GENOMIC DATA ANALYSIS USING MACHINE LEARNING METHODS FOR DISEASE AND DISEASE-GENE PREDICTION

by

Nuriye Özlem Özcan Şimşek

B.S., Computer Engineering, Middle East Technical University, 2005

M.S., System and Control Engineering, Boğaziçi University, 2010

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2024

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisors, Prof. Fikret Gürgen and Prof. Arzucan Özgür, for their kindness, intellectual support and guidance. During my PhD studies, I have always felt their support and followed their invaluable advice on every problem. I am really grateful that I had the opportunity to study with them.

In addition to my advisors, I would also like to thank my thesis progress committee: Prof. Tunga Güngör, Prof. Olcay Taner Yıldız, for their insightful comments and suggestions. I would also like to thank my defence committee: Prof. Tunga Güngör, Prof. Olcay Taner Yıldız, Assist. Prof. İnci Meliha Baytaş, Prof. Banu Diri, for their attention, time and valuable comments.

This research is supported by Boğaziçi University Research Fund Grant Number 13242. I would like to thank them for providing financial support in purchasing the servers and datasets used for the studies in this thesis and in attending conferences.

I would also like to thank my colleagues at MediaLab and TabiLab for taking the time to follow and listen to my thesis presentations, as well as for their valuable discussions, technical and motivational help.

Last but not least, I would like to thank my family. I am really grateful that my parents, my father Prof. Mustafa Özcan and my mother Nedret Özcan, have always supported me throughout my whole education life and in the completion of my PhD thesis. I thank my husband Ozan Şimşek for his love and constant support and encouragement during these studies. Last but most precious, I thank to my 4-year-old son Mert Şimşek for the joy and patience he brings to my life.

ABSTRACT

GENOMIC DATA ANALYSIS USING MACHINE LEARNING METHODS FOR DISEASE AND DISEASE-GENE PREDICTION

Genomic diseases are caused by certain mutations or combinations of mutations in the DNA. This combination can be different in each patient and the impact of each mutation on the disease is different. In this study, we focus on the genomic causes of diseases. We have defined two research problems. One is to detect the disease from the genetic code, which is represented as a list of mutations. The second is to identify associations between diseases and genes.

In this thesis, we focused on cancer and proposed three frameworks. In one approach, the list of mutations was modeled as a document and the mutations were modeled as words in the document. Based on this assumption, we proposed representation models for these patient mutation documents and used them for disease prediction. For another approach, we modeled a novel heterogeneous graph environment in which patients and genes/proteins are the nodes and mutations define the edges. Both approaches to the problem resulted in significantly better classification performances with the selected algorithms, demonstrating the success of our novel designs for the input mutations. The parameters of the two classification frameworks were analyzed and a list of the most effective genes for disease prediction was generated for each of these systems. These genes were found to be studied as causal genes or target genes in the cancer literature. In addition, this list of effective genes was transferred to the gene expression domain as a gene selection algorithm and was found to increase the rate of true prediction of disease. The proposed systems have been tested on cancer data but can be easily adapted to other genomic diseases.

ÖZET

HASTALIK VE HASTALIKLA İLİŞKİLİ GEN TAHMİNİ İÇİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILAN GENOMİK VERİ ANALİZİ

Genomik hastalıklara DNA'daki belirli mutasyonlar veya mutasyon kombinasyonları neden olur. Bu kombinasyon her hastada farklı olabileceği gibi, her mutasyonun hastalık üzerindeki etkisi de farklıdır. Bu çalışmada hastalıkların genomik nedenlerine odaklanıyoruz. İki araştırma problemini tanımladık. Birincisi, mutasyonların bir listesi olarak temsil edilen genetik koddan hastalığı tespit etmektir. İkincisi, hastalıklar ve genler arasındaki ilişkileri belirlemektir.

Bu tezde kansere odaklandık ve üç yöntem önerdik. Bir yaklaşımda mutasyonların listesi bir belge olarak modellendi ve mutasyonlar da belgedeki kelimeler olarak modellendi. Bu varsayıma dayanarak, hasta mutasyon belgeleri için temsil modelleri önerdik ve bunları hastalık tahmini için kullandık. Başka bir yaklaşım için, hastaların ve genlerin/proteinlerin düğüm olduğu ve mutasyonların kenarları tanımladığı yeni bir heterojen grafik ortamı modelledik. Soruna yönelik her iki yaklaşım da, seçilen algoritmalarla karşılaştırıldığında, önemli ölçüde daha iyi sınıflandırma performanslarıyla sonuçlandı ve bu da giriş mutasyonları için yeni tasarımlarımızın başarısını gösterdi. İki sınıflandırma çerçevesinin parametreleri analiz edildi ve bu sistemlerin her biri için hastalık tahmininde en etkili genlerin bir listesi oluşturuldu. Bu genlerin kanser literatüründe nedensel genler veya hedef genler olarak çalışıldığı görülmüştür. Ek olarak, etkili genlerin bu listesi, bir gen seçim algoritması olarak gen ekspresyon alanına aktarıldı ve hastalığın tahmin oranını arttırdığı görüldü. Önerilen sistemler kanser verileri üzerinde test edilmiştir ancak diğer genomik hastalıklara da kolaylıkla uyarlanabilir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . ix
LIST OF TABLES . . . x
LIST OF SYMBOLS . . . xii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiii
1. INTRODUCTION . . . 1
1.1. Problem Statement . . . 1
1.2. Challenges . . . 3
1.3. Motivation . . . 4
1.4. Publication Notes . . . 5
1.5. Thesis Overview . . . 6
2. BACKGROUND . . . 9
2.1. Related Genomic Concepts . . . 9
2.1.1. DNA Sequencing . . . 9
2.1.2. Genes and Mutations . . . 9
2.1.3. Proteins . . . 10
2.2. Related Data Sources . . . 10
2.2.1. TCGA . . . 11
2.2.2. UKBiobank . . . 11
2.3. Related Machine Learning Concepts . . . 12
2.3.1. Logistic Regression . . . 12
2.3.2. Feed-Forward Neural Networks . . . 12
2.3.3. Graph-Based Neural Networks . . . 13
3. STATISTICAL REPRESENTATION MODELS FOR MUTATION INFO . . . 16
3.1. Background . . . 16
3.2. Methods . . . 19

3.2.1. Dataset . . . 19
3.2.2. Data Representation Models . . . 20
3.2.2.1. Binary Mutation Model . . . 20
3.2.2.2. C-score Based Mutation Model . . . 20
3.2.2.3. Tf-idf Based Mutation Model . . . 20
3.2.2.4. Tf-rf Based Mutation Model . . . 22
3.2.2.5. BM25-tf-idf Based Mutation Model . . . 22
3.2.2.6. BM25-tf-rf Based Mutation Model . . . 23
3.2.3. Implementation and Experiment Design . . . 24
3.2.3.1. Machine Learning Models . . . 24
3.2.3.2. Evaluation Strategy . . . 24
3.2.3.3. Parameter Tuning for The Representation Models . . . 25
3.2.3.4. Parameter Tuning for The Classification Models . . . 25
3.3. Results . . . 25
3.3.1. Selection of Gene Sets . . . 25
3.3.2. Comparison of The Data Representation Models . . . 26
3.3.3. Class-Based Comparison of Experiment Results . . . 27
3.3.4. Location-Based Comparison of Gene Mutations . . . 28
3.4. Discussion . . . 29
3.5. Final Remarks . . . 31
4. GENE SELECTION METHOD FOR GENE EXPRESSION DATA . . . 46
4.1. Background . . . 46
4.2. Methods . . . 48
4.2.1. Dataset . . . 48
4.2.2. Feature Selection for Gene Expression . . . 48
4.2.3. Implementation of Methods and Experiments . . . 49
4.3. Results and Discussion . . . 50
4.3.1. Comparison with a Previous Study . . . 50
4.3.2. Gene Selection Results . . . 50
4.4. Final Remarks . . . 52
5. GNNMUTATION: A HETEROGENEOUS GRAPH FRAMEWORK . . . 57

5.1. Introduction 57
5.2. Materials and Methods 59
5.2.1. Dataset 60
5.2.2. GNNMutation 60
5.2.2.1. Heterogeneous Graph 60
5.2.2.2. Patient Feature Vectors 61
5.2.2.3. GNN Model 61
5.2.2.4. Explainer Model 61
5.2.3. Implementation and Experiment Design 62
5.3. Results and Discussion 63
5.3.1. Comparison of Different Numbers of Protein-Patient Edges . . . 63
5.3.2. Comparison of Different Protein-Patient Edge Strategies 63
5.3.3. Comparison of Algorithms 64
5.3.4. Discussion of Explainer Results 65
5.4. Final Remarks 66
6. CONCLUSION 72
6.1. Discussion 72
6.2. Future Work 74
REFERENCES 75

LIST OF FIGURES

Figure 3.1. Overview of the proposed system. 33

Figure 3.2. The effect of the smoothing parameter k 34

Figure 3.3. The heat map of the most effective genes in NN - breast cancer . . 44

Figure 3.4. The heat map of the most effective genes in NN - lung cancer . . . 45

Figure 4.1. Overview of the proposed system. 53

Figure 4.2. F-score and feature count comparison of experiment results. 56

Figure 5.1. Overview of GNNMutation 68

Figure 5.2. Heterogeneous graph of patients and genes 69

Figure 5.3. Explainer results for breast cancer. 71

Figure 5.4. Explainer results for prostate cancer. 71

LIST OF TABLES

Table 3.1. The list of cancer types and sample counts in BOUN10CANCER . 32

Table 3.2. Parameter tuning results for parameter k in the BM25-tf formula . 34

Table 3.3. The range/list of parameters used in the parameter tuning phase . 35

Table 3.4. The best parameters found as a result of parameter tuning phase . 36

Table 3.5. Machine learning experiment test results on the gene sets 37

Table 3.6. Machine learning experiment test results - 1 38

Table 3.7. Machine learning experiment test results - 2 39

Table 3.8. Machine learning experiment test results - 3 40

Table 3.9. Machine learning experiment test results - 4 41

Table 3.10. Class based experiment test results with NN 42

Table 3.11. Machine learning experiment test results - exonic and intronic . . . 43

