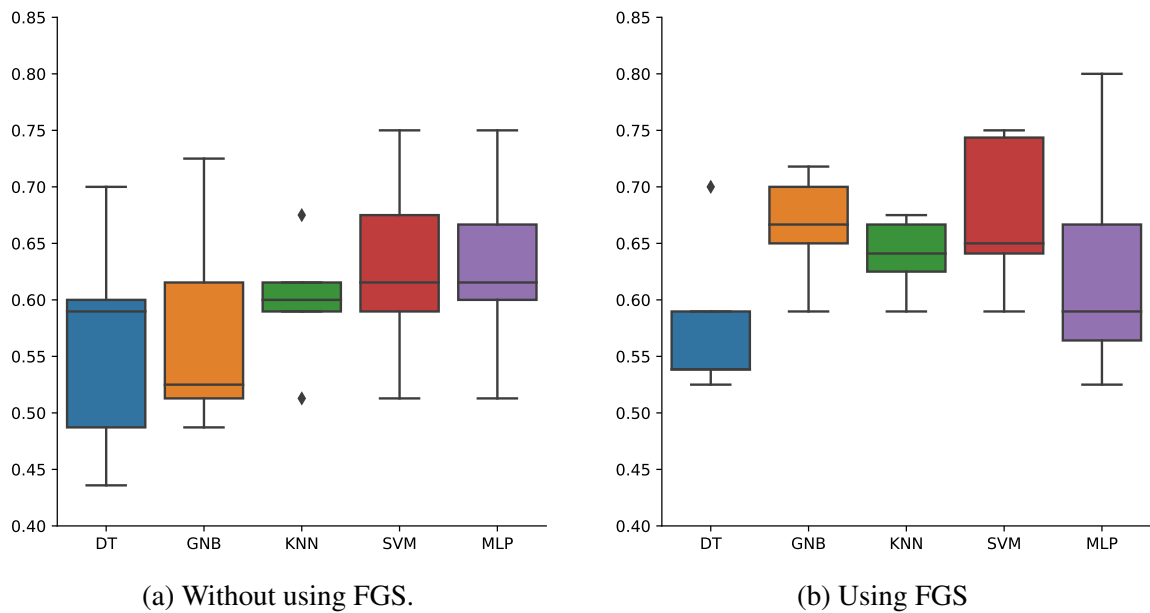


Fig. 5.12 Comparing accuracy score when using and omitting FGS (TCGA6)

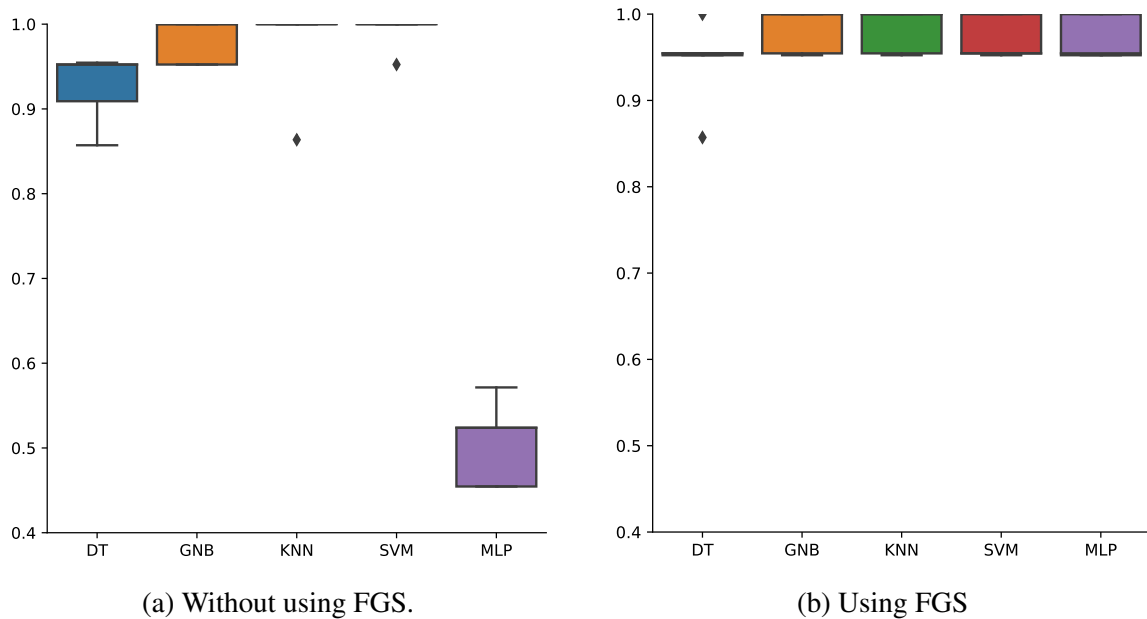


Fig. 5.13 Comparing accuracy score when using and omitting FGS (GSE10072)

genes from 16383 to 37. When the MLP algorithm was used individually, the results were 60.9%, 50.6%, 80%, and 60.5% for accuracy, precision, recall, and f1-score, respectively, while the results were 99.4%, 100%, 98.8%, and 99.4% for accuracy, precision, recall, and f1-score, respectively, when FGS and MLP were employed together. In summary, the FGS method reduced the number of selected genes that would be used as identifiers for training the classifiers and enhancing the performance of some of the classifiers. Additionally, even though some classifiers have not been improved such as SVM and GNB in terms of accuracy, the complexity of the classifier and the time taken through training were reduced. Most importantly, the outcomes of those classifiers were not reduced. As a result, while this approach failed to enhance accuracy in SVM and GNB, it was not having a negative impact either. Instead, it had a positive impact on reducing the number of genes, which aids in preventing the issue of overfitting and reducing the complexity of the classifier.

Fig. 5.15 shows that applying the FGS approach to five classifier methods slightly improved the results, but they are still substandard, even though the number of genes decreased from 33298 to 150. The findings of this dataset motivated the development of a novel classifier algorithm capable of effectively classifying this data. As seen in both figures, the accuracy gained in the five classifier algorithms was unsatisfactory before and after using FGS. FGS employed to select informative genes of lung cancer (GSE66499) datasets that would be used to train the classifier models. FGS indicates improving accuracy

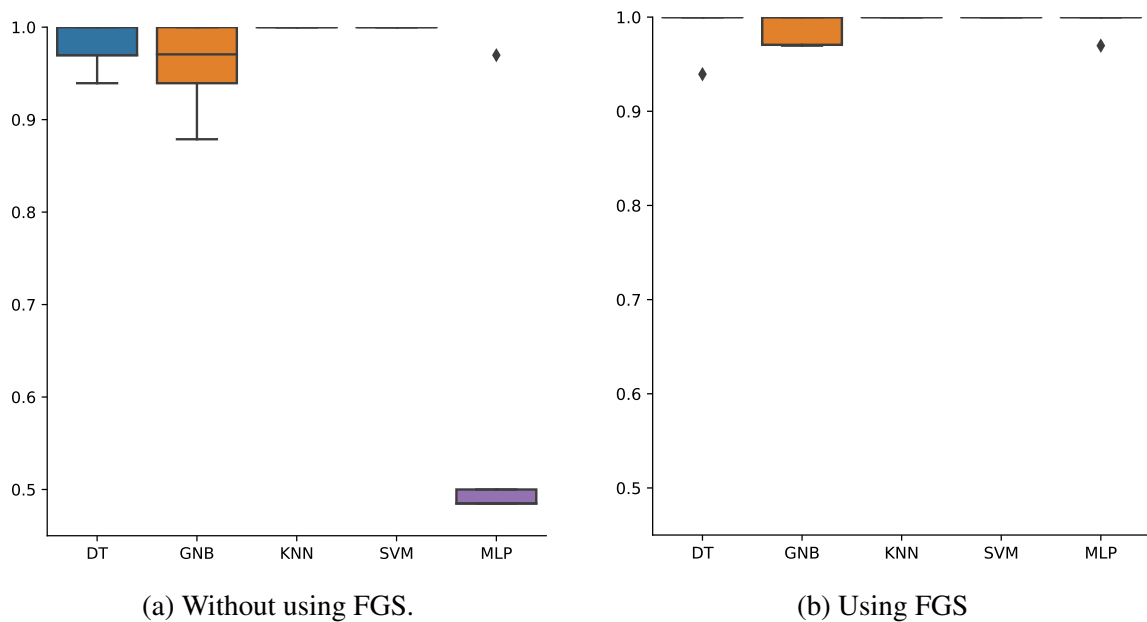


Fig. 5.14 Comparing accuracy score when using and omitting FGS (GSE75037)

using classical classifiers is not always possible, necessitating the development of a new classifier method to overcome these limitations.

This is another challenging dataset (GSE84437) for which FGS was unable to increase the accuracy of the five classifier algorithms. Furthermore, the accuracy reached in both situations (before and after employing FGS) is too poor as demonstrated in Fig. 5.16. Even though FGS has reduced the number of genes from 48710 to 105. When FGS and SVM were used together, the maximum accuracy was 37.8%. To cope with this data and properly classify cancer, developing a novel classifier algorithm is becoming necessary as described in the following development of the thesis.

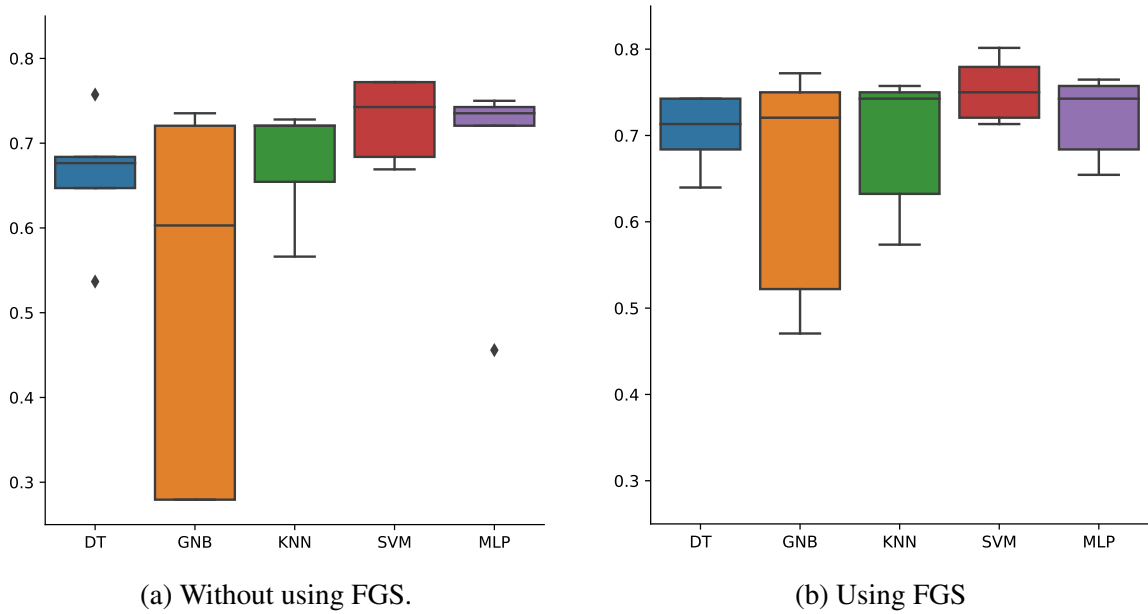


Fig. 5.15 Comparing accuracy score when using and omitting FGS (GSE66499)

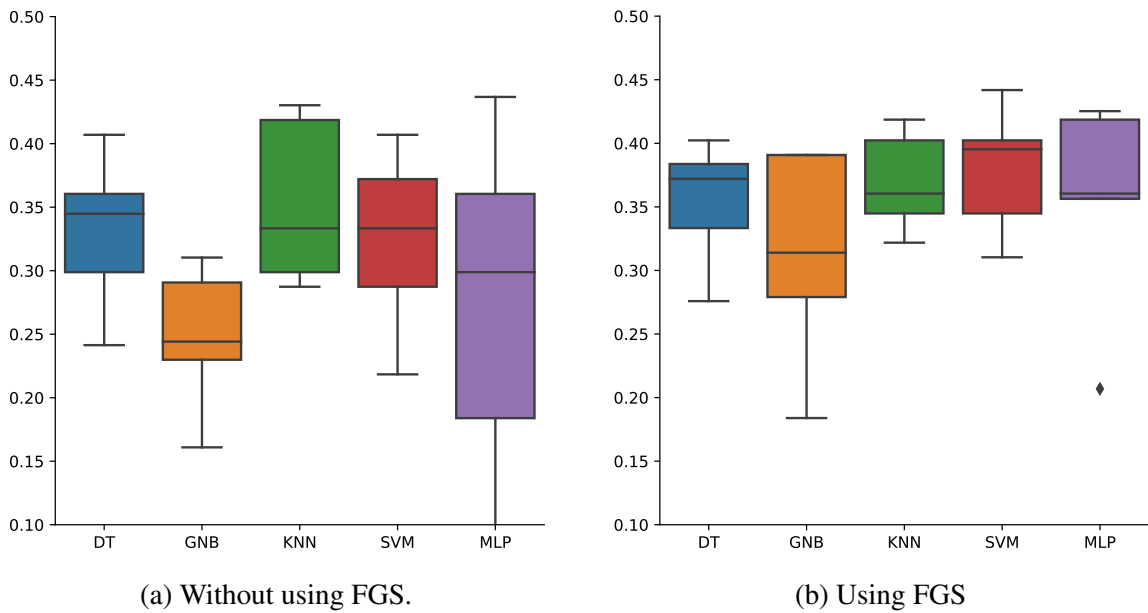


Fig. 5.16 Comparing accuracy score when using and omitting FGS (GSE84437)

### 5.3 Experimentation of applying FC

The thesis developed a new method for cancer classification which is FC that aims to classify cancer accurately and the ability to increase the generalisation of the algorithm to classify

cancer types and obtain the best accuracy achievement in all given datasets. To ensure that FC works as expected twelve datasets of gene expression have been employed that include a small and large volume of samples, as well as binary and multi-class labels. This section of the thesis tries to highlight three major points which are the datasets used, results obtained, and results discussion.

### **5.3.1 The datasets used to examine FC efficiency.**

Twelve datasets were used to train and test the FC. Datasets include (GSE45827, GSE33630, GSE19804, GSE14520, GSE53757, GSE10072, GSE66499, GSE84437, TCGA1, TCGA2, TCGA6, and GSE43580 ).