Table 4.1. The list of cancer types and sample counts in our dataset 52

Table 4.2. Machine learning experiment results - gene expression and VCF . . 54

Table 4.3. Machine learning experiment results of gene expression data 55

Table 5.1. Classification results - different number of protein-patient edges . . 69

Table 5.2. Classification results - different protein-patient edge strategies . . . 70

Table 5.3. Classification test accuracy results 70

LIST OF SYMBOLS

a	The number of samples in class c
b	The number of samples in other classes
c	Class c
g	Gene g
GB	Gigabyte
k	A smoothing parameter for term-frequency
MB	Megabyte
N	Size N
R^n	The set of all ordered n-tuples of real numbers
s	Sample s
x	The input at hand
x'	The baseline input
α	The interpolation constant to perturb features by
∂	Partial derivative

LIST OF ACRONYMS/ABBREVIATIONS

A	Adenine
C	Cytosine
CADD	The Combined Annotation Dependent Depletion
CNV	Copy Number Variation
CPU	Central Processing Unit
CT	Computed Tomography
DNA	Deoxyribonucleic Acid
FPKM	Fragments Per Kilobase Of Transcript Per Million Reads
FPR	False Positive Rate
G	Guanine
GAT	Graph Attention Neural Networks
GCN	Graph Convolutional Networks
GIN	Graph Isomorphism Networks
GNN	Graph Neural Networks
GPU	Graphics Processing Unit
GRC	The Genome Reference Consortium
IDF	Inverse Document Frequency
IG	Integrated Gradients
KNN	K-Nearest Neighbor
LR	Logistic Regression
max	Maximum
MCC	Matthews Correlation Coefficient
miRNA	MicroRNA
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Images
mRMR	Minimum Redundancy Maximum Relevance
mRNA	Messenger RNA
NB	Naive Bayes

NIH	The National Institutes of Health
NN	Neural Network
NSCLC	Non Small Cell Lung Cancer
Perceptron	One-Layer-Perceptron
PPI	Protein-Protein Interaction
RF	Relevance Frequency
RNA	Ribonucleic Acid
Roc-Auc	Area Under The Receiver Operating Curve
SCLC	Small Cell Lung Cancer
SNV	Single Nucleotide Variant
SVM	Support Vector Machine
T	Thymine
TCGA	The Cancer Genome Atlas
TF	Term Frequency
TNBC	Triple Negative Breast Cancer
UKBiobank	The UK Biobank
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

1. INTRODUCTION

A genetic disease is caused by a change in the DNA sequence [1]. These changes are called mutations. The occurrence of complex genetic diseases is often due to multiple gene mutations. Investigating which combination of mutations leads to the development of a disease is a complex problem. This is because each patient may have a different combination. In order to specify the general causal DNA alterations, a descriptive dataset must be analysed that represents a large set of human samples. Until the last 20 years, this analysis was not possible. The Human Genome Project [2] generated the first sequence of the human genome in 2003. Since then, there are huge DNA sequencing databases that make it possible to perform genomic studies on large numbers of patients.

The mutations in the human DNA can be detected by comparing the target DNA sequence with the current reference genome, GRCh38.p14, published by the Genome Reference Consortium (GRC) [3]. As a result of this comparison, we obtain a comprehensive list of mutations. Not all mutations cause diseases. Moreover, the effect of each mutation is not equal in the development of diseases. The actual causal changes in the DNA can be detected with a sensitive and precise system. The performance of such a system can be improved with a diverse set of training samples. As more and more human sequences become available, more successful computational frameworks for sequence analysis are being developed.

1.1. Problem Statement

The main problems addressed in this thesis are the prediction of diseases based on mutation information and the detection of causal or target genes for the development of diseases. We chose cancer as our target disease because it is one of the leading causes of death worldwide [4]. The mortality rate from cancer can be reduced through early diagnosis and personalised treatment. Computer systems that specialise in detecting

certain types of cancer can be used as an aid in diagnosing the disease. In this way, the rate of early diagnosis increases and the cost of diagnosis decreases. Computer-based systems for the detection of causal and target genes can lead to improvements in personalised treatments. The research problems are defined in detail in this section.

Cancer is usually diagnosed by the detection of a tumour. A biopsy is taken to examine the tumour and determine whether it is benign or malignant. A biopsy means an operation. However, the genetic code of a patient can be determined using only a blood sample. We proposed systems that can analyse the genomic code and decide on the disease. These systems can help the health profesionals in diagnosis and reduce the number of unnecessary biopsies. In this thesis, we focused on DNA sequencing data. A whole genome sequence of a patient can be very huge in size. When we think of a dataset consisting of thousands of patients, this pushes us to the limits of practical computation. Therefore, we chose a smaller data format for our studies. We utilized DNA mutation information in variant call format (VCF) files [5]. A VCF file of a patient lists all DNA mutations in the base sequence file. Since we want to predict a disease based on the mutation information, this data format is sufficient for us. We proposed and developed machine learning frameworks that process mutation data to predict whether a patient has cancer or not and to determine the type of cancer.

The second research problem in this thesis is the detection of causal or target genes for diseases. These genes can be utilized in genetic tests or in the development of drugs. Early diagnosis can be made with the help of diagnostic genetic tests and regular screening. These tests look for certain genomic variations. The development of genetic tests is based on genome research, which uncovers the causal effects of genes. Using the extensive genomic data available in biobanks, computer systems can help to analyse and extract valuable genomic information and the most effective genes. In addition, these computer systems can also detect target genes for drug development. This enables the development of specialised and personalised treatments for each patient. We added explainer blocks to our disease prediction frameworks to recognise the most effective genes and mutations for cancer prediction.

In this thesis, we addressed the genomic disease prediction and causal or target gene prediction tasks, experimenting with the human DNA mutation information. We tested our proposed systems with cancer specific data. However, these systems can be trained and applied to any genomic diseases.

1.2. Challenges

There are many challenges for the prediction of diseases based on mutation information and the detection of causal or target genes for the development of diseases. Some of these challenges are summarized as followings:

(i) Restricted access: For genomic data analysis, the source of data is generally a biobank. Biobanks collect detailed health and genomic data from patients, even biological samples can be included. The data is provided anonymously but the access wrights are restricted and controlled by biobank data access management teams.

(ii) Big data and storage: The raw genomic data is enormous. A single 30x precision whole genome sequencing (WGS) file is about 100GB in size for a single patient. In addition, there is also the output data generated by representation models and algorithms tested throughout the thesis study. For a successful genomic analysis study, a sufficient and stable data storage plan is required.

(iii) Computation time and memory cost of big data: Once the storage plan for the huge genomic data is finalised, a suitable server for data processing must be selected. Since we are dealing with large amounts of data, we need powerful CPUs and GPUs. In addition, the genomic data or a suitable representation of the data must fit into the server's memory so that the selected algorithm can be executed on the data.

(iv) Suitable data type for the research problems: If no adequate storage system can be constructed, another suitable data type, which is smaller, can be chosen for the research problems. For the case of this thesis, we could not provide storage for the actual sequencing data. Therefore, we decided to work with VCF files,

which only list DNA mutations that are changes in the target sequence compared to the reference genome. For the VCF files used in this thesis, the size of a single VCF file may vary between 1MB and 1GB, depending on the precision and the coverage of the source sequence. This is still big data when thousands of patients are considered. But compared to the raw sequencing data, the size of VCF files is more manageable.

(v) Representation model: After selecting VCF files as the input data type, the computational and memory costs are still present. Pre-processing of the data is therefore urgently required. Statistical models are applied to summarise the mutation information.

(vi) Explainable classification algorithm: The second research problem is the identification of causal or target genes for the diseases. The proposed frameworks predict on diseases based on the gene mutation information. Therefore, we need to select algorithms that are explainable in order to trace back the algorithm and recognise the most effective genes and mutations in disease prediction.

1.3. Motivation

Genomic analysis involves comparing the genomes of many different people to find genetic markers associated with a particular phenotype or risk of disease [6]. Prior to the development of biobanks, these studies were conducted on small datasets curated by hospitals or medical groups and generally involved statistical analysis. Now that thousands of genomic samples are available, computational frameworks are used for these tasks. Software frameworks save time and also enable more detailed computations of the data.

Different data types are used in genome-wide studies. Gene expression profile data are frequently used. They show the expression levels of genes for the patient. The expression level of a gene provides information about its effective state and shows whether it is active or suppressed. This provides valuable information about the development of diseases. DNA methylation data is also used in genomic studies. DNA

methylation can alter gene activity without changing the DNA sequence [7]. DNA mutations are also used in genome-wide studies. Single nucleotide variants (SNVs) and copy number variations (CNVs) are the most commonly used types of mutations that have important implications for human disease and evolution. There are other types of genomic data, but we have listed the most commonly used ones here.

In this thesis, we wanted to find the roots of diseases directly in the genetic code, i.e. the DNA. As the size of raw DNA sequencing data is enourmous, we decided to utilize gene mutations, as these are the altered parts of the DNA and are thought to be the actual genetic causes of disease. We work with VCF files, which are created based on the DNA sequence and contain a list of mutations. As far as we know, there is no other disease association study that works with the whole VCF files as the main source of genomic information. In Chapter 3, we used the mutations in VCF files. We proposed representation methods and decision frameworks to link the mutation information in the VCF to the target disease. After this initial study, we searched the cancer literature and found that gene expression data is widely used for disease detection. Therefore, in Chapter 4, we proposed our study using VCF data for a novel gene selection method for gene expression data. In the course of these studies, we also learnt about other biological molecules and structures involved in the development of genomic diseases. Genes encode proteins and proteins and their interactions play an important role in cell function. We wanted to use the knowledge about proteins and combine it with mutations to search for disease-causing situations. Therefore, in Chapter 5, we used protein interaction networks and mutations to model a graph representation and detect diseases based on this graph.