### **5.3.2 The results achieved with FC.**

This section investigates the comparison of FC to five classical classifier algorithms on twelve datasets when FGS used. The full details are presented in Table5.4 including the datasets used for training and testing the approach, and the achieved accuracy, precision, recall, and f1-score. The results demonstrate that developing the new FGS-FC greatly contributed to enhancing the model performance in all the datasets employed. It not only reduced the number of selected genes but provided accurate classification. In summary, this experiment aims to show the effectiveness of developing FC to classify expression cancer data. Especially, with the datasets accomplished poor results when classical classifiers and FGS used as described in previous experimentation. To achieve that, FC was compared to five classical classifiers in  $k=5$  when FGS used. The findings demonstrated that the developed FC outperformed five classical classifiers. FC achieved accuracy scores ranging (from 92.8% to 100%), precision ( from 95% to 100%), recall (81.4%-100%), and f1-score (85%-100%) for all employed datasets.

In general, The outcomes indicated that all employed datasets were improved when FC was used compared to classical classifiers. There are some of these datasets highly enhanced with FC such as (GSE84437, GSE66499, GSE43580, TCGA2, TCGA3, and TCGA6). Whereas, the rest of the other datasets improved by a lower percentage compared to the mentioned datasets because they were already achieved good results when FGS used. In conclusion, FC met the goal of its development to overcome the challenges that were identified when FGS and classical classifiers were applied together.

Table 5.4 A comparison of five classical classifiers vs Fuzzy classifier method

Datasets	N-Gene	Gene Selection	Classifiers	AC %	Pre %	Rec %	F1 %
GSE45827	29873	No	DT	85.8	83	82.6	81.5
			GNB	89	92.7	88.8	89
			KNN	85	88.4	87.7	86.9
			SVM	94.8	96.3	95.8	95.8
			MLP	20.6	6	17.9	7
GSE45827	68	FGS	DT	89.6	90.9	89.6	88.8
			GNB	91.6	94.5	92	92.8
			KNN	96.7	97.59	97.38	97.36
			SVM	97.4	98	97.66	97.75
			MLP	98	98.8	98	98.3
GSE45827	68	FGS	FC	100	100	100	100
GSE33630	23518	No	DT	87.6	77.6	81	79
			GNB	90.4	93.7	89.7	90
			KNN	91.4	87.7	86.5	86.3
			SVM	93.3	95.3	92	92.4
			MLP	72.3	55.6	64.5	58.5
GSE33630	76	FGS	DT	93.3	93.4	93.5	92.5
			GNB	92.3	88.3	89.8	88.8
			KNN	94	96	92.8	93
			SVM	94	96	92.8	93
			MLP	92.3	88.3	89.9	88.8
GSE33630	76	FGS	FC	100	100	100	100
GSE19804	45782	No	DT	89	89.9	88.3	88.9
			GNB	92.5	95	90	91.9
			KNN	90.8	88	95	91.3
			SVM	95.8	96.6	95	95.7
			MLP	50	20	40	26.6
GSE19804	36	FGS	DT	90.8	94.5	86.66	90
			GNB	95.8	96.7	95	95.7
			KNN	96.66	96.79	96.66	96.66
			SVM	96.66	96.79	96.66	96.66
			MLP	96.66	96.79	96.66	96.66
GSE19804	36	FGS	FC	100	100	100	100
TCGA1	972	No	DT	91	87	85.3	85.8
			GNB	94	89.7	92	90.7
			KNN	88	83.3	81.5	81.9
			SVM	93.6	91	88.9	89.8
			MLP	94	90.8	89.8	90
TCGA1	25	FGS	DT	91.7	88	87	86.5
			GNB	92.4	87.7	90.8	89
			KNN	93.6	89.4	90	89.6
			SVM	94	90.5	90.77	90.5
			MLP	95	92.3	91.6	91.6
TCGA1	25	FGS	FC	97	95	94	95

Datasets	N-Gene	Gene Selection	Classifiers	Ac %	Pre %	Rec %	F1 %
GSE43580	54675	No	DT	85.3	87.3	83.4	84.7
			GNB	84.6	91.5	75	82
			KNN	79.3	73	93	81.5
			SVM	83.3	84.3	80.6	82.3
			MLP	86	86.3	84.7	85
GSE43580	28	FGS	DT	79.3	77.6	82	79.4
			GNB	84.66	92.8	75	82.3
			KNN	83.3	91.95	72.57	80.46
			SVM	86.66	98	73	83.8
			MLP	78	77.3	78	77.6
GSE43580	28	FGS	FC	98	98	97	98
GSE14520	13425	No	DT	90	90.6	88.9	89.7
			GNB	95	95.6	94.4	94.8
			KNN	94	91	97.6	94
			SVM	97	96.4	97.6	97
			MLP	86.7	76.5	96.7	76.5
GSE14520	23	FGS	DT	95	95.4	94.9	95
			GNB	96.6	96	97	96.59
			KNN	96.6	96	97	96.59
			SVM	96.85	96	97.68	96.8
			MLP	95.5	95.6	95.3	95.3
GSE14520	23	FGS	FC	99	98	99	98
GSE53757	23516	No	DT	95.7	94.6	97	95.8
			GNB	95	95.6	94	95
			KNN	95.7	95.6	95.7	95.6
			SVM	95	94.8	96	95
			MLP	93.7	98.5	89	93.3
GSE53757	78	FGS	DT	95.8	93.7	98.5	95.9
			GNB	97.8	97	98.5	97.8
			KNN	97.8	97	98.5	97.8
			SVM	97	96	98.5	97
			MLP	96.5	96	97	96.5
GSE53757	78	FGS	FC	100	100	100	100
GSE10072	13298	No	DT	94.3	93.5	96	94.3
			GNB	98	100	95.7	97.7
			KNN	97	95.3	100	97.3
			SVM	99	100	98	98.9
			MLP	50	18	4	25
GSE10072	52	FGS	DT	93.4	93.56	93.77	93
			GNB	98	100	96	97.8
			KNN	99	100	98	98.9
			SVM	98	100	96	97.89
			MLP	98	98	96	96.9
GSE10072	52	FGS	FC	100	100	100	100

Datasets	N-Gene	Gene Selection	Classifiers	Ac %	Pre %	Rec %	F1 %
GSE84437	48710	No	DT	33	20	23	19.7
			GNB	24.7	28.5	24.4	20.6
			KNN	35.3	19.7	23.7	19
			SVM	32.3	18.3	24	18.6
			MLP	27	10	23.4	12.6
GSE84437	105	FGS	DT	35.3	22.4	24.5	21
			GNB	31	34	42	29.4
			KNN	36.96	28.5	28.3	25.6
			SVM	37.89	33	29.7	26
			MLP	35.3	32.3	32.4	29.46
GSE84437	105	FGS	FC	92.8	95.6	81.4	85
GSE66499	33298	No	DT	66	38.5	32	34.6
			GNB	52.3	38	50	27.3
			KNN	67.7	42.4	17.8	19.7
			SVM	72.7	62	28.9	37
			MLP	68	59.3	20.5	19.3
GSE66499	150	FGS	DT	70.4	49.9	24.7	31.4
			GNB	64.7	53.4	56.8	44.6
			KNN	69	55	26.3	31.4
			SVM	75	69	36.8	44.5
			MLP	72	53.3	37.3	42.6
GSE66499	150	FGS	FC	95	95	92.4	93.4
TCGA2	20531	No	DT	83.7	83.4	80	81.3
			GNB	77	77.6	77.7	77
			KNN	73	76	64.5	66.3
			SVM	82.7	82	82.5	82
			MLP	84.4	84.4	83	82
TCGA2	116	FGS	DT	80.7	76.5	74.3	74.8
			GNB	81	76.4	80	77.56
			KNN	82.98	84	75	77.68
			SVM	86.3	87	82.6	84.3
			MLP	86.3	86	84.88	85
TCGA2	116	FGS	FC	97	98	97	97
TCGA6	768	No	DT	58	55.8	58.4	56.8
			GNB	57.3	54.5	69	60.5
			KNN	59.8	56.8	68	61.3
			SVM	62.8	63.3	56.3	59
			MLP	62.8	63	57.5	59
TCGA6	28	FGS	DT	58.34	55.9	59.8	56
			GNB	66.48	61	84	70.56
			KNN	63.94	59.87	75.73	66.46
			SVM	67.48	64	75.67	68.8
			MLP	65.44	63.8	64	63
TCGA6	28	FGS	FC	98	99	98	98

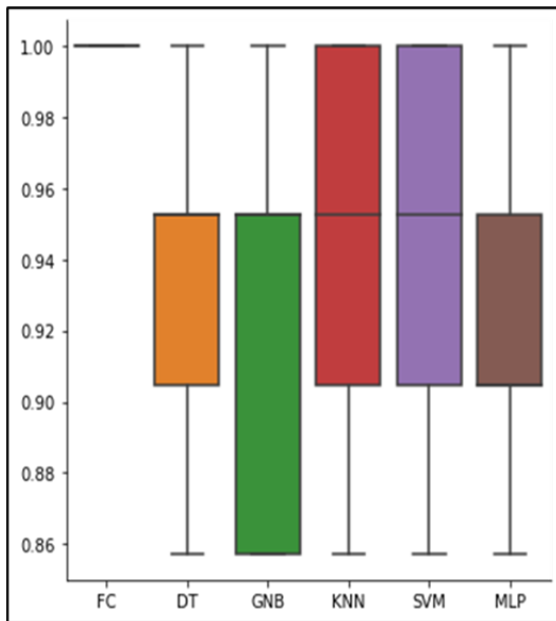


Fig. 5.17 FC vs classical classifiers (GSE33630) employing FGS

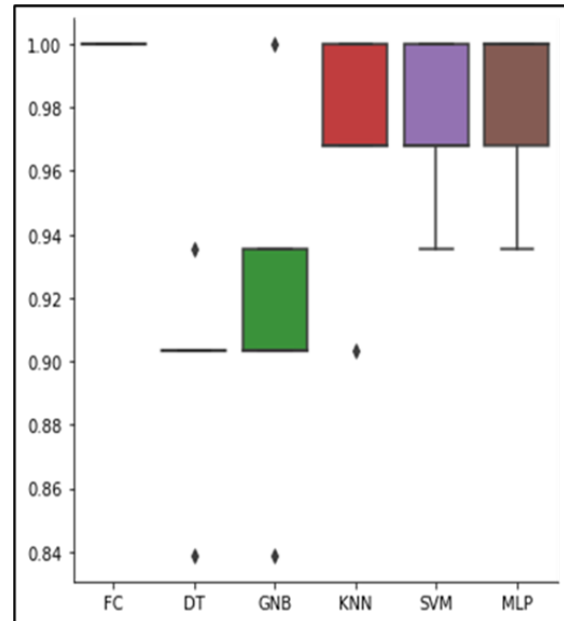


Fig. 5.18 FC vs classical classifiers (GSE45827) employing FGS

### 5.3.3 Discuss FC results

The results demonstrated that the developed FC method overcomes the other classifiers in terms of accuracy and other evaluation metrics when applied to thyroid cancer (GSE33630) as described in Table 5.4. When using classical classifiers, SVM and KNN had the maximum 94 % accuracy, 96 % precision, 92.8 % recall, and 93 %, and f1-score. Whereas FC achieved 100% for all evaluation metrics. Comparing the accuracy of FC against classical classifier with k=5 described in Fig 5.17.

Fig 5.18 depicts the lowest accuracy attained while using DT, which was 89.6%, and the best accuracy achieved when using MLP, which was 98.6%. While the FC was achieved with a 100% accuracy rating. As a consequence, when used to classify breast cancer datasets (GSE45827), FC performed better than the other classifiers.

A comparison of five classifier algorithms against FC for classifying Gastric cancer (GSE84437) in k=5 is shown in Fig 5.19. This dataset is regarded as a challenging dataset because the accuracy obtained by the five applied classifiers, which ranged from 31% to 37.89% when classical classifiers with FGS used, was extremely low. FC method significantly improved the accuracy by reaching 92.8%. FC demonstrates that enables accurate classifying of cancer expression, even in the worst scenarios. A comparison of five classical classifiers to FC that attempt to classify the lung cancer expression dataset (GSE66499) in k cross-validation with k=5 is shown in Fig 5.20. The findings indicate that all of the classifier

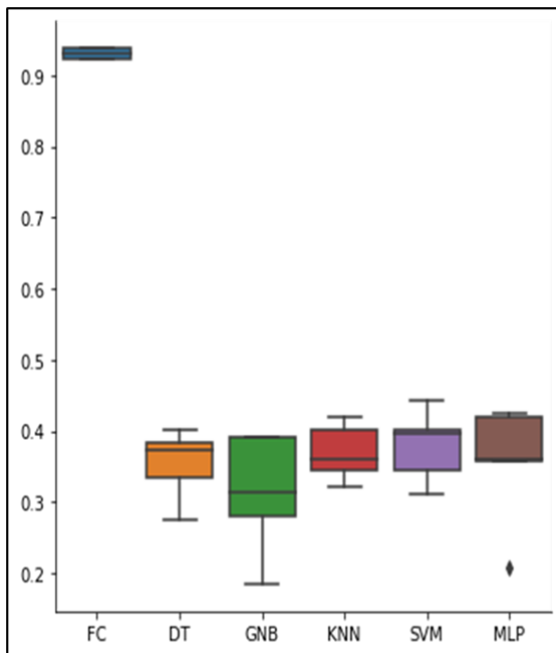


Fig. 5.19 FC vs classical classifiers (GSE84437) employing FGS

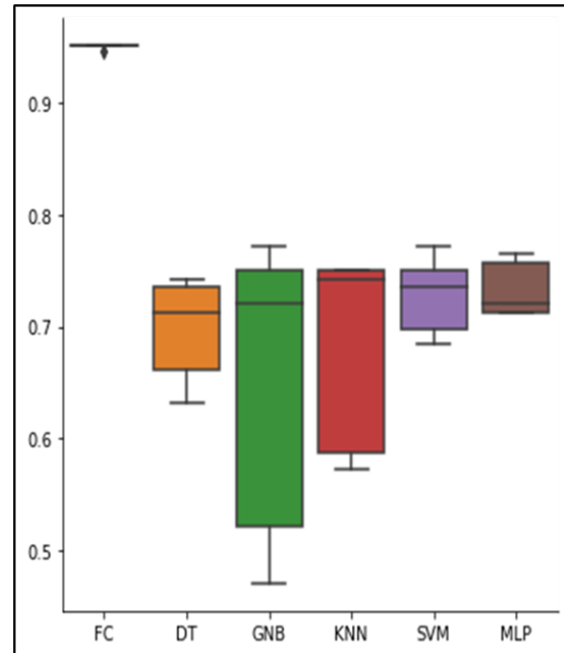


Fig. 5.20 FC vs classical classifiers (GSE66499) employing FGS

models used had poor accuracy. There is no classifier approach that was able to effectively identify data even though FGS used. This dataset is also regarded as a challenging dataset to classify. The accuracy of the preserved data ranged from 69.7% to 73.3% for the five classical classifiers. FC produced promising results, scoring 95%, 95%, 92%, and 93% for accuracy, precision, recall, and F1-score, respectively. FC demonstrated the capacity to handle some datasets that can not be resolved when using classical classifiers with FGS.

Fig 5.22 illustrates the achieved accuracy scores of five classifier techniques and FC for five cancer types (TCGA1)  $k=5$  when FGS used. The acquired accuracy of the five-classifier was near to each other, with the lowest DT being 91.7% and the highest MLP being 95%. Even though the accuracy is 95%, the precision, recall, and f1-score with MLP were 92.3%, 91.6%, and 91.6%, respectively. The developed model tried to increase not only the accuracy but also the other factors used to assess a model. FC improved all evaluation metrics, including 97% accuracy, 95% precision, recall, and f1-score. In short, the developed model improved by 3 for accuracy, 4.2 for precision, 5.2 for recall, 5 for f1-score when compared to MLP alone, and 9 when compared to DT classifier.

FGS and classical classifiers failed to accurately classify the lung cancer dataset (GSE43580). Even though the FGS technique reduced the number of identifying genes used to train the model, adequate accuracy was not reached. As a result, employing the FC approach to deal with this data is critical. The best accuracy attained in the five classifier approaches

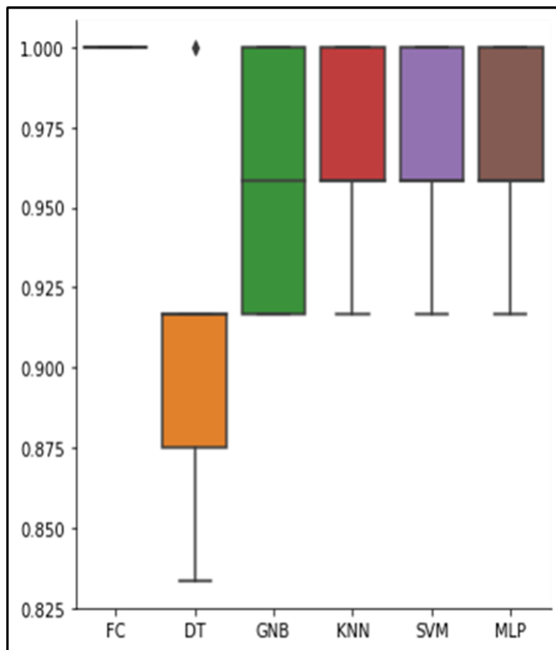


Fig. 5.21 FC vs classical classifiers (GSE19804) employing FGS

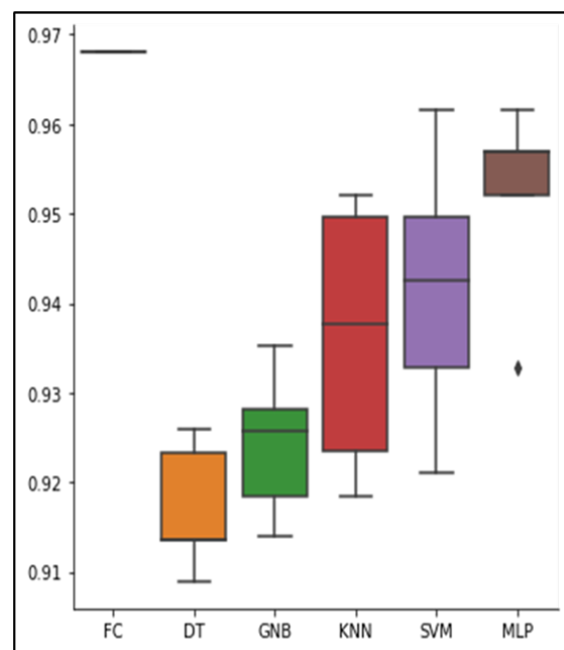


Fig. 5.22 FC vs classical classifiers (TCGA1) employing FGS

was 86.6% when SVM was used, while FC achieved 98% accuracy as shown in Fig 5.23. FC increased the accuracy by 11.4% when compared to the best accuracy achieved by the five classical classifiers. Noteworthy, FC improved the recall from 73% when SVM was used to 97% when FC was used. Good accuracy alone does not qualify a classifier as good, other assessment measures, such as precision, recall, and f1-score, are required to give the preferability for a classifier. Fig 5.24 compares five classifier techniques for classifying liver cancer (GSE14520) k cross-validation with k=5. By averaging the accuracy scores across k=5. The results revealed that FC achieved the highest average accuracy of 99%, followed by SVM with 96.86%, KNN with 96.6%, GNB with 96.6%, MLP with 95.5%, and 95%. These findings suggest that FC outperformed the other classifiers in terms of accuracy, highlighting its potential as a promising choice for our classification task.

Fig 5.25 compares the accuracy scores of five classifiers against FC to classify kidney cancer (GSE53757) with k=5 using FGS to select 78 out of 23516 genes. The achieved accuracy scores for the five classifiers ranged from 95.8% to 97.8%, While FC achieved an accuracy score of 100%. FC also showed effectiveness with this dataset which It is outperforming the other classifiers. Five classical classifiers were compared to FC to classify the lung cancer dataset (GSE10072) when FGS used. The average accuracy scores in k=5 were ranging from 93.4% as the lowest with DT and 99% as the highest accuracy when KNN was used while FC achieved 100%. The findings indicate that FC overcomes the other

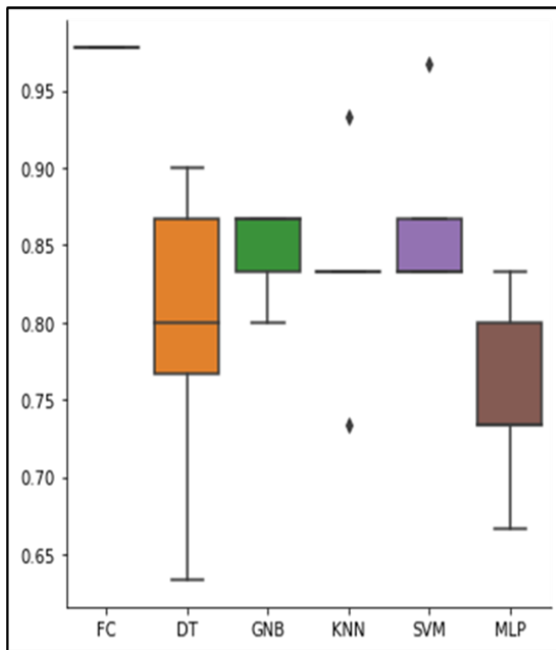


Fig. 5.23 FC vs classical classifiers (GSE43580) employing FGS

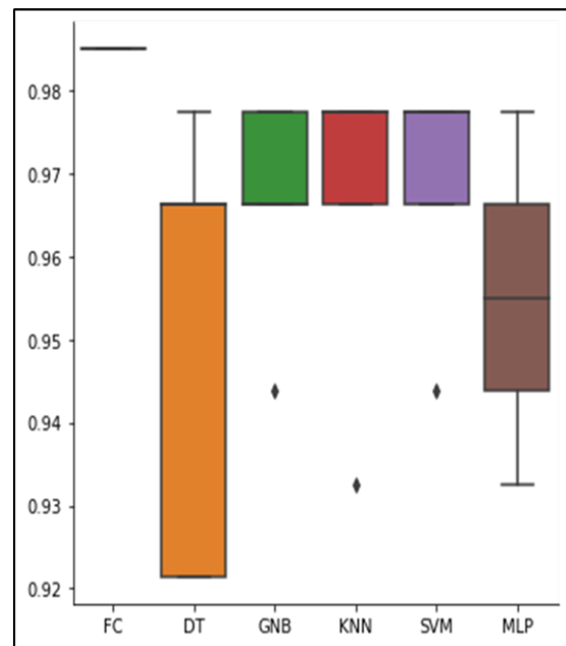


Fig. 5.24 FC vs classical classifiers (GSE14520) employing FGS

classifiers as described in Fig 5.26. Even though, the results achieved by classical classifiers were good, however, FC increase the accuracy. FC increased the accuracy by 1% compared to KNN, and by 6.6% compared to DT.

Five classical classifiers and FC were compared to classify breast cancer subtypes (TCGA2) with  $k=5$  when FGS used. The results indicate that classical algorithms failed to accurately classify the data. The highest accuracy scores were 86.3% when MLP and SVM were used. FC demonstrated that outperformed other classifiers by reading 97% as described in Fig 5.27.

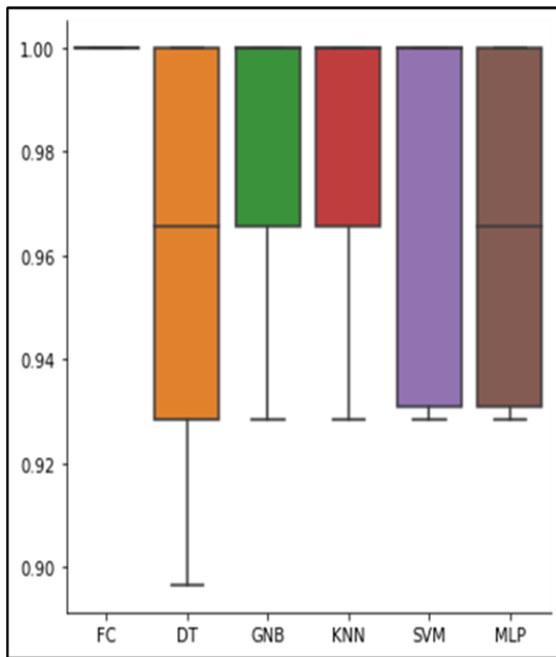


Fig. 5.25 FC vs classical classifiers (GSE53757) employing FGS

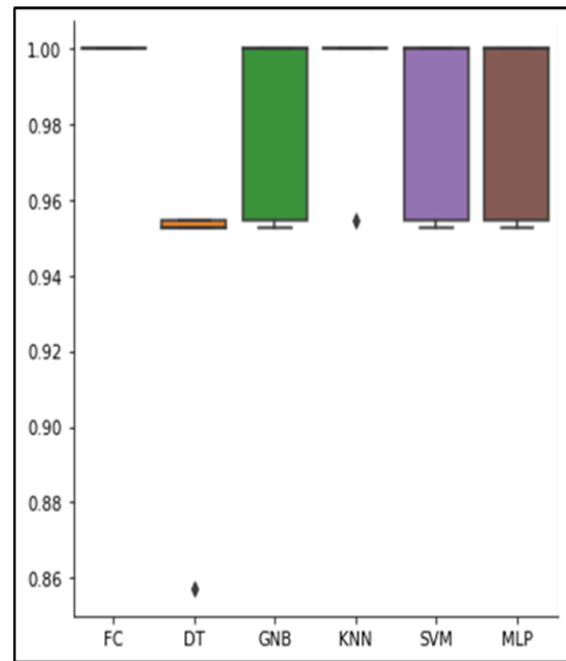


Fig. 5.26 FC vs classical classifiers (GSE10072) employing FGS

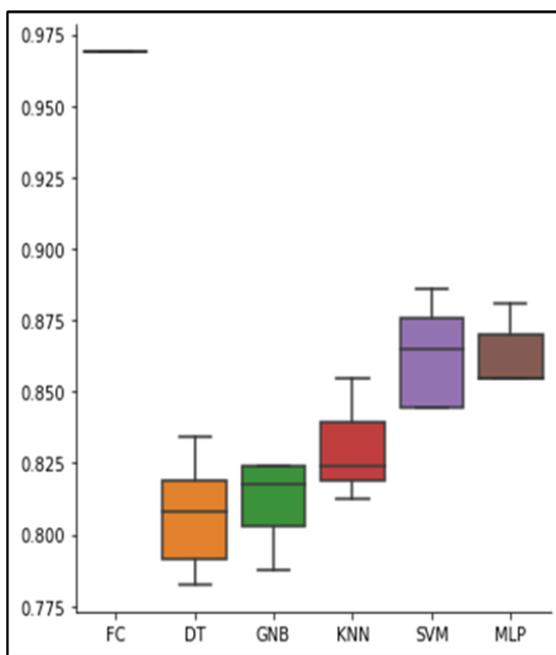


Fig. 5.27 FC vs classical classifiers (TCGA2) employing FGS

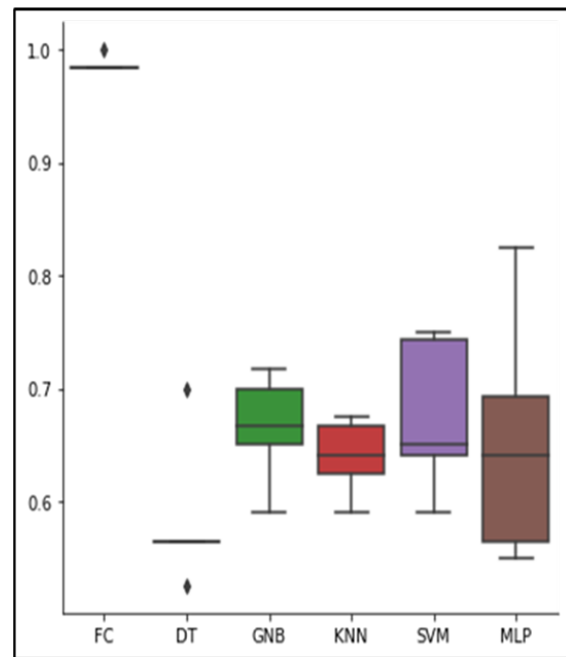


Fig. 5.28 FC vs classical classifiers (TCGA6) employing FGS

FC was developed to address datasets with poor accuracy when using FGS with classical classifiers used. Fig 5.28 compared the accuracy scores for five classifiers against FC in  $k=5$  when FGS used to classify nine cancer types. The findings revealed that the accuracy of the five classifiers, which ranged from 58% to 67%, was extremely low, but FC demonstrated its efficiency by obtaining 98% accuracy. These results showed that FC performed much better than other classifiers in this dataset.

## 5.4 Experimentation of applying FGSWP

FGSWP was developed to achieve two major goals: first, analysing the selected genes that have been identified by the fuzzy gene selection method to determine whether or not these genes have an influence on accuracy. Second, further reducing chosen genes can increase or at least maintain accuracy as it was before removal. That is, obtaining fewer genes with no influence on accuracy accomplishment. To assess the efficacy of developing this approach, it was trained and tested using cancer expression datasets. It also used classical classifier algorithms and the fuzzy classifier approach to demonstrate the impact of both.

### 5.4.1 The datasets used to examine FGSWP efficiency

To test and train the developed model (FGSWP), six datasets were employed (GSE53757, GSE45827, GSE33630, TCGA2, GSE10072, and GSE43580). The datasets are smaller and larger cohorts, as well as having binary and multiclass labels. Five distinct classifier models were used to test and train the datasets.