1.4. Publication Notes

Parts of the work in this thesis have appeared in the following publications:

(i) Özcan Şimşek, N.Ö., A. Özgür, F. Gürgen, "Statistical Representation Models for Mutation Information within Genomic Data", BMC Bioinformatics, Vol. 20,

No. 324, pp. 1-13, 2019 [8] (Chapter 3).

(ii) Özcan Şimşek, N.Ö., A. Özgür, F. Gürgen, "A Novel Gene Selection Method for Gene Expression Data for the Task of Cancer Type Classification", Biology Direct, Vol. 16, No. 7, pp. 1-5, 2021 [9] (Chapter 4).

(iii) Özcan Şimşek, N.Ö., A. Özgür, F. Gürgen, "GNNMutation: a Heterogeneous Graph-Based Framework for Cancer Detection Based on DNA Mutations", (Submitted to Bioinformatics)(Chapter 5).

1.5. Thesis Overview

In this thesis, we focused specifically on finding the genomic roots of diseases. We defined two research problems for this topic. Firstly, we aim to predict the disease using the DNA mutation information as our input. Secondly, we aim to address the specific parts of the input that lead the frameworks for disease prediction. By addressing both problems, we proposed three frameworks that can successfully discriminate between disease types and between patients with or without the disease. The proposed frameworks are tested on cancer specific data. However, these systems can be adapted to target any genomic diseases.

This thesis demonstrates that VCF files can be utilized as an alternative representation of DNA sequences and can lead to state-of-the-art performance in disease prediction. Chapter 3 presents and compares the proposed preprocessing methods for VCF files. In this study, we model the VCF file as a document and mutations as words inspired from the field of information retrieval. We proposed representation models based on this assumption and compare their effectiveness on cancer prediction. It is also shown that the proposed method can be utilized to detect causal or target genes for each cancer type in the study. Chapter 4 presents a transformation study of the previous method as a gene selection method for gene expression data and the resulting performance in disease prediction. Chapter 5 introduces a novel graph structure for the representation of VCF data and shows the resulting performance of this representation in both disease prediction and causal or target gene detection.

The main contributions of this thesis are summarized below.

(i) We transferred our knowledge from the field of information retrieval to mutations. We modelled VCF files as documents consisting of mutations as words. We calculated term weighting methods on the mutation words and used them to represent patients. We wanted to reflect the different effects of each mutation on disease development. We performed a comparison of these data representation schemes for the task of cancer type classification. We proposed a neural network architecture for disease classification and showed that the proposed method achieves significantly better performance in discriminating cancer types compared to other machine learning methods (Chapter 3) [8].

(ii) A novel and disease specific gene selection method for gene expression data was proposed and tested. We extracted the most effective genes for disease development based on DNA mutations and transferred this information to the gene expression domain as a novel gene selection method. Our test results showed that the new feature selection method leads to equivalent or better performance compared to other feature selection methods (Chapter 4) [9].

(iii) A novel heterogeneous graph structure has been proposed for the representation of patients and proteins for disease classification. We modelled patients and proteins as separate nodes in the graph. The edges between proteins were defined according to the PPI network. The edges between patients and proteins were defined based on DNA mutation information. If there is a mutation in a gene, we add an edge from the target protein of this gene to the patient. Thus, the proposed graph structure was able to represent the protein interactions and gene mutations together. We processed this heterogeneous graph with GAT, which can distinguish the effects of neighbours on a target node. The proposed framework led to a leap in classification performance compared to other machine learning methods (Chapter 5).

(iv) We analysed our proposed frameworks for disease classification after training and identified the most effective features, in our case genes, for prediction. In a literature search on cancer research, we found that almost half of the identified

genes were analysed and reported as causal or target genes for the corresponding cancer types. This showed us that the proposed methods make predictions based on true causal genes. In addition, the rest of the identified most effective genes can be investigated as candidate genes for cancer in laboratory studies (Chapter 3, Chapter 5) [8].

2. BACKGROUND

This chapter presents the background information on the basic concepts used in this thesis study. The first section describes the genomic concepts related to the studies. The second section describes the genomic data sources used in the studies. The third section describes the main algorithms used in the studies.

2.1. Related Genomic Concepts

2.1.1. DNA Sequencing

DNA is the molecule that carries the genetic information for the development and functioning of an organism [1]. Human DNA consists of four bases: Adenine (A), Cytosine (C), Guanine (G) or Thymine (T). DNA sequencing means determining the order of these four chemical building blocks [1]. Human DNA contains more than 3 billion base pairs (bp) [10]. We can think of the DNA sequence as a long text written in a small alphabet with only four letters; A, C, F and T. The length of this text is more than 3 billion characters. This means that a single human DNA sequence comprises at least 3GB of data, if we use one byte for the representation of one base.

Whole genome sequencing (WGS) is the sequencing of the entire DNA parts. For a 30x precision for WGS, the resulting file is about 100GB in size. There is also whole exome sequencing (WES), in which only the coding part of the DNA, which makes up %1 of the total DNA, is sequenced. For a 100x precision for WES, the resulting file is about 8GB.

2.1.2. Genes and Mutations

Genes contain the information that determines physical and biological characteristics [1]. They are the part of DNA that carry the codes for the production of

proteins. There are almost $20,000$ protein-coding genes in the human DNA [1]. There can be changes in gene sequences and they are called mutations. Mutations usually don't have major results [1]. But some mutations or combination of mutations lead to the development of genomic diseases.

Variant calling is the comparison of the target DNA sequence with the reference genome. The reference genome can be regarded as the standard DNA sequence without mutations. This comparison reveals the alterations in the target DNA sequence. VCF [5] is the standard file format for storing the mutation data. In this file, each data line corresponds to a mutation and provides detailed information about that mutation. By using a 30x precision for WGS, the resulting VCF file is about 1GB. By using a 100x precision for WES, the resulting VCF file is about 0.1GB.

2.1.3. Proteins

Genes in the DNA are used to encode proteins, which is called the expression of the gene. Proteins determine and control the functioning of cells. Mutations in the gene sequence alter the resulting protein which can change biological processes. In addition, the level of gene expression determines the amount of proteins produced and this can also effect biological processes. Therefore, gene expression levels are popularly studied for disease analysis.

Proteins are biological molecules that interact with each other. As proteins can alter biological processes, their interactions can also effect biological operations. Therefore, protein-protein interactions are analysed and saved as protein-protein interaction (PPI) networks. These interaction networks are popularly studied for disease analysis.

2.2. Related Data Sources

The data sources used in this thesis are The Cancer Genome Atlas (TCGA) and UK Biobank (UKBiobank). They are described in detail in the following subsections.

2.2.1. TCGA

TCGA [11] is a project launched by The National Institutes of Health (NIH) [12] in 2005. It is a cancer specific genomic data collection project. They concluded sample collection in 2013. They published their Pan-Cancer Atlas [13,14] study in 2018. They collected over 20,000 cancer samples spanning 33 cancer types. Some of the available data types are clinical data, copy number data, DNA sequencing data, imaging, methylation, miRNA sequencing, mRNA sequencing and gene expression data. Some data types are provided with open access. They can be downloaded directly from the Genomic Data Commons Data Portal [15] supported by NIH. Some data types are provided with controlled access. Researchers have to be registered and get access permission for TCGA dataset from NIH. Gene expression data is served with open access but VCF files are served with controlled access. Therefore, we were registered to NIH and were approved to access TCGA controlled data.

2.2.2. UKBiobank

UKBiobank [16] was founded by the Wellcome Trust medical charity [17] in 2006. It is a large-scale biomedical database containing biological samples, genetic and health information from half a million UK participants. The data can only be used by authorised researchers. According to UKBiobank, more than 30,000 researchers worldwide are registered to access UKBiobank data and more than 9,000 scientific papers have been published about their studies using the data. UKBiobank provides online access to data and biological samples in response to specifically authorised requests. UKBiobank offers a variety of data types including hospital reports, biological samples, questionare results, lifestyle information and genomic data. For the genomic part, it provides WGS and WES data for the whole cohort. We had authorised access to the first 200,000 of WES data. The sequences are obtained from blood samples, unlike TCGA where tumour samples are used. This is because the UKBiobank contains information on patients with and without a cancer diagnosis. This allows us to compare not only different types of cancer, but also cancer patients with non-cancer patients.

2.3. Related Machine Learning Concepts

The machine learning techniques that we apply in this thesis are logistic regression, feed-forward and graph-based neural networks. They are briefly explained in this section.

2.3.1. Logistic Regression

Logistic regression is a supervised machine learning algorithm used for classification tasks. It defines a coefficient for each input feature value and decides on the class probability with a sigmoid function on the weighted sum of the coefficients and the features and an extra bias term.

If we think of a genomic disease as the result of a combination of mutations. Each mutation has different degree of impact on the development of a disease. We can model the relationship between mutation and disease as an LR problem. Mutations are the features of a patient. When the LR system is trained with truth values (class labels), the system learns the importance levels of each mutation as a coefficient of that feature. The same approach can be applied to gene expression values. Since we consider each gene expression as a feature, its impact ratio can be calculated as coefficients in an LR system. With the use of coefficients, the system is easy to interpret. If the coefficient value for feature A is greater than the coefficient value for feature B, this can be interpreted as the impact of feature A on the target class is higher than feature B.

2.3.2. Feed-Forward Neural Networks

The architecture of a feed-forward neural network consists of three types of layers: the input layer, the hidden layers, and the output layer. Each layer consists of units, called neurons, and the layers are interconnected by weights. In the case of classification, the output class is the cumulative result of the weights in each hidden layer and the input features. By adding layers, a more complex relationship between

the input features and the target class can be defined. This architecture still has no cycles or loops and is therefore still an easy to interpret multi-layer system.

When we defined our study in Chapter 3, there was no definition of explainer modules. Therefore, we wanted to work with algorithms where we could easily interpret the reasons for the predictions. The feed-forward neural network happened to be the answer to our search for an explainable method that is also capable of defining complex relationships. Our test results prove the success of the proposed feed-forward neural network module compared to related algorithms.