### 5.4.2 FGSWP's results with Classical Classifiers

As mentioned earlier, the main objective of developing this approach is to reduce the number of genes selected by FGS without compromising accuracy. The experimental results demonstrate that the FGSWP method substantially reduced the gene count while maintaining or even improving the accuracy levels across the majority of the datasets analysed. It is noteworthy that these outcomes were obtained through the use of classical classifiers. Notably, significant reductions in the number of selected genes were observed in GSE33630, GSE45827, and GSE10072 datasets, as indicated in Table 5.5. By employing FGSWP on GSE33630, the gene count decreased from 76 to 17, demonstrating a substantial improvement. Similarly, in the case of GSE45827, the number of genes reduced from 68 to 30 upon applying FGSWP. In the case of GSE10072, the number of genes reduced from 52 to 5 upon applying FGSWP. Additionally, FGSWP led to increased accuracy scores with certain datasets, such as GSE43580

and GSE45827. Table 5.5 depicts the comparison of evaluation metrics for five classical classifiers employing FGS and FGSWP is shown in the Table. Old AC refers to the accuracy scores gained with FGS, whereas New AC represents the accuracy scores obtained with FGSWP. Similarly, the other evaluation metrics.

Fig 5.29 presents a comparison between FGS and FGSWP across five classifiers for the classification of lung cancer (GSE43580) using k cross-validation with  $k=5$ . The results demonstrate that FGSWP reduced the number of genes from 28 to 8. Furthermore, it significantly improved the accuracy scores for the employed classifiers. Specifically, when using FGS and MLP, the accuracy, precision, recall, and f1-score were observed to be 78%, 77.3%, 78%, and 77.6% respectively. However, when FGSWP and MLP were used, these metrics improved to 86.6%, 90.4%, 83.5%, and 86% respectively.

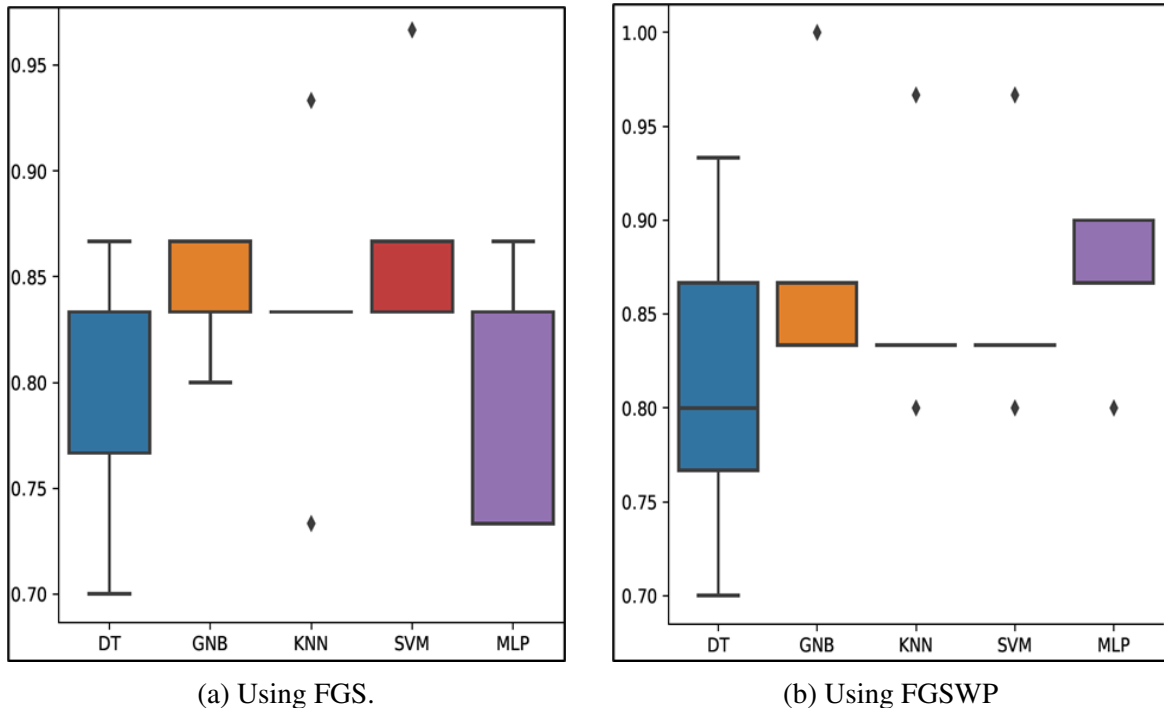


Fig. 5.29 A comparison of FGS vs FGSWP in five classifiers for (GSE43580)

The Fig. 5.30 illustrates a comparison of accuracy scores obtained from k cross-validations with  $k=5$  using two approaches: FGS and FGSWP. Among the five classifiers tested, the DT model exhibited a slight increase in accuracy score, while the remaining classifiers maintained a consistent performance. As mentioned previously, FGSWP aims to reduce the number of selected genes identified by the FGS approach. In line with this objective, the

Table 5.5 A comparison of FGS vs FGSWP across five classical classifiers.

Dataset	FGS	FGSWP	Classifier	Old AC %	New AC %	Old Pre %	New Pre %	Old Rec %	New Rec %	Old F1 %	New F1 %
GSE53757	78	69	DT	95	96.4	92.4	94.6	98.5	98.5	95.3	96.5
			GNB	97.8	96.5	97	97	98.5	95.7	97.8	96.3
			KNN	97.8	97.8	97	97	98.5	98.5	97.8	97.8
			SVM	97	97	96	96	98.5	98.5	97	97
			MLP	97	97	97	97	97	97	97	97
GSE33630	76	17	DT	93.3	94	93.4	94.6	93.5	94	92.5	93.7
			GNB	92.3	91.4	88.3	87.5	89.8	89	88.8	88
			KNN	94	92.3	96	88.3	92.8	90	93	88.8
			SVM	94	93.3	96	90	92.8	90.5	93	89.5
			MLP	93.3	94	95.3	96	92	92.8	92.4	93
GSE45827	68	30	DT	89.6	92.9	90.9	94.8	89.6	92.3	88.8	92.8
			GNB	91.6	92.9	94.5	94.4	92	94	92.8	94
			KNN	95.4	98	96.5	98.3	96	98	96	98
			SVM	98.7	98.7	99	99	98.8	98.8	98.9	99
			MLP	98.7	99.3	99.3	99.4	98.8	99.4	98.9	99.4
TCGA1	25	18	DT	91.7	91	88	87.5	87	86.8	86.5	86
			GNB	92.4	91.5	87.7	87	90.8	89.7	89	88
			KNN	93.6	93.3	89.4	89.6	90	89.4	89.6	89
			SVM	94	93.4	90.5	90	90.7	89.4	90.5	89.5
			MLP	95.3	94.6	92.4	91	91.8	91	91.7	90.7
GSE10072	52	5	DT	93.4	95.3	93.5	93.5	93.7	98	93	95.3
			GNB	98	96	100	96	96	96	97.8	96
			KNN	95.3	97	94.8	98	96	96	95	96.9
			SVM	98	97	100	98	96	96	97.8	97
			MLP	97	96	100	94.8	94	98	96.7	96
GSE43580	28	8	DT	80	81.3	78.7	82	82	79.3	80	80.4
			GNB	84.6	88	92.8	97	75	77	82.3	85.7
			KNN	83.3	85.3	92	96.6	72.5	72.4	80.4	82.3
			SVM	86.6	85.3	98	98	73.8	71	83.8	82
			MLP	78	86.6	77.3	90.4	78	83.5	77.6	86

developed FGSWP successfully reduced the gene count from 78 to 69, effectively excluding nine irrelevant genes.

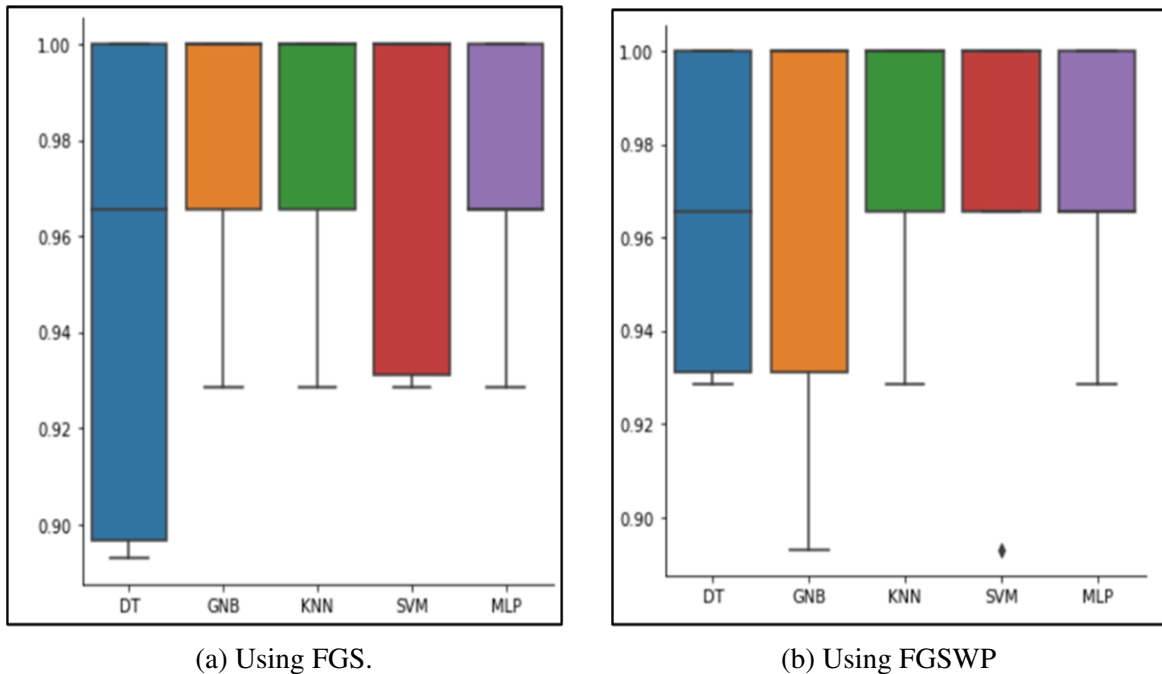


Fig. 5.30 A comparison of FGS vs FGSWP in five classifiers for (GSE53757)

Fig. 5.31 illustrates a minor gain in accuracy attained, notably in the MLP and DT classifiers when FGSWP was used. This was accomplished by reducing the number of genes in the data used to classify thyroid (GSE33630) from 76 to merely 17. In summary, this approach demonstrated its effectiveness by eliminating 59 genes without affecting the degree of accuracy in classifying thyroid cancer.

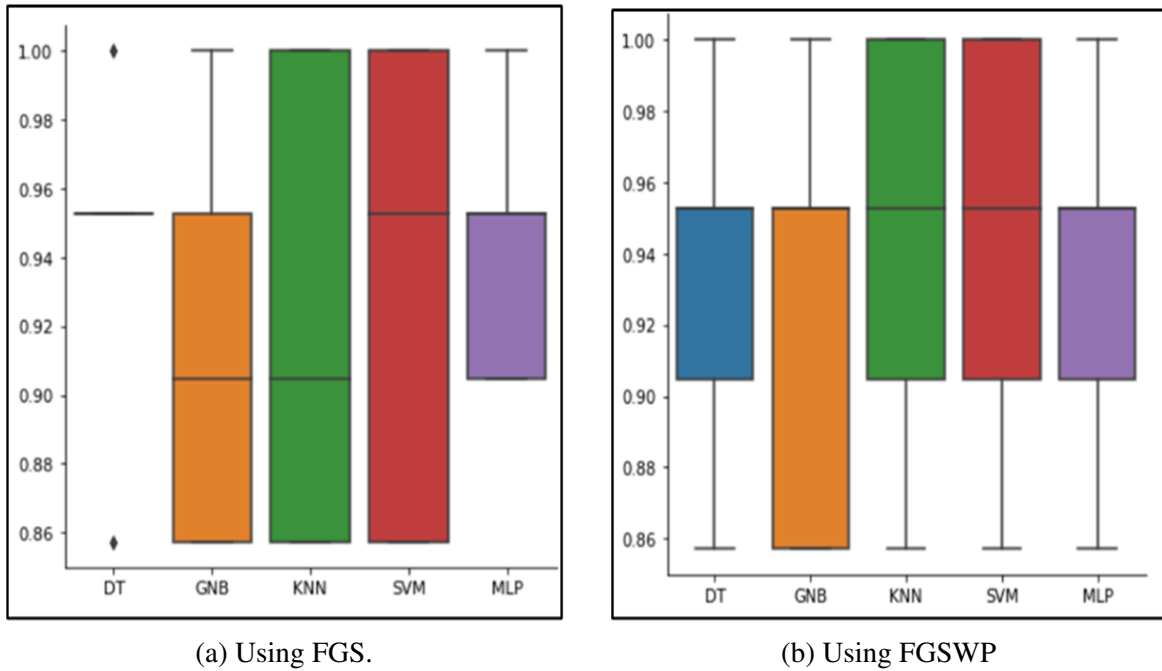


Fig. 5.31 A comparison of FGS vs FGSWP in five classifiers for (GSE33630)

The use of FGSWP resulted in a reduction of genes from 68 to 30, indicating the removal of 38 genes. Notably, Fig. 5.32 illustrates slight improvements in most of the classifiers used. In the breast cancer subtype classification dataset (GSE45827), FGSWP not only reduced the number of genes that were selected through using FGS with the same accuracy, which is the primary objective of this technique but also surpassed it by improving accuracy across different classifiers techniques. This suggests that FGSWP has the potential to enhance accuracy by utilizing the smallest possible number of genes for training the classifier. In conclusion, the FGSWP approach demonstrates the capability to either achieve comparable accuracy with a smaller gene count than FGS or outperform FGS in terms of accuracy while utilizing fewer genes. In either scenario, it offers a distinct advantage.

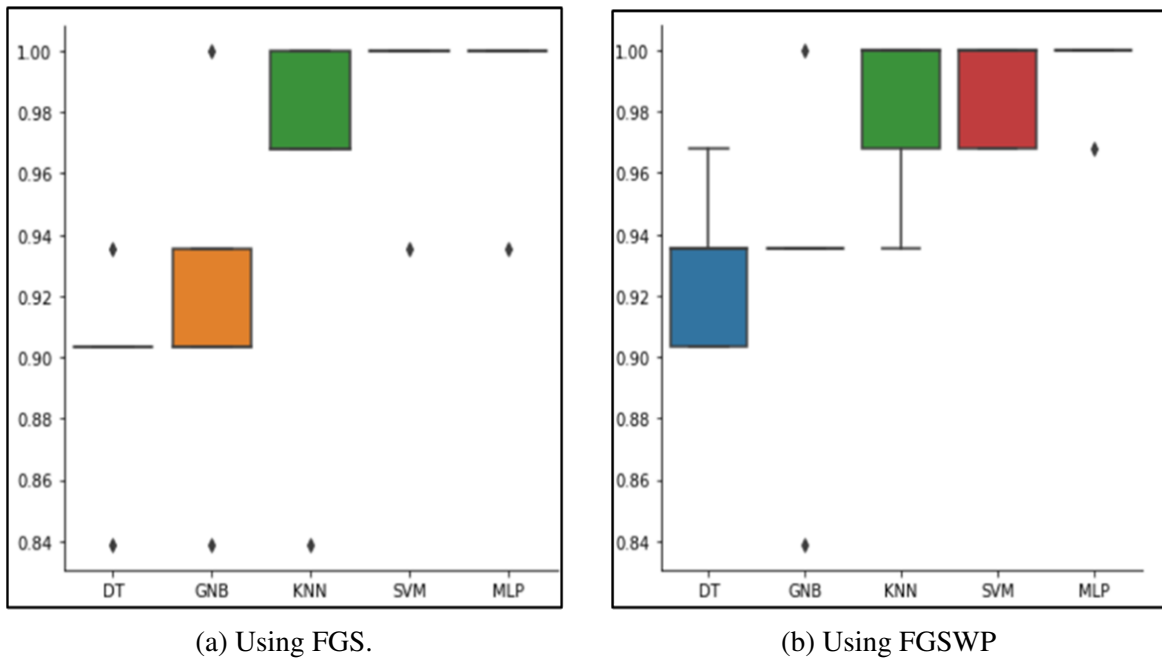


Fig. 5.32 A comparison of FGS vs FGSWP in five classifiers for (GSE45827)

A comparison of the five classifiers when using FGS and FGSWP in  $k=5$  was described in Fig. 5.33. Although, the number of selected decreased from 25 to 18 genes when FGSWP was used. However, the findings acquired from analysing five cancer types (TCGA1) revealed that there is no substantial difference, either positive or negative, in the degree of accuracy when FGSWP was employed. As a result, the developed FGSWP is efficient in attaining the stated aim of reducing the number of genes while retaining classification accuracy. Based on it, certain additional benefits were realised, including reduced training time, and less classifier complexity.

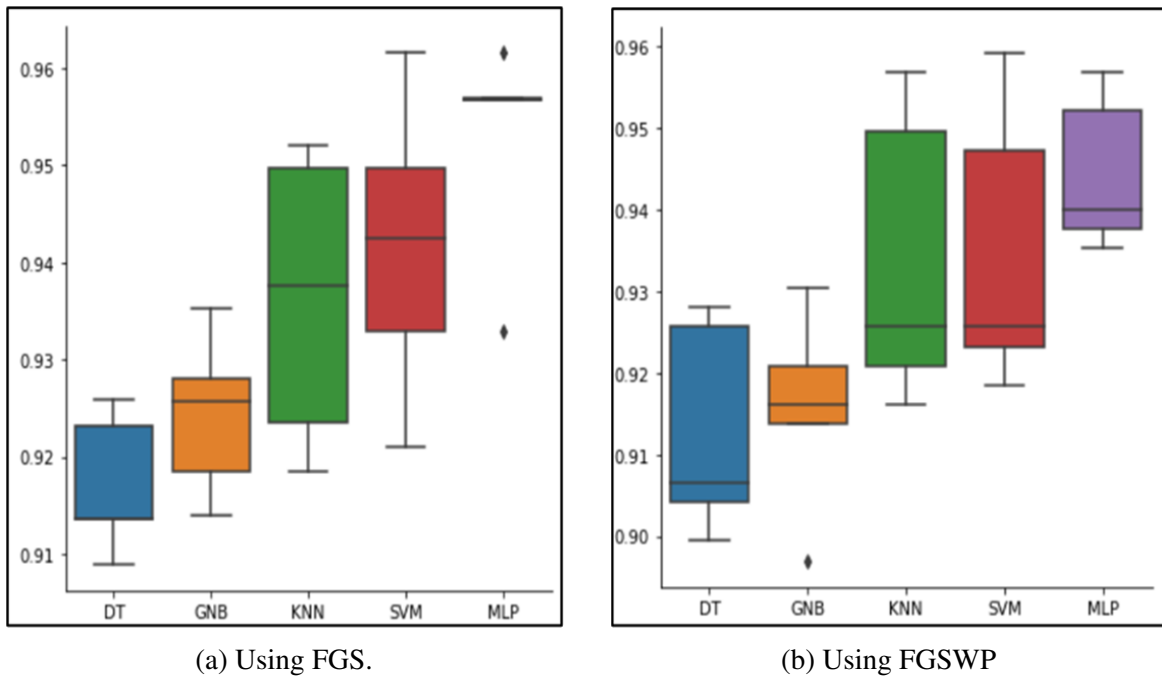


Fig. 5.33 A comparison of FGS vs FGSWP in five classifiers for (TCGA1)

Fig. 5.34 presents a comparison between FGS and FGSWP for classifying lung cancer (GSE10072) using  $k=5$ . Despite a significant reduction in the number of genes from 52 with FGS to 5 with FGSWP, the accuracy of the majority of employed classifiers remained unchanged. These results indicate that the developed FGSWP successfully achieved its objective of reducing the number of genes while maintaining the same level of accuracy.

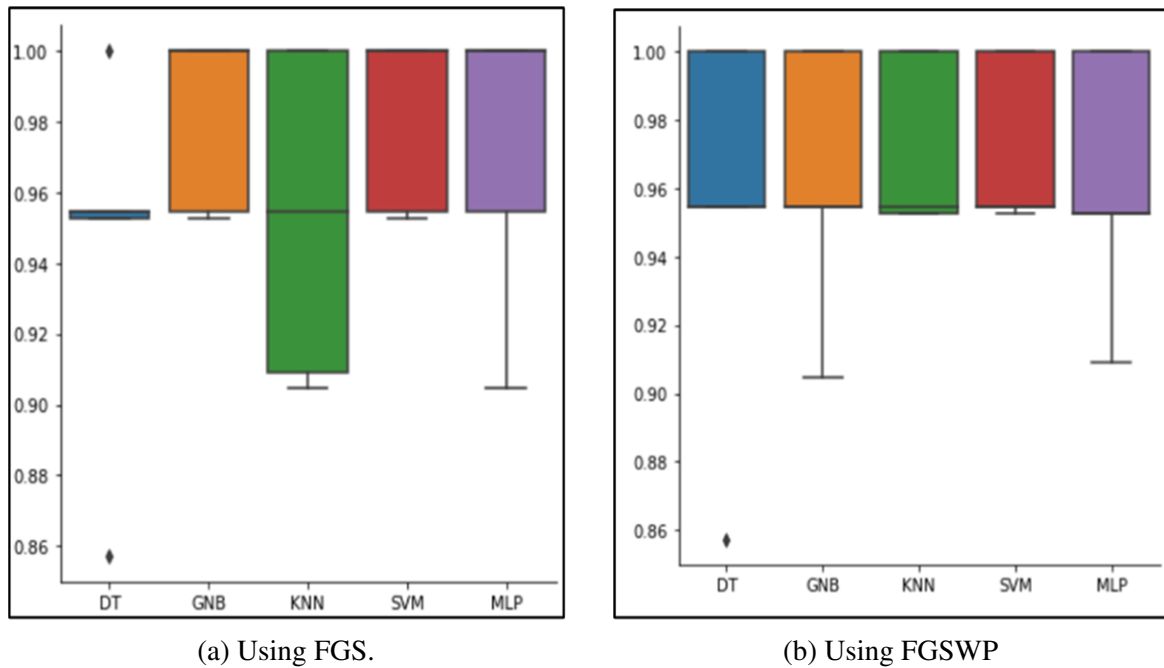


Fig. 5.34 A comparison of FGS vs FGSWP in five classifiers for (GSE10072)

### 5.4.3 Results of FGSWP with FC

FGSW demonstrates that it is successful to meet the goal of its development when employing the FC method. Using FGSWP led to keeping or improving the accuracy with less number of genes compared to the FGS method in all employed datasets. This experiment illustrates the effectiveness of using FGSWP with the FC method. The results indicate that FGSWP highly reduces the number of genes for the majority of employed datasets. Particularly with GSE33630 from 76 to 17, GSE33630 from 68 to 30, GSE66499 from 150 to 99, TCGA4 from 298 to 201 and GSE10072 from 52 to 5. While the rest of the other datasets, the number of genes was reduced, but at a lower rate compared to mentioned datasets. In summary, the developing FGSWP reduced the number of genes selected by FGS for all employed datasets in different proportions while keeping the accuracy for the majority of datasets used. Additionally, FGSWP improved the results with lung cancer (GSE43580) data. The results were 91%, 94%, 89%, and 90 for accuracy, precision, recall, and f1-score, respectively when FGS, while the results were 93%, 95%, 92%, 93% for accuracy, precision, recall, and f1-score respectively when FGSWP used as described in Table 5.6.

Table 5.6 A comparison of FGS vs FGSWP using FC method.

Dataset	FGS	FGSWP	Classifier	Old AC %	New AC %	Old Pre %	New Pre %	Old Rec %	New Rec %	Old F1 %	New F1 %
GSE53757	78	69	FC	100	100	100	100	100	100	100	100
GSE66499	150	99	FC	95	95	95	95	92.4	93	93.4	94
GSE84437	105	98	FC	92.8	93	95.6	95.7	81.4	83	85	86
GSE14520	23	17	FC	99	99	98	98	99	99	98	98
GSE19804	36	28	FC	100	100	100	100	100	100	100	100
TCGA2	116	101	FC	97	97.3	98	98	97	97	97	97.2
GSE33630	76	17	FC	100	100	100	100	100	100	100	100
TCGA6	28	22	FC	98	98	99	99	98	98	98	98
GSE45827	68	30	FC	100	100	100	100	100	100	100	100
TCGA1	25	18	FC	97	96	95	95	95	94	95	94
GSE10072	52	5	FC	100	100	100	100	100	100	100	100
TCGA4	298	201	FC	98.6	99	98	97	98	99	98	98
GSE43580	28	8	FC	91	93	94	95	89	92	90	93
TCGA7	11	8	FC	99	99	98.7	98.7	98.8	98.8	98.7	98.7
GSE77314	12	10	FC	99	99	98	98	100	100	99	99
TCGA5	15	10	FC	99	99	99	100	97	100	98.7	99

#### 5.4.4 Comparison of Findings

The section describes the comparison of the developed Multidimensional Fuzzy Deep Learning (MDFDL) model against the other published works using the same cancer expression data. The findings demonstrated that the developed model surpasses over other published works using the same datasets in terms of evaluation metrics or the number of genes used to train the models.

In summary, the developed model (MDFDL) shows that, when compared to other studies, it is capable of significantly increasing the accuracy of cancer classification while reducing the number of genes involved. According to the results, MDFDL performed much better than earlier research that analysed these datasets using different classifier and feature selection approaches. With all datasets used by previous research, the accuracy was between 70% and 99%; however, the accuracy was between 95% and 100% for all datasets using the developed MDFDL. Furthermore, the number of genes ranged from 33298 to 10 in previous studies, while, from 150 to 7 with the developed model. In terms of accuracy achieved and the number of chosen genes used to train the classifier algorithms, the developed model provides advantages over earlier studies. Enhancing the accuracy, less classifier complexity, saving training time, and reducing the overfitting issue are the considerable contributions of

selecting an optimal subset of genes. The outcomes of comparing the developed model to other published works are described in Table 5.7.

Table 5.7 Comparing the developed model against prior studies .

Dataset	No.Genes	Approaches	Ac %	Pre %	Rec %	F1%	Reference
GSE66499	33298	CNN	81	88	78	74	[129]
	150	MDFDL	95	95	92	93	The proposed work
GSE43580	43	MCSF +RF	88	82	97	89	[102]
	8	MDFDL	95	95	92	93	The proposed work
GSE33630	17	mRMR+KNN	91.3	No	No	No	[110]
	No	PCA + RF	92	92	89	No	[167]
	17	MDFDL	100	100	100	100	The proposed work
GSE19804	10	HLR+SVM	94	No	No	No	[163]
	11	MDFDL	100	100	100	100	The proposed work
GSE14520	1253	DRE-DNN	82	83.3	95	88.9	[137]
	23	MDFDL	96	96	96	96	The proposed work
TCGA6	768	DRE-DNN	70	77.3	70.8	73.9	[137]
	7	MDFDL	100	100	100	100	The proposed work
GSE10072	19	ReliefF+NB	95	No	No	No	[100]
	5	MDFDL	100	100	100	100	The proposed work
TCGA5	67	ReliefF +RF	83.6	No	No	No	[114]
	12	MI+RFE+SVM+RF	97.9	97.6	97.7	97.6	[168]
	194	KL divergence+DNN	99	98	100	No	[169]
	10	MDFDL	99	100	97	98	The proposed work
TCGA2	No	CNN	88.4	88.3	88.4	88.2	[140]
	No	gcForest	92	No	No	No	[109]
	116	MDFDL	97	98	97	97	The proposed work
TCGA1	971	BPSO-DT+CNN	96	94.96	95	95	[164]
	18	MDFDL	96	95	94	94	The proposed work
GSE45827	20	CFS+NB	89.7	92	89.7	90	[165]
	38	Rough set +SVM	96.86	96.9	97.34	97.8	[166]
	30	MDFDL	100	100	100	100	The proposed work
TCGA7	49	GGA+ELM	98.8	No	No	No	[170]
	No	Intersection+SVM	94.4	78	83.5	80.9	[171]
	11	MDFDL	100	100	100	100	The proposed work

### 5.4.5 Synthesis of Findings

This chapter presented three experiments to evaluate the developed approaches (FGS, FGSWP, and FC) and compared integrated all these methods in one model called MDFDL with the previous studies. Additionally, this chapter involves the initial experiment of the thesis that was done using FTFS. The findings demonstrated that these approaches have achieved

promising results, both individually and in combination in one model. The accuracy scores of employing FTFS with MLP were ranging from 95.5% to 100% for six cancer expressions data 70% of the data for training and 30% of the data for testing. Additionally, the number of selected genes were ranging from 99 to 31 for the six cancer expression data employed. Even though FTFS achieved promising findings in terms of evaluation metrics and the number of genes, however, it has some limitations that were identified previously (Chapter 3 section 3.1). To overcome these limitations, FGS was developed. FGS was evaluated by employing sixteen cancer expression data using k cross-validation with k=5 rather than splitting the data into training and testing randomly. Cross-validation provides a more realistic evaluation of how well a model will perform on unseen data and assist ensure that the developed model is capable of generalising beyond the training data. FGS reduced the number of genes in all employed datasets as well as it successfully improved the results for twelve datasets when used with classical classifiers. While the rest datasets were improved, however, the accuracy is still poor. Even though FGS reduced the number of genes with these datasets. The average results for the successfully improved datasets were 97%, 97.3%, 96.5%, and 96.77% for accuracy, precision, recall, and f1-score respectively.