To identify the roots of the predictions, we performed explainer tests after training phase is complete and the system coefficients are decided. Our input was a vector of gene features. To determine the cumulative effect of the hidden layer weights for a gene, we tested the system with one-hot inputs. Each time, only a single gene feature is set to 1 and the others set to 0. Since our tests were 10-fold cross validation, we had 10 trained models and tested each one-hot gene vector 10 times, once each trained model. Then we take the average of the prediction class probabilities. In the end, we obtained a list of coefficients reflecting the effect of each gene for each disease class. We plotted these coefficients with a heatmap of genes and diseases and searched for the most effective genes for each class prediction.

2.3.3. Graph-Based Neural Networks

For our study in Chapter 5, we used PPI networks to model the biological interaction system. Since we have a defined network, we wanted to model the disease prediction problem as a graph-based problem. We know that diseases arise from a combination of mutations which can be different for each patient, and that the effect of each mutation in the combination is different for the development of the disease. Therefore, we needed different weights for the different interactions. For this modelling of the problem, we used Graph Attention Neural Networks (GAT) for classification. GAT introduces the attention mechanism as a replacement for the normalized convolution

operation in Graph Convolutional Networks (GCN). In this way, each neighbour has a different effect on the target node and this is defined by attention weights. The attention weights are not statically defined, but are learnt during the training of the model. By applying GAT to our problem of disease prediction based on mutations, the learnt attention weights reflect the importance of each mutation for disease development.

GAT is defined in [18]. It is proposed for node classification on graph structures. The idea behind GAT is to compute the hidden representations of each node in the graph with its neighbors following a self-attention strategy. For this purpose, a new graph attention layer is proposed. The input of this layer is a set of node features, $h = \{h_1, h_2, ..., h_N\}$, where N is the number of nodes in the input graph. After processing these inputs, this layer produces a new set of features as output, $h' = \{h'_1, h'_2, ..., h'_N\}$. A shared attention mechanism, a, is used to compute attention coefficients, e,

$$e_{i,j} = a(Wh_i * Wh_j), \tag{2.1}$$

where W is a weight matrix, i and j are graph nodes. $e_{i,j}$ represents the impact level of node j on node i. The attention coefficients are then normalized accross all choices of j bu using $softmax$ function

$$\alpha_{ij} = softmax_j(e_{i,j}). \tag{2.2}$$

After these coefficients are decided, the new features are calculated as

$$h'_i = \sigma(\sum_{j \in N_i} \alpha_{ij} W h_j), \tag{2.3}$$

where σ is a nonlinearity function and N_i is some neighbourhood of node i.

We defined a novel heterogeneous graph structure for the problem of disease prediction. We used the PPI network as the basis of the graph. The proteins were the nodes and the edges were defined based on interactions. The protein-protein edges were defined as undirected. We added patients as another type of node in the graph and defined edges between patient nodes and protein nodes based on gene mutations. The protein-patient edges were directed from the protein node to the patient node. An edge from a protein to a patient is defined when there is a mutation in a gene in the

DNA sequence of that patient and that gene is responsible for the production of that protein. For example, if a patient has a mutation in the BRCA1 gene, then in our graph we add an edge from the BRCA1 protein to this patient in our graph. We thus expressed both the protein interactions and the DNA mutations in our heterogeneous graph structure. By processing the resulting graph with GAT, the different effects of protein interactions and gene mutations on the disease were taken into account. The test results in Chapter 5 show the success of this representation in disease prediction.

As a final step of the study, which is described in detail in Chapter 5, we added an explainer block to analyse the trained GAT classifier. We selected Captum explainer since it could be applied on heterogeneous graphs.

3. STATISTICAL REPRESENTATION MODELS FOR MUTATION INFORMATION

As DNA sequencing technologies are improving and getting cheaper, genomic data can be utilized for diagnosis of many diseases such as cancer. Human raw genome data is huge in size for computational systems. Therefore, there is a need for a compact and accurate representation of the valuable information in DNA. The occurrence of complex genetic disorders often results from multiple gene mutations. The effect of each mutation is not equal for the development of a disease. Inspired from the field of information retrieval, we propose using the term frequency and term weighting measures to weight genes based on their mutations. The underlying assumption is that the more mutations a gene has in patients with a certain disease and the less mutations it has in other patients, the more discriminative that gene is.

We evaluated the proposed representations on the task of cancer type classification. The best performing classification system is further utilized for causal gene analysis. Examples from the most effective genes that are used for decision making are found to be in the literature as target or causal genes. The details of this study is presented in the following sections.

3.1. Background

Complex diseases with genetic components arise from different combinations of the mutations in DNA. With the help of the decreasing cost of sequencing technologies, large scale sequencing datasets are being curated. Machine learning methods can be helpful in analyzing the huge genome data. However, a suitable representation technique for the sequencing data is still a problem to be solved. In this paper, we propose using statistical methods from the field of information retrieval for the representation of mutation information in DNA. The proposed representation methods are evaluated for the task of cancer type classification.

As stated in the 2017 report of National Center for Health Statistics [19], cancer is the second among top leading causes of death. Cancer is a group of diseases and each cancer type is labeled by the primary area of the body where the cancer cells arise. Each cancer type in general has a different set of causal genes and the disease emerges from the combination of various mutations of these genes [6]. The treatment is planned according to the primary site. Late diagnosis prevents the application of treatments and often results in the loss of the patient. Accordingly, the unknown or wrong analysis of the primary site and late diagnosis are major problems for cancer patients. The use of genomic data for diagnosis might help both to recognize the disease in early stages and to accurately classify the primary site.

Cancer classification has been primarily based on the morphological appearance of the tumor. Medical images such as magnetic resonance images (MRI) [20–22], X-ray and computed tomography (CT) images [23, 24], as well as histopathology images [25, 26] have been utilized for cancer diagnosis and classification. Medical images provide valuable information, especially about tumors, but they represent a restricted area. Therefore, there is a need for a more comprehensive data source.

Another commonly used data type for cancer classification is gene expression data. While a number of studies utilizing gene expression data have addressed the classification of cancer types [27–31], this type of data is highly sensitive to the microarray experiment setup and in general suffers from low accuracy and robustness [32]. In addition, due to the high dimensionality of gene expression data, gene selection methods are commonly applied prior to classification [33, 34]. The feature selection step may eliminate genes that in general have minor effects on disease generation while still being significant for the diagnosis of particular cancer types for some patients.

The biotechnology improvements and the automation of sequencing systems have increased the speed and lowered the cost of human DNA sequencing, which enabled the usage of this data type for disease diagnosis. The variants or mutations in the DNA of an individual can be identified by comparing the DNA sequence of the individual

to the DNA sequence of a reference genome maintained by The Genome Reference Consortium [3] and stored in a variant call format (VCF) file [5]. In recent studies, the binary representations of mutation data obtained from sources of manually curated somatic mutation profiles have been utilized for cancer classification [35,36]. However, binary representation is a limited way of describing the data. It highlights the genes with mutations, but treats them as equal. The distinction of common, rare and disease causing mutations is not expressed with the binary representation. Therefore, methods such as C-score from the Combined Annotation Dependent Depletion (CADD) framework [37,38] have been developed for weighting gene mutations. Recently, the sum of C-scores of gene mutations has been successfully used to cluster breast cancer patients and predict the stage of the disease [39].

In this paper, we propose adapting and using term weighting techniques (tf-idf, tf-rf and BM25) from the information retrieval field for weighting genes based on mutation information. As far as we know, these techniques have not been used on variant data before. The proposed gene weighting techniques are evaluated for the task of cancer type classification. Our results demonstrate that the best performing information retrieval based model (bm25-tf-rf) outperforms the C-score based approach. When the best performing classification model is analyzed, the most effective genes in the classification of certain cancer types are found to have been also proposed as causal or target genes in the previously published studies. These literature findings support the effectiveness of our representation models.

Our work brings the following contributions:

(i) Term weighting methods from the field of information retrieval have been proposed for the representation of mutation information within genomic data.
(ii) A comparison of these data representation schemes for the task of cancer type classification has been performed using a wide range of machine learning methods.
(iii) The best performing representation and classification model are utilized for causal gene analysis.

3.2. Methods

In this section, we describe the proposed data representation models as well as their utilization with machine learning algorithms for cancer type classification. An overview of the developed system is shown in Figure 3.1. The phenotypes occur as a result of DNA mutations. In our system, we take VCF files, which hold the DNA mutation information, as input. These gene mutations are weighted by using the proposed representation models. The genomic data representation vector is then processed with a wide range of machine learning algorithms for the task of cancer type classification. So, the VCF data constitute the set of observations and the genes, which are weighted based on mutation information, are the features in our classification model for learning the given cancer classes. The first output of the system is the prediction of the cancer type. The second output is the most effective genes in the classification process, which are obtained by analyzing the most accurate representation and classification model.

3.2.1. Dataset

The Cancer Facts and Figures 2017 annual report [40] states the leading sites of cancer. According to this list and the common cancer types from The National Cancer Institute [6], we created a dataset of ten cancer types, which are observed frequently and account for above half of the estimated cancer caused deaths. The list of selected cancer types and the sample counts for each cancer type are provided in Table 3.1. We downloaded a total of 7028 VCF files for ten cancer types from The Cancer Genome Atlas (TCGA) [11]. ANNOVAR [41] is used for gene-based annotation of the VCF files. From the annotated files, we selected the exonic and intronic mutations as they include specific gene labels. This selection of mutation types resulted in $16,383$ distinct genes. As a result, our dataset, named as BOUN10CANCER, has 7028 samples with mutation information of $16,383$ distinct genes and a class label for each sample representing the cancer type.

3.2.2. Data Representation Models

3.2.2.1. Binary Mutation Model. In the binary model, a gene is represented with 1 if there is a mutation in that gene, and it is represented with 0 otherwise. Hence, the resulting data set has the gene labels as features and binary values for each feature. The binary value for each gene feature is extracted from the annotated VCF file. If the gene label exists in the annotated file, the value is 1, and it is 0 otherwise. If there are more than 1 mutation for a gene, the feature value for that gene is still set to 1.