Although, FGS reduced the number of genes successfully with all employed datasets and improved the results for most data used. However, there are some datasets even though the number of genes lowered and the accuracy increased, they are still under the level required when FGS and classical classifiers were used. Additionally, there is no classifier that can continuously achieve the highest accuracy in all employed datasets. To overcome these challenges, the FC method was developed. To show the effectiveness of using FC, twelve datasets were used including the datasets that were not successfully classified when FGS and classical classifiers were used. FC demonstrated that it continuously achieved the highest results in all employed datasets compared to classical classifiers. FC reached average results were 98%, 98.2%, 96.5%, and 97% for accuracy, precision, recall, and f1-score respectively for all employed datasets.

This chapter also includes the experiment of using FGSWP in six datasets to more reducing the number of genes that were selected by FGS and keep the accuracy the same or improve. The findings indicate that FGSWP greatly reduced the number of genes for all employed datasets. FGSWP was evaluated using FC and classical classifiers. The results in both demonstrated that FGSW kept the accuracy and other evaluation metrics the same when FGS used with a lower number of genes for most employed datasets. Furthermore, it slightly enhanced the accuracy of some datasets.

The final experiment was comparing the previously published works against the combination of all developed approaches ( FGSWP and FC) in one model namely multidimensional

fuzzy deep learning (MDFDL). MDFDL was compared to prior studies using the same datasets and different approaches. MDFDL demonstrated that outperformed previous studies that used the same datasets in terms of selected genes and evaluation metrics.

# Chapter 6

## General Discussion and Future Directions

Cancer is a group of diseases with considerable morbidity and mortality and some tumour types have increased dramatically in incidence due to smoking, drinking, poor diet, physical inactivity, and air pollution all of which are risk factors [172]. Whilst treatment option has improved somewhat, late-stage cancer patients have a particularly poor prognosis. For these reasons, efforts are concentrated on developing effective novel therapies and developing new technologies for accurate diagnosis. As a computer science thesis, the project focused on approaches that have been developed for accurate diagnosis, which may assist scientists to speed up their job. Based on this, machine learning techniques and statistical approaches (Feature selection methods) were used to analyse data derived from different cancer, the most notable of which are images (CT scan and MRI) and, more recently, gene expression data. ML approaches have trained the model with retrospective data and then evaluated it prospectively to see how well the model was trained. While feature selection approaches try to choose an important subset of genes that are substantially informative for distinguishing differences across classes (i.e. normal and cancer tissue).

In this study, cancer expression data were analysed to obtain reliable cancer classification using a subset of genes and to mitigate the technical limitations that have occurred in prior studies. Microarray and RNA-seq approaches are two commonly used technologies for measuring the expression of each gene in tissue from normal and malignant diseases. Both tools have the same purpose, which is to identify the extent of gene expression across thousands of genes, but there is a technical difference between them. This thesis used both tools to evaluate the developed approaches. The freely available gene expression data are distinguished by a limited number of samples with a large number of genes, i.e., high dimensional datasets [173]. As a result, the probability of overfitting, the complexity of a classifier, and the time required throughout the training stage are all quite high. These limitations have a negative impact on the performance of any classifier. It becomes necessary to handle these drawbacks

by developing gene selection methods that reduce the dimensionality, the complexity of the classifier, and the time required during the training stage. Furthermore, our experiments indicate that no one classifier can continuously obtain the best accuracy for all provided datasets. However, each classifier can achieve superior accuracy for specific types of data.

Cancer expression data is characterised by high dimensionality which is a small number of samples and a high number of genes [174]. That leads to an overfitting issue, complexity, and time-consuming for a classifier model and inaccurate cancer classification [175]. Accordingly, high dimensionality and inaccurate cancer classification are the challenges that this thesis aims to overcome. To achieve that, four approaches were developed (FTFS, FGS, FGSWP, and FC). FTFS method was developed to reduce the number of genes that are used as identifiers for training a classifier. Even though it produced good results in terms of reducing the number of genes and increasing the accuracy, however, it has some limitations (explained in the methodology chapter section 3.1). To avoid these limitations, a novel FGS method was developed to select informative genes.

The result indicates that the FGS method presented promising results in terms of reducing the number of genes and improving the performance of the classifiers for the majority of employed datasets. However, it achieved poor accuracy for some datasets when applied to classical classifiers even though reduce the number of genes. Additionally, our experiment indicated that no one classifier can achieve the highest results continuously. For example, GNB had the highest results for kidney cancer data (GSE53757) while, MLP accomplished the highest results for lung cancer (GSE19804), and five cancer types (TCGA1). Additionally, GNB and KNN had the highest results for liver cancer (GSE14520) while SVM and KNN achieved the highest results for thyroid cancer (GSE33630). The FGS method with classical classifiers achieved poor results with these datasets (GSE84437, GSE43580, GSE66499, and TCGA6). Consequently, an FC was developed to generalise the classifier so that it can achieve the best accuracy across all provided datasets, and a fuzzy gene selection wrapper plus was tested to reduce the number of genes without lowering the accuracy. Briefly, the thesis developed the FTFS approach and evaluated it using classical classifiers as a preliminary experiment. Then, the thesis developed three efficient methods (FGS+FGSWP+FC) and combined them into a single model termed MDFDL.

## 6.1 Summary of Findings

### 6.1.1 Implications of FTFS Findings

FTFS was developed to reduce the number of genes and enhance the performance of classifiers. The results demonstrated that FTFS reduced the number of genes for the six datasets employed. For the six cancer expression data used, the number of chosen genes varied from 99 to 31 rather than thousands of genes in the original data. Moreover, the results were improved for most of the data used when compared to omitting using FTFS. Notably, when FTFS and MLP were used together. The results ratings for six cancer expressions ranged from (95.5% to 100%) accuracy, (94.4% to 100%) precision, (94% to 100%) recall, and (95.7% to 100%) f1-score, with 70% of the data used for training and 30% used for testing when FTFS and MLP were used. Despite having obtained encouraging results in terms of evaluation metrics and the number of genes, FTFS has certain previously noted shortcomings (Chapter 3 section3.1).

### 6.1.2 Implications of FGS Findings

FGS was developed to make cancer expression datasets less dimensional and to boost the classifier's performance. The development of FGS was evaluated using sixteen cancer expression datasets with classical classifiers. The results indicated that FGS highly reduced the number of genes (less dimensional ) across all employed datasets. Furthermore, FGS greatly improved the evaluation metrics for twelve datasets, while the remaining datasets even if the results improved, however, they are still poor. The average results for the successfully improved datasets (twelve) were 97%,97.3%, 96.5%, and 96.77% for accuracy, precision, recall, and f1-score respectively. Although, FGS reduced the number of genes successfully with all employed datasets and improved the results for most data used. However, there are some datasets even though the number of genes lowered and the accuracy increased, they are still under the level required when FGS and classical classifiers were used. Additionally, there is no classifier that can continuously achieve the highest accuracy in all employed datasets. For example, GNB had the highest results for kidney cancer data (GSE53757) while, MLP accomplished the highest results for lung cancer (GSE19804), and five cancer types (TCGA1). Additionally, GNB and KNN had the highest results for liver cancer (GSE14520) while SVM and KNN achieved the highest results for thyroid cancer (GSE33630). The FGS method with classical classifiers achieved poor results with these datasets (GSE84437, GSE43580, GSE66499, and TCGA6).

### 6.1.3 Implications of FC Findings

FC was developed to accurately classify cancer expression data for different types of cancer and continuously achieved the highest results for all given datasets. On the other hand, it aims to generalise the classifier rather than each classifier accomplishing the best accuracy for a particular dataset. It also aims to enhance the results for those datasets that were unable FGS and classical classifiers to accurately classify them. FC demonstrated that it continuously achieved the highest results in all (twelve) employed datasets compared to classical classifiers. FC reached average results were 98%, 98.2%, 96.5%, and 97% for accuracy, precision, recall, and f1-score respectively for all employed datasets. Notable improvements were with these datasets ( GSE43580, GSE84437, GSE66499, TCGA1, and TCGA6) when FC was employed compared to classical classifiers as described in Table5.4. Consequently, the findings indicate that FC successfully achieved the goal of its development.

### 6.1.4 Implications of FGSWP Findings

FGSPW was developed to reduce the number of selected genes using FGS and maintains the accuracy and other evaluation metrics as the same as previously or improve them. Six cancer expression data were used to evaluate the FGSWP using classical classifiers and the FC method. The findings indicate that FGSWP reduced the number of genes for all employed datasets and kept the accuracy and other evaluation metrics when classical classifiers and FC were used for most employed datasets. Furthermore, it slightly improved the achievement results of some datasets such as. More importantly, the number of genes high decreased in some datasets without impact on the results. For example, when the following datasets were used, the number of genes selected using FGS was significantly higher than when FGSWP was used GSE33630: 76 genes vs. 17 genes, GSE53757: 78 genes vs. 69 genes, GSE45827: 68 genes vs. 30 genes, TCGA1: 25 genes vs. 18 genes, GSE10072: 52 genes vs. 5 genes, and GSE43580: 28 genes vs. 8 genes. The results showed that FGSWP was able to reduce the number of genes selected by up to 82% without sacrificing accuracy and other evaluation metrics.

### 6.1.5 Implications of MDFDL Findings

To ensure that the developed model produces better results when compared to earlier studies that used the same datasets. In the thesis, all developed approaches (FGS, FGSWP, and FC) were combined into the multidimensional fuzzy deep learning (MDFDL) model. When MDFDL was compared to fourteen other published works that used the same datasets, it

showed that it had produced better results. It was demonstrated through comparison that MDFDL improved in terms of reducing the number of genes used for training and evaluation metrics as shown in Table 5.7.

## **6.2 Future Directions**

### **6.2.1 Integrating different inputs**

The discriminate performance of current learning algorithms can be enhanced by adding further input characteristics, such as DNA methylations and mutations. Indeed, the complete influence of gene expression can not be represented by the genetic sequence alone [173]. Consequently, combining methylations and mutations with RNA-Seq data can result in characteristics that help with tumour classification.

### **6.2.2 Advancements in cancer biomarkers**

Developing advanced approaches to identify biomarkers for types of cancer is a key future direction in the study of cancer-related biomarkers [176]. For instance, doing functional pathway analysis of related genes for the various cancer types can be aided by the techniques provided for IntPath [177] and others [178]. Focusing on genetic and transcriptome changes may be helpful in distinguishing tumours that have similar clinical characteristics. The omics data include specifics on several approaches for ovarian and other malignancies, as well as renal cell carcinoma, for finding important genes and pathways that may help with prognostic and diagnostic predictions. For a better understanding of the prognosis for different cancer datasets, optical genome mapping and structural variant analysis (at a region of DNA known as copy number variations, which might contain inversions, balanced translocations, or genomic imbalances) may be used [176].

### **6.2.3 Interpretable ML**

It is crucial to emphasise the development of interpretable machine learning models, which aid in understanding the decision-making process by the used mathematical techniques and offer reasons for instances in which the models may falter [176]. Increased focus in this area should be placed on interpretable and explicable models that emphasise the local and global features of ML models based on counterfactual or feature attribution.

# References

- [1] Sean Blandin Knight, Phil A Crosbie, Haval Balata, Jakub Chudziak, Tracy Hussell, and Caroline Dive. Progress and prospects of early detection in lung cancer. *Open biology*, 7(9):170070, 2017.
- [2] Alper Emre Celik, Jawad Rasheed, and Amani Yahyaoui. Machine learning approaches for lung cancer prediction. In *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 540–543, 2022.
- [3] Shilpi Shandilya and Chaitali Chandankhede. Survey on recent cancer classification systems for cancer diagnosis. *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2590–2594, 2017.
- [4] Jayadeep Pati. Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach. *IEEE Access*, 7:4232–4238, 2019.
- [5] Souvik Ghatak, Syrina F. Mehrabi, Lubna M. Mehdawi, Shakti Ranjan Satapathy, and Anita Sjölander. Identification of a novel five-gene signature as a prognostic and diagnostic biomarker in colorectal cancers. *International Journal of Molecular Sciences*, 23(2), 2022.
- [6] Ramazani A Bodaghi A, Fattahi N. Biomarkers: Promising and valuable tools towards diagnosis, prognosis and treatment of covid-19 and other diseases. *Heliyon*, 9(2), 2023.
- [7] S. Vanjimalar, D. Ramyachitra, and P. Manikandan. A review on feature selection techniques for gene expression data. In *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–4, 2018.
- [8] Esra’a Alhenawi, Rizik Al-Sayyed, Amjad Hudaib, and Seyedali Mirjalili. Feature selection methods on gene expression microarray data for cancer classification: A systematic review. *Computers in Biology and Medicine*, 140:105051, 2022.
- [9] Luca Zanella, Pierantonio Facco, Fabrizio Bezzo, and Elisa Cimetta. Feature selection and molecular classification of cancer phenotypes: A comparative study. *International Journal of Molecular Sciences*, 23(16):1422–0067, 2022.
- [10] Thomas Rincy N and Roopam Gupta. Feature selection techniques and its importance in machine learning: A survey. In *2020 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6, 2020.

- [11] Zhuo Wang, Huan Li, Bin Nie, Jianqiang Du, Yuwen Du, and Yufeng Chen. Feature selection using different evaluate strategy and random forests. In *2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, pages 310–313, 2021.
- [12] Souvik Sarkar, Sidhant Singh, Rohit Kumar, and Debraj Chatterjee. Comparison between different machine learning algorithms. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, volume 6, pages 240–244, 2021.
- [13] W.J. Zhang, Guosheng Yang, Yingzi Lin, Chunli Ji, and Madan M. Gupta. On definition of deep learning. In *2018 World Automation Congress (WAC)*, pages 1–5, 2018.
- [14] Bengio Y. Hinton LeCun, Y. Deep learning. *Nature*, (521):436–444, 2015.
- [15] Wan Zhu, Longxiang Xie, Jianye Han, and Xiangqian Guo. The application of deep learning in cancer prognosis prediction. *Cancers*, 12(3), 2020.
- [16] Fadoua Rafii, M’hamed Aït Kbir, and Badr Dine Rossi Hassani. Mlp network for lung cancer presence prediction based on microarray data. In *2015 Third World Conference on Complex Systems (WCCS)*, pages 1–6, 2015.
- [17] Laiqa Rukhsar, Waqas Haider Bangyal, Muhammad Sadiq Ali Khan, Ag Asri Ag Ibrahim, Kashif Nisar, and Danda B. Rawat. Analyzing rna-seq gene expression data using deep learning approaches for cancer classification. *Applied Sciences*, 12(4), 2022.
- [18] Ahmed Fawzy Gad, Ahmed Fawzy Gad, and Suresh John. *Practical computer vision applications using deep learning with CNNs*. Springer, 2018.
- [19] Zhang J. Humaidi A.J. et al Alzubaidi, L. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53), 2021.
- [20] Diyar Qader Zeebaree, Habibollah Haron, and Adnan Mohsin Abdulazeez. Gene selection and classification of microarray data using convolutional neural network. In *2018 International Conference on Advanced Science and Engineering (ICOASE)*, pages 145–150, 2018.
- [21] Halah Almazrua and Hala Alshamlan. A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data. *IEEE Access*, 10:71427–71449, 2022.
- [22] Nivedhitha Mahendran, PM Durai Raj Vincent, Kathiravan Srinivasan, and Chuan-Yu Chang. Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. *Frontiers in genetics*, 11:603808, 2020.
- [23] Sarah Osama, Hassan Shaban, and Abdelmgeid A. Ali. Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. *Expert Systems with Applications*, 213:118946, 2023.

- [24] Lin Sun, Xiaoyu Zhang, Yuhua Qian, Jiucheng Xu, and Shiguang Zhang. Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. *Information Sciences*, 502:18–41, 2019.
- [25] Zhang Yaping and Zhou Changyin. Gene feature selection method based on relieff and pearson correlation. In *2021 3rd International Conference on Applied Machine Learning (ICAML)*, pages 15–19, 2021.
- [26] Waleed Ali and Faisal Saeed. Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. *Processes*, 11(2), 2023.
- [27] G Manikandan and S Abirami. A survey on feature selection and extraction techniques for high-dimensional microarray datasets. *Knowledge Computing and its Applications: Knowledge Computing in Specific Domains: Volume II*, pages 311–333, 2018.
- [28] Nivedhitha Mahendran, PM Durai Raj Vincent, Kathiravan Srinivasan, and Chuan-Yu Chang. Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. *Frontiers in genetics*, 11:603808, 2020.
- [29] Gregg B. Whitworth. Chapter 2 - an introduction to microarray data analysis and visualization. In *Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis*, volume 470 of *Methods in Enzymology*, pages 19–50. Academic Press, 2010.
- [30] Krishanpal Anamika, Srikant Verma, Abhay Jere, and Aarti Desai. Transcriptomic profiling using next generation sequencing-advances, advantages, and challenges. *Next generation sequencing-advances, applications and challenges*, 9:7355–7365, 2016.
- [31] Kaliyappan K Palanisamy M. Govindarajan R, Duraiyan J. Microarray and its applications. *Pharm Bioallied Sci*, 4:310–312, 2012.
- [32] Ciurlionis R Buck WR Mittelstadt SW Blomme EAG Liguori MJ Rao MS, Van Vleet TR. Comparison of rna-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Front Genet.*, 9, 2019.
- [33] Yongjun Piao and Keun Ho Ryu. Detection of differentially expressed genes using feature selection approach from rna-seq. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 304–308, 2017.
- [34] Li J Ye F Samuels DC Shyr Y Guo Y, Sheng Q. Large scale comparison of gene expression levels by microarrays and rnaseq using tcga data. *PLoS ONE*, 8(8), 2013.
- [35] Motieghader H. Masoudi-Nejad A Masoudi-Sobhanzadeh, Y. Featureselect: a software for feature selection based on machine learning approaches. *BMC Bioinformatics*, 20(170), 2019.

- [36] Amandeep Kaur, Kalpna Guleria, and Naresh Kumar Trivedi. Feature selection in machine learning: Methods and comparison. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 789–795, 2021.
- [37] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.
- [38] Zhang Y Liu J Yu B Liu X Dehmer M Liu S, Xu C. Feature selection of gene expression data for cancer classification using double rbf-kernels. *BMC Bioinformatics*, 19(396), 2018.
- [39] Jianyu Miao and Lingfeng Niu. A survey on feature selection. *Procedia computer science*, 91:919–926, 2016.
- [40] Silvia Cateni, Valentina Colla, and Marco Vannucci. A hybrid feature selection method for classification purposes. In *2014 European Modelling Symposium*, pages 39–44, 2014.
- [41] Majed A. Alenizi and Hussein Y. Abu Mansour. A new intelligent hybrid feature selection method. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6, 2018.
- [42] A. Jović, K. Brkić, and N. Bogunović. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, 2015.
- [43] C. E. SHANNON. A mathematical theory of communication. *The Bell System Technical*, 17:379–423, 2000.
- [44] Hongfang Zhou. Feature selection based on mutual information with correlation coefficient. *Applied Intelligence*, 52:5457–5474, 2022.
- [45] Nimrita Koul and Sunilkumar S Manvi. Ensemble feature selection from cancer gene expression data using mutual information and recursive feature elimination. In *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pages 1–6, 2020.
- [46] Nadir Omer Fadl Elssied, Othman Ibrahim, and Ahmed Hamza Osman. A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7:625–638, 2014.
- [47] Yujia Zhai, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. A chi-square statistics based feature selection method in text classification. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 160–163, 2018.
- [48] Sujan Ray, Khaldoon Alshouli, Anupam Roy, Ali AlGhamdi, and Dharma P. Agrawal. Chi-squared based feature selection for stroke prediction using azureml. In *2020 Intermountain Engineering, Technology and Computing (IETC)*, pages 1–6, 2020.

- [49] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.
- [50] Peng H., J Ding C. Minimum redundancy feature selection from microarray gene expression data. *Bioinform Comput Biol*, 3(2):185–205, 2005.
- [51] Ghalwash M. Filipovic N. et al Radovic, M. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(9), 2017.
- [52] Qiyu Wang. Support vector machine algorithm in machine learning. In *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 750–756, 2022.
- [53] Fereshteh Falah Chamasemani and Yashwant Prasad Singh. Multi-class support vector machine (svm) classifiers – an application in hypothyroid detection and classification. In *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, pages 351–356, 2011.
- [54] Aashi Maharjan. *Machine Learning Approach for Predicting Cancer Using Gene Expression*. PhD thesis, University of Nevada, Las Vegas, 2020.
- [55] Muhammad Ali Farooq, Peter Corcoran, Cosmin Rotariu, and Waseem Shariff. Object detection in thermal spectrum for advanced driver-assistance systems (adas). *IEEE Access*, 9:156465–156481, 2021.
- [56] Laura Auria and Rouslan A Moro. Support vector machines (svm) as a technique for solvency analysis. 2008.
- [57] Lucian-Ovidiu Fedorovici and Florin Dragan. A comparison between a neural network and a svm and zernike moments based blob recognition modules. In *2011 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 253–258, 2011.
- [58] Chapter 6 - machine learning for soil moisture assessment. In Ramesh Chandra Poonia, Vijander Singh, and Soumya Ranjan Nayak, editors, *Deep Learning for Sustainable Agriculture*, Cognitive Data Science in Sustainable Computing, pages 143–168. Academic Press, 2022.
- [59] Mona Al Hamad and Ahmed M. Zeki. Accuracy vs. cost in decision trees: A survey. In *2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 1–4, 2018.
- [60] Kurebayashi R. Kobayashi Ohta, S. Minimizing false positives of a decision tree classifier for intrusion detection on the internet. *Journal of Network and Systems Management*, 16:399–419, 2008.
- [61] Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24:175–186, 2014.

- [62] Bahzad Charbuty and Adnan Mohsin Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2021.
- [63] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995.
- [64] Baoxun Xu, Joshua Zhexue Huang, Graham Williams, Qiang Wang, and Yunming Ye. Classifying very high-dimensional data with random forests built from small subspaces. 8(2), 2012.
- [65] Yueyue Xiao, Wei Huang, and Jinsong Wang. A random forest classification algorithm based on dichotomy rule fusion. In *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 182–185, 2020.
- [66] Thanh-Tung Nguyen, Joshua Zhexue Huang, and Thuy Thi Nguyen. Unbiased feature selection in learning random forests for high-dimensional data. *The Scientific World Journal*, 2015, 2015.
- [67] Dahai Zhang, Liyang Qian, Baijin Mao, Can Huang, Bin Huang, and Yulin Si. A data-driven design for fault detection of wind turbines using random forests and xgboost. *IEEE Access*, 6:21020–21031, 2018.
- [68] Hartatik, A Purnomo, R Hartono, and H Munawaroh. Naïve bayes approach for expert system design of children skin identification based on android. *IOP Conference Series: Materials Science and Engineering*, 333(1):105–112, mar 2018.
- [69] Kumar V. Ross Quinlan J. et al Wu, X. Top 10 algorithms in data mining. *Knowl Inf Syst*, 14:1–37, 2008.
- [70] Ali Haghpanah Jahromi and Mohammad Taheri. A non-parametric mixture of gaussian naive bayes classifiers based on local independent features. In *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, pages 209–212, 2017.
- [71] Berke Akkaya and Nurdan Çolakoğlu. Comparison of multi-class classification algorithms on early diagnosis of heart diseases. 2019.
- [72] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. *UAI'95 Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.
- [73] N. Sebe, M.S. Lew, I. Cohen, A. Garg, and T.S. Huang. Emotion recognition using a cauchy naive bayes classifier. In *2002 International Conference on Pattern Recognition*, volume 1, pages 17–20 vol.1, 2002.
- [74] Kevin P Murphy et al. Naive bayes classifiers. *University of British Columbia*, 18(60):1–8, 2006.
- [75] Ali Haghpanah Jahromi and Mohammad Taheri. A non-parametric mixture of gaussian naive bayes classifiers based on local independent features. In *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, pages 209–212, 2017.

- [76] Kashvi Taunk, Sanjukta De, Srishti Verma, and Aleena Swetapadma. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1255–1260, 2019.
- [77] Aireza Naser Sadrabadi, Seyed Mahmood Znjirchi, Habib Zare Ahmad Abadi, and Ahmad Hajimoradi. An optimized k-nearest neighbor algorithm based on dynamic distance approach. In *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–7, 2020.
- [78] Alka Rani. *Recent Trends in Computational Intelligence Enabled Research*. ScienceDirect, 2021.
- [79] Sadegh Bafandeh Imandoust and Mohammad Bolandraftar. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. 2013.
- [80] Pdraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers-a tutorial. *ACM computing surveys (CSUR)*, 54(6):1–25, 2021.
- [81] Rui Xie, Jia Wen, Andrew Quitadamo, Jianlin Cheng, and Xinghua Shi. A deep auto-encoder model for gene expression prediction. *BMC genomics*, 18:39–49, 2017.
- [82] Xiaonan Zou, Yong Hu, Zhewen Tian, and Kaiyuan Shen. Logistic regression model optimization and case analysis. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pages 135–139, 2019.
- [83] R Harine Rajashree and M Hariharan. A study on ensemble methods for classification. In *Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication: Proceedings of MDCWC 2020*, pages 127–136. Springer, 2021.
- [84] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1):41–51, 2018.
- [85] Jagpreet Chhatwal, Oguzhan Alagoz, Mary J Lindstrom, Charles E Kahn Jr, Katherine A Shaffer, and Elizabeth S Burnside. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *AJR. American journal of roentgenology*, 192(4):1117, 2009.
- [86] K Nagaiah, K Madan Mohan, and Mohan Aswala. Efficient feature selection approach for breast cancer classification using machine learning. In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–5, 2023.
- [87] Fadi Alharbi and Aleksandar Vakanski. Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10(2):173, 2023.
- [88] Jelmar Quist, Lawson Taylor, Johan Staaf, and Anita Grigoriadis. Random forest modelling of high-dimensional mixed-type data for breast cancer classification. *Cancers*, 13(5):991, 2021.

- [89] Sarah M. Ayyad, Ahmed I. Saleh, and Labib M. Labib. Gene expression cancer classification using modified k-nearest neighbors technique. *Biosystems*, 176:41–51, 2019.
- [90] Cesare Alippi and Manuel Roveri. Virtual k-fold cross validation: An effective method for accuracy assessment. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2010.
- [91] Ch Anwar Ul Hassan, Muhammad Sufyan Khan, and Munam Ali Shah. Comparison of machine learning algorithms in data classification. In *2018 24th International Conference on Automation and Computing (ICAC)*, pages 1–6, 2018.
- [92] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 01 2002.
- [93] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
- [94] Husna Ayardenta et al. A clustering approach for feature selection in microarray data classification using random forest. *Journal of Information Processing Systems*, 14(5):1167–1175, 2018.
- [95] Mehrdad Rostami, Saman Forouzandeh, Kamal Berahmand, Mina Soltani, Meisam Shahsavari, and Mourad Oussalah. Gene selection for microarray data classification via multi-objective graph theoretic-based method. *Artificial Intelligence in Medicine*, 123:102228, 2022.
- [96] Jiande Wu and Chindo Hicks. Breast cancer type classification using machine learning. *Journal of personalized medicine*, 11(2):61, 2021.
- [97] Aiguo Wang, Ning An, Guilin Chen, Jing Yang, Lian Li, and Gil Alterovitz. Incremental wrapper based gene selection with markov blanket. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 74–79. IEEE, 2014.
- [98] C. Arun Kumar and S. Ramakrishnan. Binary classification of cancer microarray gene expression data using extreme learning machines. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–4, 2014.
- [99] Suleyman Vural, Xiaosheng Wang, and Chittibabu Guda. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC systems biology*, 10(3):263–276, 2016.
- [100] Arturo López Pineda, Henry Ato Ogoe, Jeya Balaji Balasubramanian, Claudia Rangel Escareño, Shyam Visweswaran, James Gordon Herman, and Vanathi Gopalakrishnan. On predicting lung cancer subtypes using ‘omic’ data from tumor and tumor-adjacent histologically-normal tissue. *BMC cancer*, 16:1–11, 2016.
- [101] Cancer Genome Atlas Research Network et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543, 2014.