This representation model is applied by using two different gene lists. The first model is constructed with the known causal genes for the selected cancer types. The causal genes are obtained from OMIM [42] by using the MSC tool [43] of HGNC [44]. We extracted 434 causal genes for the cancer types in the BOUN10CANCER dataset. The second model is constructed by using all mutated genes in the annotated VCF files. We extracted $16,383$ mutated genes for the cancer types in the BOUN10CANCER dataset.

3.2.2.2. C-score Based Mutation Model. The CADD framework [37, 38] is a Support Vector Machine (SVM) based framework which calculates C-scores for variants. The C-score integrates diverse annotations and creates a single score for a variant. In the C-score based mutation model, the sum of C-scores for all mutations in a gene are used as the feature value for that gene. For example, if a gene has two mutations in a sample, the feature value for that gene in that sample is the sum of the C-scores of these two mutations. The sum of C-scores approach was also used in [39] for breast cancer patient classification. We evaluate this approach for cancer type classification and compare it with the proposed information retrieval based mutation models described in the following subsections.

3.2.2.3. Tf-idf Based Mutation Model. Complex diseases are in general developed from the combination of various mutations in the genes. Each mutation may influence the

evolution of the disease at different levels. In order to express these differences, we proposed to utilize the tf-idf (term frequency-inverse document frequency) weighting method. Tf-idf is a term weight calculation technique used commonly in the information retrieval and text mining research areas. In [45], tf-idf is defined as a statistical measure, which is used to evaluate how important a word is to a document in a collection by checking the distribution and frequency of the word's occurrences.

In our context, the tf-idf value measures how important a gene mutation is to a sample in a collection of samples. Mutations in genes that are found in most samples have low tf-idf values, whereas genes with rare mutations are granted higher weights. With this strategy, we aim to increase the impact of the existence of rare mutations and suppress the effects of common mutations in the classification task, since common mutations may not be a sign of a disease.

In this model, instead of binary values, the calculated $tf\text{-}idf$ weights of the genes are used as feature values. The main equation of $tf\text{-}idf$ is presented as

$$tf\text{-}idf_{g,s} = tf_{g,s} * idf_g. \tag{3.1}$$

The $tf\text{-}idf$ value for a gene g and sample s is the multiplication of tf, that is term frequency, and idf, that is inverse document frequency, values. The tf value for a gene g and sample s is taken as the count of mutations of gene g in sample s. The higher the number of mutations for a gene in a sample, the more tf weight is assigned to this gene. The df, that is document frequency, value for a gene g is taken as the count of samples in the collection that contain mutations of gene g. For a sample collection of size N, idf of gene g is calculated as shown in

$$idf_g = \log(N/df_g). \tag{3.2}$$

Intuitively, the more samples in the collection have mutations in gene g, the less discriminating power this gene will have as a feature in cancer type classification. So it is assigned a lower idf score.

3.2.2.4. Tf-rf Based Mutation Model. A mutation can be rare in the collection, however, it may be effective for samples with particular cancer types. In order to account for the class information, $tf\text{-}rf$ (term frequency-relevance frequency) based data representation is adapted. Similarly to $tf\text{-}idf$, $tf\text{-}rf$ is also used in information retrieval and text mining. Unlike $tf\text{-}idf$, $tf\text{-}rf$ is a supervised statistical measure proposed in [46]. It is used to evaluate how important a word is to a class of documents in a collection. In $tf\text{-}rf$, a word may have different weight values for different classes.

In our context, the $tf\text{-}rf$ value measures how important a gene mutation is to a sample by using the information of its class label. If the particular gene mutation is encountered more in one class compared to the other classes, the corresponding rf and $tf\text{-}rf$ values are higher than for the other classes. As shown in

$$tf\text{-}rf_{g,s} = tf_{g,s} * rf_{g,c}, \tag{3.3}$$

the $tf\text{-}rf$ value for a gene g and sample s is the multiplication of tf, that is term frequency, and rf, that is relevance frequency, values. The tf value is computed in the same way as in $tf\text{-}idf$. The rf value of gene g and class c is calculated as in

$$rf_{g,c} = \log(2 + a/\max(1, b)), \tag{3.4}$$

where a is the number of samples in class c which contain mutation in gene g, and b is the number of samples in other classes which contain mutation in gene g.

3.2.2.5. BM25-tf-idf Based Mutation Model. BM25, often called Okapi, is a ranking function used by search engines to rank matching documents according to their relevance to a given search query [47]. For our task of weighting genes based on mutation information, the term frequency definition in BM25 is used instead of the classic term frequency in $tf\text{-}idf$. As shown in

$$bm25\text{-}tf\text{-}idf_{g,s} = bm25\text{-}tf_{g,s} * idf_g, \tag{3.5}$$

$bm25\text{-}tf\text{-}idf$ value for a gene g and sample s is the multiplication of $bm25\text{-}tf$, that is BM25 definition of term frequency, and idf, that is inverse document frequency, values.

$bm25\text{-}tf$ value for a gene g and sample s is calculated as in

$$bm25\text{-}tf_{g,s} = ((k+1) * tf_{g,s})/(k * ((1-b) + b * (L_s/L_{ave})) + tf_{g,s}). \quad (3.6)$$

In this equation, L_s and L_{ave} are the length of sample s and the average sample length for the whole collection, respectively. We model the samples with the same features. Therefore, in our representation model, L_s is equal to L_{ave}. When we use this equality, Equation (3.6) is simplified to

$$bm25\text{-}tf_{g,s} = ((k+1) * tf_{g,s})/(k + tf_{g,s}). \quad (3.7)$$

k is used as a smoothing parameter for tf. The idf definition is the same as in $tf\text{-}idf$.

3.2.2.6. BM25-tf-rf Based Mutation Model. For $bm25\text{-}tf\text{-}rf$, the term frequency definition in BM25 is used instead of the classic term frequency in $tf\text{-}rf$. In

$$bm25\text{-}tf\text{-}rf_{g,s} = bm25\text{-}tf_{g,s} * rf_{g,c}, \quad (3.8)$$

the $bm25\text{-}tf\text{-}rf$ value for a gene g and sample s is the multiplication of $bm25\text{-}tf$, that is BM25 definition of term frequency, and rf, that is relevance frequency, values. The $bm25\text{-}tf$ value is computed in the same way as in $bm25\text{-}tf\text{-}idf$. The rf definition is the same as in $tf\text{-}rf$.

The effect of the smoothing parameter k is illustrated in Figure 3.2. In this chart, the tf and $bm25\text{-}tf$ values for different values of k are shown when the number of gene mutations changes in the range from 1 to 10. The figure demonstrates that, the tf values, which are represented by empty circles, keep increasing as the number of mutations increases. Even a point mutation may be significant for the occurrence of a certain disease. Therefore, a gene with n mutations is not necessarily n times more important than a gene with 1 mutation for disease detection. As shown in Figure 3.2, the smoothing parameter k in $bm25\text{-}tf$ dampens the effect of high tf values.

3.2.3. Implementation and Experiment Design

3.2.3.1. Machine Learning Models. A wide range of machine learning algorithms are applied to investigate the effects of the proposed mutation based DNA representation models in the task of disease classification. Naive Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic Regression (LR), One-Layer-Perceptron (Perceptron) and Feed-Forward Multilayer Neural Network (NN) are run on the prepared datasets.

For the Feed Forward NN, the model is composed of two or more fully connected layers. Except the last layer, the number of nodes is halved at each layer. If the first layer has N units, then the second layer has N/2 units. With this strategy, each layer represents the information from the previous layer with less units. After each fully connected layer, a dropout is applied. As there are 10 classes, the last layer has 10 nodes with softmax activation function. Categorical cross-entropy is employed as the loss function. The number of epochs is 50 and the batch size is 50.

All experiments are implemented with Python. For the traditional machine learning algorithms, the scikit-learn library [48] is used. The feed forward network model is implemented with Keras [49] on Tensorflow [50] backend.

3.2.3.2. Evaluation Strategy. The input datasets are first divided into 80% training and 20% test sets. Parameter tuning is accomplished using 10-fold cross-validation on the training set. Testing is also accomplished in 10-folds. In each fold, the model with the best parameters is trained with one of the training sets from the initial cross-validation experiment, which was performed over the 80% of the data, and testing is performed over the test set (the initially separated 20% of the data). By this strategy, the instances in the independent test set are not included in any step of either parameter tuning or training, and the effects of minor changes of the training data are also indicated. The reported results are the micro-averaged scores and standard deviations on the independent test set. Accuracy, f-score, precision, recall, false posi-

tive rate (FPR) and area under the receiver operating curve (roc-auc) are used as the performance measures.

3.2.3.3. Parameter Tuning for The Representation Models. For BM25-tf, a range of k values between 0.6 and 2 are used in the parameter tuning phase. The BM25-tf-rf representation model and the Feed Forward NN are used in the parameter tuning setup. The classification results for different values of k are presented in Table 3.2 and the best performing row is shown in bold. k=0.8 leads to the best accuracy and f-score values. Therefore, this value is used in our experiments for the BM25-tf calculations.

3.2.3.4. Parameter Tuning for The Classification Models. A parameter tuning phase is applied for each machine learning algorithm and data representation model. The best parameter set is used in the test phase. The default values are used for the parameters that are not tuned. NB and LR are applied with the default parameters. The range or list of values used in parameter tuning for the other machine learning algorithms are presented in Table 3.3. The best parameters for the classification models are presented in Table 3.4.

3.3. Results

3.3.1. Selection of Gene Sets

The success of the mutation based data representation schemes also depends on the selected gene list. The initial experiments are applied on the binary representation to observe the classification performances over the causal and full gene sets. LR, SVM with linear kernel and Perceptron are selected as pilot algorithms. As shown in Table 3.5, the accuracy values are between 33% and 37% for the binary causal gene dataset and between 66% and 69% for the binary full gene dataset. It is observed that, the addition of extra gene information nearly doubles the classification accuracy. The mutation data of the additional genes increased the accuracy. The dramatic increase

can be observed in f-score results too. This result can be interpreted as being an indication of the existence of new (currently unknown) causal genes. As the accuracy of the classification models are enhanced with additional genes, the rest of the experiments are applied on the full gene datasets.

3.3.2. Comparison of The Data Representation Models with Machine Learning Experiments

Machine learning algorithms are applied in order to explore the effects of the proposed data representations. Table 3.6, Table 3.7, Table 3.8 and Table 3.9 list the results of the machine learning experiments. The table is designed to compare the data representation models for each algorithm. The row with the best result is shown in bold for each algorithm and the overall best performance is highlighted with a surrounding box.

The accuracy scores of NB, KNN and SVM with polynomial kernel are below 57% and the f-score results are below 59%. The remaining 5 algorithms obtain accuracy and f-score levels above 60% (except SVM-rbf with tf-idf). We will focus on these better performing algorithms. The BM25-tf-rf representation scheme leads to the best accuracy and f-score results for all of the 5 algorithms. In addition, the BM25-tf-rf data representation results in nearly 2 to 4 percent accuracy and f-score improvement and 0.01 to 0.05 roc-auc improvement compared to the second best representation for all of the 5 successful algorithms. When we consider the FPR results, BM25-tf-rf leads to lowest values in all of the 5 successful algorithms.

For the Feed-Forward NN f-score results with binary and BM25-tf-rf representations, the paired t-test produces a $p\text{-}value < 2.5e{-}10$, and with the tf-rf and BM25-tf-rf representations, the paired t-test produces a $p\text{-}value < 1.1e-05$. These results show that the additional statistical information hidden in the BM25-tf-rf representation provides significant gain compared to the other representation models. When we compare the Feed-Forward NN f-score values for the C-score and BM25-tf-rf models, the paired

t-test produces a $p\text{-}value < 3.2e-06$. This significant difference states that, although the BM25-tf-rf scheme doesn't utilize the various properties of mutations that are expressed in C-score, it is more successful for the differentiation of cancer types with its class-based statistical approach for the mutations. We can conclude that the BM25-tf-rf scheme is a suitable representation tool for the mutation information in VCF files for the cancer type classification task.

The most accurate algorithm (76.44% accuracy and 76.95% f-score) is the Feed-Forward NN with the BM25-tf-rf representation scheme, despite the extra network cost. The precision and recall results for the NN on BM25-tf-rf representation are similar with the accuracy value. The roc-auc result is also the highest compared to the other results in Table 3.6, Table 3.7 and Table 3.8.

For the LR and NN f-score results with the BM25-tf-rf representation scheme, the paired t-test produces a $p\text{-}value < 2.35e-05$. For the Perceptron and NN results with BM25-tf-rf, the paired t-test produces a $p\text{-}value < 1.7e-06$. Thus, the multilayer feed forward neural network model is found to be significantly more accurate than the single layer perceptron and LR with BM25-tf-rf.

3.3.3. Class-Based Comparison of Experiment Results

BM25-tf-rf based representation leads to improved performance results compared to the other representations with almost all machine learning algorithms in our experiments. In addition, the multilayer feed forward neural network model achieves better cancer type classification performance compared to the other machine learning algorithms in our experiments for all data representations except tf-idf. Therefore, we used the results for the NN algorithm with the BM25-tf-rf representation model for further discussions on class-level performance.

Table 3.10 lists the class-based performance metrics. The cancer types are presented in the order of descending sample count. From this list, it is observed that the

classification performance differs for each class. The results show that the success level does not entirely depend on the number of samples in the dataset, as there are fewer samples for Skin cancer than Thyroid cancer, but the f-score for Skin cancer is higher. The proposed model doesn't perform well for cancer types such as Prostate and Thyroid. Other cancer types such as Lung, Breast, Colorectal, and Skin are classified with better f-scores. This suggests that there may be more distinctive and class specific mutations in these cancer types, which the BM25-tf-tr scheme can model successfully.

3.3.4. Location-Based Comparison of Gene Mutations

In the previous sections all exonic and intronic mutations in the dataset have been utilized to compare the data representation and classification models. The results have shown that the BM25-tf-rf is the best performing representation model and the multilayer feed forward neural network is the best performing classification model in our experiments. By using these best models, a new experiment setup is created to explore the effect of the location of the mutations in the classification result. The exonic and intronic mutations are used separately. The experiment results are presented in Table 3.11. When only the exonic mutations are used, the classification performance decreases dramatically to 54.56% accuracy and 55.52% f-score. This decrease can be dependent on the fact that only 15% of all mutations are exonic. When only intronic mutations are used, the classification performance decreases to 74.39% accuracy and 75.54% f-score. This relatively lower decrease in the performance can be explained by the vast majority of mutations being intronic. The paired t-test produces a $p\text{-}value < 4.8e-02$ for the f-score results of the Feed-Forward NN with BM25-tf-rf representation with only intronic and all mutations. The utilization of all exonic and intronic mutations for input representation leads to statistically significant improvement in f-score performance. Similar to recent studies stating that malignancy-driving mutations can also occur outside the coding region [51,52], our location based comparison results support the need for further research in non-coding variants.

3.4. Discussion

The main goal of genomic studies for diseases is to propose target or causal genes. BM25-tf-rf is found to be superior compared to binary, tf-idf and C-score for the representation of DNA mutations for the task of cancer classification. We further analyze our best model (BM25-tf-rf and Feed Forward NN) for the most effective genes in decision making.

In Figure 3.3 and Figure 3.4, the heat maps show the most effective genes in the NN model with BM25-tf-rf representation for the classification of the breast and lung cancer types, respectively. The heat maps are constructed by giving one hot vectors, where only one gene feature is set to 1 and the others are set to 0, as input to the 10-fold trained NN classifiers. The output values from the 10 output nodes corresponding to the probabilities for the 10 cancer types, are saved as a heat map for that fold. The final heat map is the average of the folds. Each output of the NN reflects the effect of the labelled gene (set to 1 in the input) to the prediction of the cancer types. If the probability value for a cancer type is high, this means that the labelled gene is more effective in the prediction of that cancer type, since in the NN calculations, the other features are cancelled as their input values are 0. If the probability value for a cancer type is low, this means that the labelled gene doesn't play an important role in the prediction of that cancer type. The genes that result in high output probability values for a cancer type are taken as more effective and active genes for the prediction of that cancer type.

For the specific heat maps in Figure 3.3 and Figure 3.4, the final heat map is sorted according to the selected class column and 20 genes, which have the highest impact on the prediction of the selected cancer type, are plotted. The values in the heat-maps represent how effective a gene is for the prediction of the corresponding cancer type. The lighter colors represent higher values which refer to more effective state.

In Figure 3.3, PRSS3, AQP3, HAPLN3, RASSF8, BRF1 and GSE1 are shown to be effective on classification of breast cancer and there is supporting evidence in the literature for their relatedness to the disease. PRSS3 was found to promote the growth of breast cancer cells [53]. AQP3 is studied in [54]. In [55], HAPLN3 was shown to be among the overexpressed genes for breast cancer. In [56], HAPLN3 was suggested to be involved in the development of breast cancer and to be a biomarker for the treatment of breast cancer. In [57], RASSF8 was proposed to be used in discrimination of benign and malignant breast tumors. In [58], they investigated whether BRF1 expression is increased in the samples of human breast cancer and their results indicated that it is overexpressed in most cases. In [59], it is reported that GSE1 is overexpressed in breast cancer and silencing of GSE1 significantly suppressed breast cancer cells.

The effect of BM25-tf-rf weighting is illustrated through an example for the HAPLN3 and BRF1 genes of two sample patients. Sample-79 is a breast cancer patient and Sample-164 is a brain cancer patient. HAPLN3 and BRF1 have been found relevant to breast cancer in the literature and they are both found to be among the most effective genes for our classification system. But there is no scientific evidence about the relationship of these genes and brain cancer. The BM25-tf-rf value for HAPLN3 in Sample-79 is 1.016021 whereas it is 0.794813 in Sample-164. There is a nearly 28 percent increase in the weight of this gene for the breast cancer sample. For BRF1, the BM25-tf-rf value in Sample-79 is 1.323967 whereas it is 1.085498 in Sample-164. There is a nearly 22 percent increase in the weight of this gene for the breast cancer sample. When we consider the Feed Forward NN classification model, the weights for these gene features is the same for both samples. Therefore, the distinction of predicted cancer type results arise from the difference in BM25-tf-rf weights. This difference effects the decision and helps to distinguish between cancer types.

In Figure 3.4, POLD1, COTL1, AXIN2, WIF1 and SLC2A1 (previous symbol GLUT1) are shown to be effective on classification of lung cancer. There are studies in the literature supporting the relatedness of these genes to lung cancer. POLD1 is studied in [60]. In [61] and [62], COTL-1 was proposed to be a biomarker or a

therapeutic target for small cell lung cancer (SCLC) patients. In [63], AXIN2 was found to play a major role in modulating lung cancer risk. It was shown that WIF1 had the potential as a methylation biomarker in the diagnosis of non small cell lung cancer (NSCLC) [64]. In [65], it is reported that lung squamous cell carcinoma, a major subtype of NSCLC, exhibits remarkably elevated glucose transporter GLUT1 expression.

Since there is literature evidence for the disease-relatedness of a subset of the most effective genes in classification using NN model with BM25-tf-rf, the other most effective genes that are not studied yet might also have causal roles in cancer development. According to these evidences, NN trained with the BM25-tf-rf representation of the mutations in the VCF files, can also be used for the purpose of finding new candidate genes for cancer types.

3.5. Final Remarks

Complex genomic diseases are caused by changes in DNA that alter cell behavior. The impact level of each mutation may be different for various diseases. In order to model this diversity, being inspired from the document representation techniques in the information retrieval domain, we proposed different mutation based statistical genomic data representation schemes.