- [102] Fei Yuan, Lin Lu, and Quan Zou. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1866(8):165822, 2020.
- [103] Pourya Naderi Yeganeh and M. Taghi Mostafavi. Use of machine learning for diagnosis of cancer in ovarian tissues with a selected mrna panel. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2429–2434, 2018.
- [104] Sara Tarek, Reda Abd Elwahab, and Mahmoud Shoman. Gene expression based cancer classification. *Egyptian Informatics Journal*, 18(3):151–159, 2017.
- [105] Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297, 1999.
- [106] Yasser El-Manzalawy. Cca based multi-view feature selection for multi-omics data integration. In *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8, 2018.
- [107] Naiyang Guan, Xiang Zhang, Zhigang Luo, and Long Lan. Sparse representation based discriminative canonical correlation analysis for face recognition. In *2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 51–56. IEEE, 2012.
- [108] O. Arandjelović. Discriminative extended canonical correlation analysis for pattern set matching. *Machine Learning volume*, 94:353–370, 2014.
- [109] Jing Xu, Peng Wu, Yuehui Chen, and Li Zhang. Comparison of different classification methods for breast cancer subtypes prediction. In *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 91–96. IEEE, 2018.
- [110] Ji Z Liu H Liu Y Peng H Wu J Fan J Xu Y, Deng Y. Identification of thyroid carcinoma related genes with mrmr and shortest path approaches. *PLoS One*, 9(4), 2014.
- [111] Sujan Ray, Khaldoon Alshouli, Anupam Roy, Ali AlGhamdi, and Dharma P. Agrawal. Chi-squared based feature selection for stroke prediction using azureml. In *2020 Intermountain Engineering, Technology and Computing (IETC)*, pages 1–6, 2020.
- [112] Padideh Danaee, Reza Ghaeini, and David A Hendrix. A deep learning approach for cancer detection and relevant gene identification. In *Pacific symposium on biocomputing 2017*, pages 219–229. World Scientific, 2017.
- [113] Sara Haddou Bouazza, Nezha Hamdi, Abdelouhab Zeroual, and Khalid Auhmani. Gene-expression-based cancer classification through feature selection with knn and svm classifiers. In *2015 Intelligent Systems and Computer Vision (ISCV)*, pages 1–6. IEEE, 2015.
- [114] Ching T. Huang S. et al Li, J. Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics*, (16), 2015.

- [115] Dejun Zhang, Lu Zou, Xionghui Zhou, and Fazhi He. Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *IEEE Access*, 6:28936–28944, 2018.
- [116] Siyabend Turgut, Mustafa Dağtekin, and Tolga Ensari. Microarray breast cancer data classification using machine learning methods. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–3, 2018.
- [117] Maxim D Podolsky, Anton A Barchuk, Vladimir I Kuznetsov, Natalia F Gusarova, Vadim S Gaidukov, and Segrey A Tarakanov. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific journal of cancer prevention*, 17(2):835–838, 2016.
- [118] Zhihua Cai, Dong Xu, Qing Zhang, Jiexia Zhang, Sai-Ming Ngai, and Jianlin Shao. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems*, 11(3):791–800, 2015.
- [119] Hasseeb Azzawi, Jingyu Hou, Russul Alnni, and Yong Xiang. Sbc: A new strategy for multiclass lung cancer classification based on tumour structural information and microarray data. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 68–73, 2018.
- [120] Dhahbi J. Chen, J.W. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *scientific reports*, 11(13323), 2021.
- [121] M Jansi Rani and Durairaj Devaraj. Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification. *Journal of medical systems*, 43:1–11, 2019.
- [122] Sherry Bhalla, Kumardeep Chaudhary, Ritesh Kumar, Manika Sehgal, Harpreet Kaur, Suresh Sharma, and Gajendra PS Raghava. Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Scientific reports*, 7(1):44997, 2017.
- [123] Rabia Aziz, CK Verma, and Namita Srivastava. A novel approach for dimension reduction of microarray. *Computational biology and chemistry*, 71:161–169, 2017.
- [124] Daniel Urda, Leonardo Franco, and José M. Jerez. Classification of high dimensional data using lasso ensembles. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2017.
- [125] Azzawi H Xiang Y Alanni R, Hou J. Deep gene selection method to select genes from microarray datasets for cancer classification. *BMC Bioinformatics*, 20(1), 2019.
- [126] Kusumastuti Cahyaningrum, Adiwijaya, and Widi Astuti. Microarray gene expression classification for cancer detection using artificial neural networks and genetic algorithm hybrid intelligence. In *2020 International Conference on Data Science and Its Applications (ICoDSA)*, pages 1–7, 2020.

- [127] Dongdong Sun, Minghui Wang, and Ao Li. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3):841–850, 2018.
- [128] Anwer Mustafa Hilal, Areej A Malibari, Marwa Obayya, Jaber S Alzahrani, Mohammad Alamgeer, Abdullah Mohamed, Abdelwahed Motwakel, Ishfaq Yaseen, Manar Ahmed Hamza, Abu Sarwar Zamani, et al. Feature subset selection with optimal adaptive neuro-fuzzy systems for bioinformatics gene expression classification. *Computational Intelligence and Neuroscience*, 2022, 2022.
- [129] Teppei Matsubara, Tomoshiro Ochiai, Morihito Hayashida, Tatsuya Akutsu, and Jose C Nacher. Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles. *Journal of bioinformatics and computational biology*, 17(03):1940007, 2019.
- [130] Ali Muhamed Ali, Hanqi Zhuang, Ali Ibrahim, Oneeb Rehman, Michelle Huang, and Andrew Wu. A machine learning approach for the classification of kidney cancer subtypes using mirna genome data. *Applied Sciences*, 8(12):2422, 2018.
- [131] Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019.
- [132] Jing Xu, Peng Wu, Yuehui Chen, Qingfang Meng, Hussain Dawood, and Muhammad Murtaza Khan. A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. *IEEE Access*, 7:22086–22095, 2019.
- [133] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
- [134] Serhat Kilicarslan, Kemal Adem, and Mete Celik. Diagnosis and classification of cancer using hybrid model based on relieff and convolutional neural network. *Medical Hypotheses*, 137:0306–9877, 2020.
- [135] Guillermo Lopez-Garcia, Jose M Jerez, Leonardo Franco, and Francisco J Veredas. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PloS one*, 15(3):e0230536, 2020.
- [136] Daniel Urda, Julio Montes-Torres, Fernando Moreno, Leonardo Franco, and José M Jerez. Deep learning to analyze rna-seq gene expression data. In *Advances in Computational Intelligence: 14th International Work-Conference on Artificial Neural Networks, IWANN 2017, Cadiz, Spain, June 14-16, 2017, Proceedings, Part II 14*, pages 50–59. Springer, 2017.
- [137] Junyi Li, Yuan Ping, Hong Li, Huinian Li, Ying Liu, Bo Liu, and Yadong Wang. Prognostic prediction of carcinoma by a differential-regulatory-network-embedded deep neural network. *Computational Biology and Chemistry*, 88:107317, 2020.

- [138] Md Mohaiminul Islam, Shujun Huang, Rasif Ajwad, Chen Chi, Yang Wang, and Pingzhao Hu. An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Computational and structural biotechnology journal*, 18:2185–2199, 2020.
- [139] Yang Guo, Shuhui Liu, Zhanhuai Li, and Xuequn Shang. Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1664–1669, 2017.
- [140] Milad Mostavi, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC medical genomics*, 13:1–13, 2020.
- [141] Lei Chen, XiaoYong Pan, Yu-Hang Zhang, Min Liu, Tao Huang, and Yu-Dong Cai. Classification of widely and rarely expressed genes with recurrent neural network. *Computational and Structural Biotechnology Journal*, 17:49–60, 2019.
- [142] Dejun Zhang, Lu Zou, Xionghui Zhou, and Fazhi He. Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *IEEE Access*, 6:28936–28944, 2018.
- [143] Feng Gao, Wei Wang, Miaomiao Tan, Lina Zhu, Yuchen Zhang, Evelyn Fessler, Louis Vermeulen, and Xin Wang. Deepcc: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis*, 8(9):44, 2019.
- [144] Bihter Das and Suat Toraman. Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized dna sequences. *Biomedical Signal Processing and Control*, 72:103317, 2022.
- [145] Rajul Mahto, Saboor Uddin Ahmed, Rizwan ur Rahman, Rabia Musheer Aziz, Priyanka Roy, Saurav Mallik, Aimin Li, and Mohd Asif Shah. A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection. *BMC bioinformatics*, 24(1):479, 2023.
- [146] Amol Avinash Joshi and Rabia Musheer Aziz. Deep learning approach for brain tumor classification using metaheuristic optimization with gene expression data. *International Journal of Imaging Systems and Technology*, page e23007, 2023.
- [147] Abdulrhman M Alshareef, Raed Alsini, Mohammed Alsieni, Fadwa Alrowais, Radwa Marzouk, Ibrahim Abunadi, Nadhemi Nemri, et al. Optimal deep learning enabled prostate cancer detection using microarray gene expression. *Journal of Healthcare Engineering*, 2022, 2022.
- [148] Lipo Wang, Yaoli Wang, and Qing Chang. Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods*, 111:21–31, 2016.
- [149] D.L. Barbour. Precision medicine and the cursed dimensions. *npj Digital Medicine volume*, 4(2), 2019.

- [150] Ali Dabba, Abdelkamel Tari, Samy Meftali, and Rabah Mokhtari. Gene selection and classification of microarray data method based on mutual information and moth flame algorithm. *Expert Systems with Applications*, 166:114012, 2021.
- [151] Mohsen Rahmanian and Eghbal G Mansoori. An unsupervised gene selection method based on multivariate normalized mutual information of genes. *Chemometrics and Intelligent Laboratory Systems*, 222:104512, 2022.
- [152] KR Kavitha, Avani Prakasan, and PJ Dhrishya. Score-based feature selection of gene expression data for cancer classification. In *2020 fourth international conference on computing methodologies and communication (ICCMC)*, pages 261–266. IEEE, 2020.
- [153] Kun Yu, Wei Li, Weidong Xie, and Linjie Wang. A hybrid feature-selection method based on mrmr and binary differential evolution for gene selection. *Processes*, 12(2):313, 2024.
- [154] Elnaz Pashaei and Elham Pashaei. Gene selection for cancer classification using a new hybrid of binary black hole algorithm. In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2020.
- [155] MAH Akhand, Md Asaduzzaman Miah, Mir Hussain Kabir, and MM Hafizur Rahman. Cancer classification from dna microarray data using mrmr and artificial neural network. *International Journal of Advanced Computer Science and Applications*, 10(7), 2019.
- [156] Martha A. Zaidan, Lubna Dada, Mansour A. Alghamdi, Hisham Al-Jeelani, Heikki Lihavainen, Antti Hyvärinen, and Tareq Hussein. Mutual information input selector and probabilistic machine learning utilisation for air pollution proxies. *Applied Sciences*, 9(20), 2019.
- [157] Artur J. Ferreira and Mário A.T. Figueiredo. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, 33(13):1794–1804, 2012.
- [158] Sanjay Yadav and Sanyam Shukla. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 78–83, 2016.
- [159] Yali Nie, Laura De Santis, Marco Carratù, Mattias O’Nils, Paolo Sommella, and Jan Lundgren. Deep melanoma classification with k-fold cross-validation for process optimization. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6, 2020.
- [160] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, 153:1–9, 2018.
- [161] Thomas Villmann, Marika Kaden, Mandy Lange, Paul Stürmer, and Wieland Hermann. Precision-recall-optimization in learning vector quantization classifiers for improved medical classification systems. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 71–77, 2014.

- [162] Arvind Kumar, Nishant Sinha, and Arpit Bhardwaj. A novel fitness function in genetic programming for medical data classification. *Journal of Biomedical Informatics*, 112:103623, 2020.
- [163] Liang Y Huang HH, Liu XY. Feature selection and cancer classification via sparse logistic regression with the hybrid  $l_{1/2} + l_2$  regularization. *PLoS One*, 11(5), 2016.
- [164] Nour Eldeen M. Khalifa, Mohamed Hamed N. Taha, Dalia Ezzat Ali, Adam Slowik, and Aboul Ella Hassanien. Artificial intelligence technique for gene expression by tumor rna-seq data: A novel optimized deep learning approach. *IEEE Access*, 8:22874–22883, 2020.
- [165] Venkatesan A Naorem LD, Muthaiyan M. Integrated network analysis and machine learning approach for the identification of key genes of triple-negative breast cancer. *J Cell Biochem*, 120(4), 2019.
- [166] Jaroslav Frnda<sup>3</sup> 4 Parame-shachari B.D.<sup>5</sup> Srinivas Konda<sup>6</sup> Sujata Patil<sup>1</sup>, Kavitha Rani Balmuri<sup>2</sup> and Jan Nedoma<sup>4</sup>. Identification of triple negative breast cancer genes using rough set based feature selection algorithm ensemble classifier. *Human-centric Computing and Information Sciences*, 12(54), 2022.
- [167] Claudia Cava, Christian Salvatore, and Isabella Castiglioni. Pan-cancer classification of gene expression data based on artificial neural network model. *Applied Sciences*, 13(13):2076–3417, 2023.
- [168] Omar Abdelwahab, Nourelislam Awad, Menattallah Elserafy, and Eman Badr. A feature selection-based framework to identify biomarkers for cancer diagnosis: A focus on lung adenocarcinoma. *Plos one*, 17(9):e0269126, 2022.
- [169] Yao W Liu, S. Prediction of lung cancer using gene expression and deep learning with kl divergence gene selection. *BMC Bioinformatics*, 175(23), 2022.
- [170] Pilar García-Díaz, Isabel Sánchez-Berriel, Juan A. Martínez-Rojas, and Ana M. Diez-Pascual. Unsupervised feature selection algorithm for multiclass cancer classification of gene expression rna-seq data. *Genomics*, 112(2):1916–1925, 2020.
- [171] Saba Bashir, Irfan Ullah Khattak, Aihab Khan, Farhan Hassan Khan, Abdullah Gani, and Muhammad Shiraz. A novel feature selection method for classification of medical data using filters, wrappers, and embedded approaches. *Complexity*, 2022, 2022.
- [172] Lim-E.L. Weeden C.E. et al Hill, W. Lung adenocarcinoma promotion by air pollutants. *Nature*, 616:159–167, 2023.
- [173] Fadi Alharbi and Aleksandar Vakanski. Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10(2):173, 2023.
- [174] Ying Zhang, Qingchun Deng, Wenbin Liang, and Xianchun Zou. An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. *BioMed research international*, 2018.

- 
- [175] Surbhi Gupta, Manoj K Gupta, Mohammad Shabaz, and Ashutosh Sharma. Deep learning techniques for cancer classification using microarray gene expression data. *Frontiers in Physiology*, 13:952709, 2022.
- [176] Kazim Y Arga and Raghu Sinha. Recent developments in cancer systems biology: Lessons learned and future directions. *Journal of Personalized Medicine*, 11(4):271, 2021.
- [177] Zhang H Yi B Wozniak M Wong L Zhou H, Jin J. Inpath—an integrated pathway gene relationship database for model organisms and important pathogens. *BMC Syst Biol*, 6, 2012.
- [178] Tan Q Kir J Liu D Bryant D Guo Y Stephens R Baseler MW Lane HC Lempicki RA Huang DW, Sherman BT. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*, 35:75–169, 2007.