We utilized VCF files, which contain mutation information in the DNA, for the classification of cancer types as a case study. Cancer, in general, results from a combination of several genomic alterations, which can be addressed in variant calls data. We evaluated the performance of the proposed data representation schemes with a wide range of machine learning algorithms. Our experiment results showed that BM25-tf-rf based representation is more successful at modeling VCF data compared to the binary, tf-idf and C-score based representation schemes. Each cancer type may develop as a result of different gene mutations. The supervised weighting approach of tf-rf successfully reflects this class-mutation relationship. The normalization effect of BM25-tf

further improves the classification performance of tf-rf. We investigated the most effective mutated genes in our proposed system for breast and lung cancers. A subset of the resulting genes have also been suggested as causal or target genes in previously published studies, which demonstrates that the proposed approach can also be used to recommend candidate genes.

The introduced data representation models are evaluated for the task of cancer type classification, which is an important problem in bioinformatics, since the appropriate treatment is determined according to the primary site. However, they can also be utilized for other genetic diseases, which we plan to investigate in our future studies.

Table 3.1. The list of cancer types and sample counts in the BOUN10CANCER dataset.

Cancer Type	**Sample Count**
Lung	1232
Breast	1080
Brain	1028
Kidney	734
Colorectal	656
Thyroid	504
Prostate	503
Skin	472
Stomach	441
Liver	378

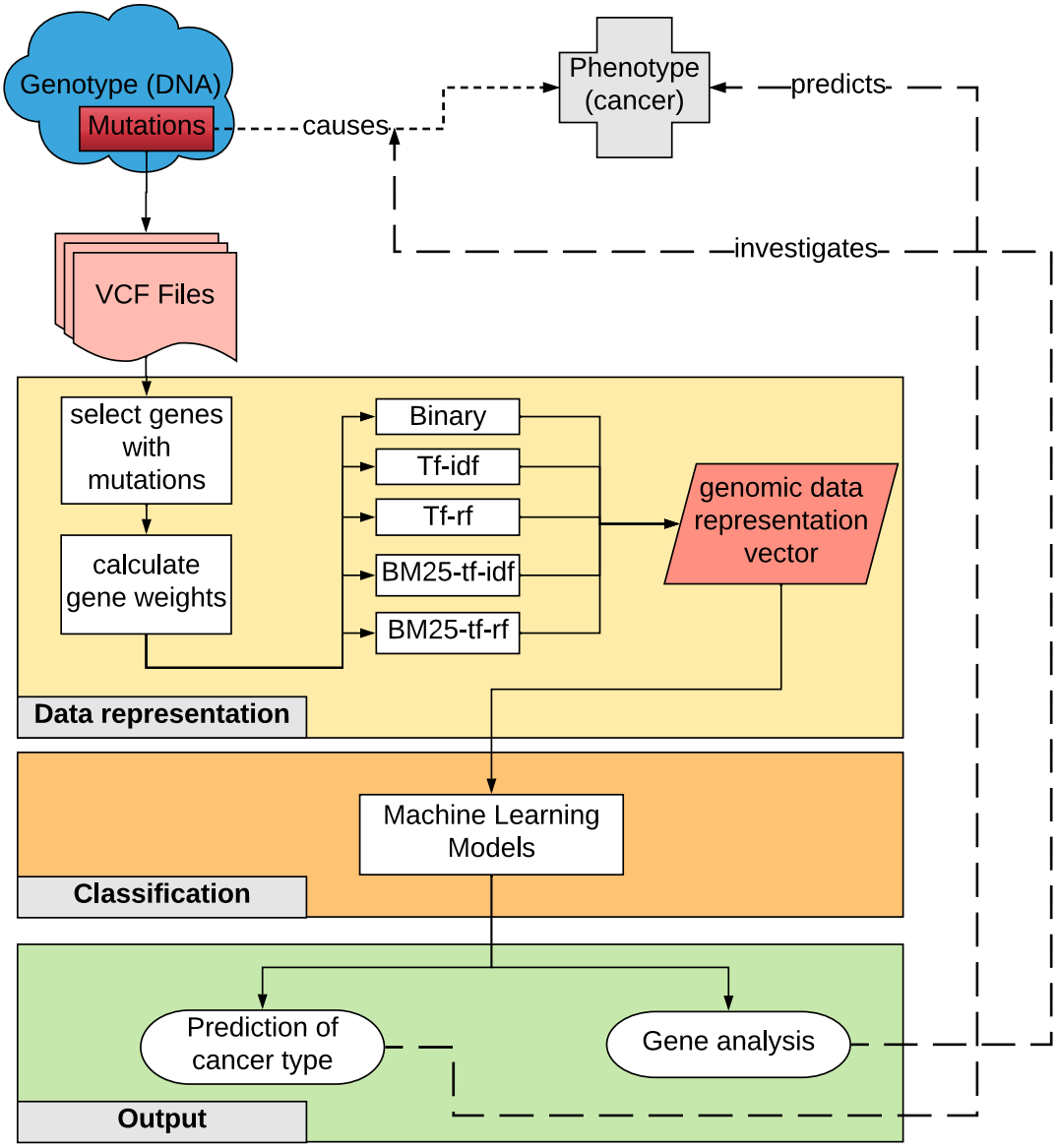


Figure 3.1. Overview of the proposed system.

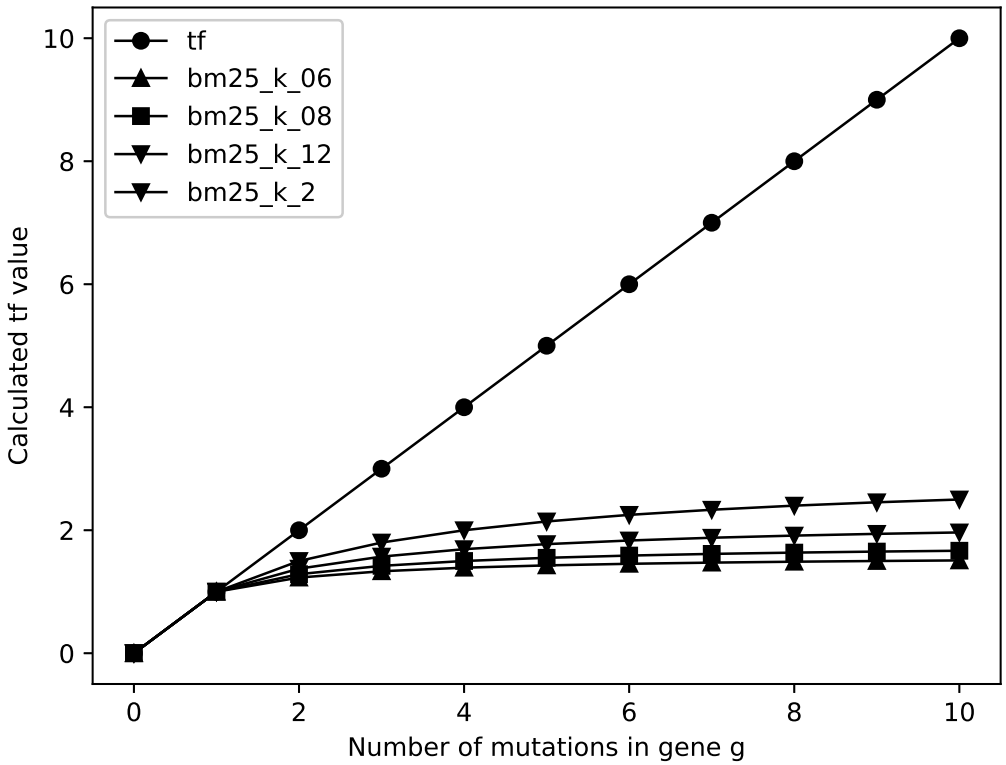


Figure 3.2. The effect of the smoothing parameter k in the BM25 calculations for term frequency.

Table 3.2. Parameter tuning results for the parameter k in the BM25-tf formula.

k	**Accuracy**	**F-Score**	**Precision**	**Recall**
0.6	75.20 ± 1.21	75.89 ± 1.14	76.59 ± 1.11	75.20 ± 1.21
0.8	**75.38 ± 1.02**	**76.18 ± 1.03**	**76.99 ± 1.10**	**75.38 ± 1.02**
1.0	75.24 ± 1.60	75.87 ± 1.55	76.51 ± 1.53	75.24 ± 1.60
1.2	74.60 ± 1.60	75.32 ± 1.49	76.07 ± 1.39	74.60 ± 1.60
1.4	74.24 ± 1.00	74.88 ± 0.76	75.54 ± 0.65	74.24 ± 1.00
1.6	74.43 ± 1.39	75.35 ± 1.28	76.30 ± 1.42	74.43 ± 1.39
1.8	74.21 ± 1.46	74.89 ± 1.40	75.58 ± 1.38	74.21 ± 1.46
2.0	74.73 ± 1.32	75.53 ± 1.21	76.36 ± 1.16	74.73 ± 1.32

Table 3.3. The range (or list) of parameters used in the parameter tuning phase for the classification models.

Algorithm	**Parameter**	**Range or Values**
KNN	k	$[2, 150]$
SVM	kernel	linear, polynomial, rbf
	polynomial degree	$[2, 5]$
	gamma	$[10^{-4}, 10^{-1}]$
	cost	$[10^{1}, 10^{4}]$
Perceptron	optimization function	Adam, SDG
	activation function	ReLU, tanh
	hidden layer size	$[10, 100]$
	max iterations	$[100, 300]$
Feed Forward NN	optimization function	Adam, SDG
	activation function	ReLU, tanh
	the number of layers	$[2, 6]$
	dropout rate	$[0.25, 0.5]$
	nodes in first layer	$[1024, 8192]$

Table 3.4. The best parameters found as a result of the parameter tuning phase for the classification models.

Algorithm	**Parameter**	**Value**	**Data Rep.**
KNN	k	50	binary
		10	c-score, tf-idf, tf-rf, bm25-tf-idf, bm25-tf-rf
SVM-poly	polynomial degree	3	binary, tf-idf, bm25-tf-idf
		2	c-score, tf-rf, bm25-tf-rf
SVM-rbf	gamma	10^{-4}	All
	cost	10^{3}	All
SVM-linear	gamma	10^{-4}	All
	cost	10^{2}	All
Perceptron	optimization function	SGD	binary
		Adam	c-score, tf-idf, tf-rf, bm25-tf-idf, bm25-tf-rf
	activation function	tanh	binary
		ReLU	c-score, tf-idf, tf-rf, bm25-tf-idf, bm25-tf-rf
	hidden layer size	100	All
	max iterations	200	binary
		300	c-score, tf-idf, tf-rf, bm25-tf-idf, bm25-tf-rf
Feed Forward NN	optimization function	Adam	All
	activation function	ReLU	All
	the number of layers	4	All
	dropout rate	0.25	All
	nodes in first layer	8192	All

Table 3.5. Machine learning experiment test results on the gene sets with the binary representation model.

Gene Set	**Algorithm**	**Data Rep.**	**Accuracy**	**F-Score**	**Precision**	**Recall**	**Roc-Auc**	**FPR**
causal	LR	binary	36.81 ± 0.45	36.36 ± 0.50	35.93 ± 0.52	36.81 ± 0.45	0.63 ± 0.03	9.03 ± 0.10
causal	SVM-linear	binary	33.53 ± 0.32	32.70 ± 0.99	31.92 ± 1.13	33.53 ± 0.32	0.62 ± 0.05	9.38 ± 0.11
causal	Perceptron	binary	36.74 ± 0.56	36.62 ± 0.83	36.52 ± 2.56	36.74 ± 0.56	0.63 ± 0.06	10.01 ± 0.10
all	LR	binary	67.19 ± 0.41	68.01 ± 0.01	68.01 ± 0.00	67.01 ± 0.01	0.78 ± 0.01	3.85 ± 0.07
all	SVM-linear	binary	68.46 ± 0.67	68.01 ± 0.01	69.01 ± 0.01	68.01 ± 0.01	0.78 ± 0.01	4.07 ± 0.09
all	Perceptron	binary	68.50 ± 0.48	69.01 ± 0.01	70.01 ± 0.01	68.01± 0.01	0.78 ± 0.03	4.07 ± 0.09

Table 3.6. Machine learning experiment test results on the data representation models of BOUN10CANCER dataset - 1.

Algorithm	**Data Rep.**	**Accuracy**	**F-Score**	**Precision**	**Recall**	**Roc-Auc**	**FPR**
NB	binary	33.84 ± 0.83	35.25 ± 0.95	37.04 ± 1.34	33.84 ± 0.83	0.62 ± 0.02	8.38 ± 0.11
	c-score	31.10 ± 0.86	32.72 ± 0.74	34.53 ± 1.43	31.10 ± 0.86	0.59 ± 0.01	8.61 ± 0.08
	tf-idf	33.34 ± 0.48	35.04 ± 0.60	37.03 ± 1.03	33.34 ± 0.48	0.62 ± 0.02	7.99 ± 0.07
	tf-rf	*38.14 ± 0.57*	*38.97 ± 0.87*	*40.08 ± 1.27*	*38.14 ± 0.57*	*0.65 ± 0.01*	*7.99 ± 0.10*
	bm25-tf-idf	32.50 ± 0.96	34.19 ± 0.87	36.08 ± 1.35	32.50 ± 0.96	0.60 ± 0.01	8.48 ± 0.10
	bm25-tf-rf	37.94 ± 0.63	38.99 ± 0.60	40.12 ± 1.24	37.94 ± 0.63	0.62 ± 0.01	7.91 ± 0.10
KNN	binary	11.54 ± 0.85	16.87 ± 0.66	31.46 ± 2.54	11.54 ± 0.85	0.50 ± 0.04	7.41 ± 0.04
	c-score	15.87 ± 0.63	22.60 ± 0.44	39.27 ± 4.21	15.87 ± 0.63	0.53 ± 0.01	7.96 ± 0.07
	tf-idf	*34.96 ± 0.66*	*37.35 ± 0.95*	*38.92 ± 0.69*	*34.96 ± 0.66*	*0.62 ± 0.03*	*8.10 ± 0.04*
	tf-rf	19.29 ± 0.44	22.23 ± 0.61	40.29 ± 0.82	19.29 ± 0.44	0.55 ± 0.02	7.57 ± 0.07
	bm25-tf-idf	12.72 ± 1.23	20.05 ± 0.58	47.32 ± 5.85	12.72 ± 1.23	0.51 ± 0.01	8.17 ± 0.37
	bm25-tf-rf	11.91 ± 1.13	19.21 ± 0.50	49.74 ± 1.58	11.91 ± 1.13	0.51 ± 0.01	7.88 ± 0.17

Table 3.7. Machine learning experiment test results on the data representation models of BOUN10CANCER dataset - 2.

Algorithm	**Data Rep.**	**Accuracy**	**F-Score**	**Precision**	**Recall**	**Roc-Auc**	**FPR**
SVM-poly	binary	17.50 ± 0.00	5.21 ± 0.00	3.06 ± 0.00	17.50 ± 0.00	0.53 ± 0.00	16.34 ± 0.00
	c-score	*56.14 ± 0.44*	*58.90 ± 0.39*	*61.96 ± 0.46*	*56.14 ± 0.44*	*0.73 ± 0.01*	*5.33 ± 0.06*
	tf-idf	17.50 ± 0.00	5.21 ± 0.00	3.06 ± 0.00	17.50 ± 0.00	0.53 ± 0.00	16.35 ± 0.00
	tf-rf	55.51 ± 0.55	56.52 ± 0.65	61.40 ± 0.53	55.51 ± 0.55	0.71 ± 0.03	5.16 ± 0.05
	bm25-tf-idf	36.36 ± 0.66	42.64 ± 0.75	51.56 ± 0.89	36.36 ± 0.66	0.62 ± 0.01	7.93 ± 0.08
	bm25-tf-rf	53.41 ± 0.27	51.46 ± 0.27	63.95 ± 0.65	53.41 ± 0.27	0.66 ± 0.01	7.38 ± 0.04
SVM-rbf	binary	66.71 ± 0.36	67.01 ± 0.00	68.01 ± 0.00	67.01 ± 0.01	0.78 ± 0.01	4.04 ± 0.09
	c-score	57.35 ± 0.30	61.31 ± 0.28	65.86 ± 1.10	57.35 ± 0.30	0.72 ± 0.01	7.09 ± 0.05
	tf-idf	50.92 ± 0.19	44.26 ± 0.20	51.64 ± 0.19	50.92 ± 0.19	0.69 ± 0.02	8.30 ± 0.03
	tf-rf	69.53 ± 0.71	69.82 ± 0.72	70.75 ± 0.71	69.53 ± 0.71	0.78 ± 0.03	3.64 ± 0.09
	bm25-tf-idf	66.17 ± 0.56	66.61 ± 0.60	67.20 ± 0.62	66.17 ± 0.56	0.78 ± 0.01	4.40 ± 0.07
	bm25-tf-rf	*73.77 ± 0.46*	*74.00 ± 0.46*	*74.96 ± 0.40*	*73.77 ± 0.46*	*0.83 ± 0.01*	*3.20 ± 0.07*

Table 3.8. Machine learning experiment test results on the data representation models of BOUN10CANCER dataset - 3.

Algorithm	**Data Rep.**	**Accuracy**	**F-Score**	**Precision**	**Recall**	**Roc-Auc**	**FPR**
SVM-linear	binary	68.46 ± 0.67	68.01 ± 0.01	69.01 ± 0.01	68.01 ± 0.01	0.78 ± 0.01	4.07 ± 0.09
	c-score	71.91 ± 0.44	72.46 ± 0.45	73.02 ± 0.44	71.91 ± 0.44	0.82 ± 0.01	3.50 ± 0.09
	tf-idf	69.54 ± 0.66	69.01 ± 0.01	70.01 ± 0.01	69.01 ± 0.01	0.78 ± 0.01	3.94 ± 0.06
	tf-rf	68.80 ± 0.62	68.01 ± 0.01	69.51 ± 0.01	69.01 ± 0.01	0.78 ± 0.01	3.74 ± 0.09
	bm25-tf-idf	66.26 ± 0.58	66.35 ± 0.60	67.94 ± 0.66	66.26 ± 0.58	0.78 ± 0.01	4.31 ± 0.07
	bm25-tf-rf	*73.44 ± 0.43*	*73.66 ± 0.45*	*74.63 ± 0.41*	*73.44 ± 0.43*	*0.83 ± 0.01*	*3.24 ± 0.07*
LR	binary	67.19 ± 0.41	68.01 ± 0.01	68.01 ± 0.00	67.01 ± 0.01	0.78 ± 0.01	3.85 ± 0.07
	c-score	73.50 ± 0.64	73.89 ± 0.92	74.29 ± 0.66	73.50 ± 0.64	0.83 ± 0.01	3.40 ± 0.08
	tf-idf	63.17 ± 0.30	60.01 ± 0.00	66.01 ± 0.01	63.01 ± 0.00	0.74 ± 0.01	5.68 ± 0.04
	tf-rf	71.51 ± 0.46	72.01 ± 0.01	73.01 ± 0.01	71.01 ± 0.01	0.81 ± 0.01	3.24 ± 0.07
	bm25-tf-idf	67.80 ± 0.45	68.20 ± 0.47	68.61 ± 0.53	67.80 ± 0.45	0.79 ± 0.01	4.09 ± 0.06
	bm25-tf-rf	*74.99 ± 0.41*	*75.19 ± 0.38*	*75.96 ± 0.37*	*74.99 ± 0.41*	*0.83 ± 0.01*	*3.03 ± 0.06*

Table 3.9. Machine learning experiment test results on the data representation models of BOUN10CANCER dataset - 4.

Algorithm	**Data Rep.**	**Accuracy**	**F-Score**	**Precision**	**Recall**	**Roc-Auc**	**FPR**
Perceptron	binary	68.50 ± 0.48	69.01 ± 0.01	70.01 ± 0.01	68.01± 0.01	0.78 ± 0.03	4.07 ± 0.09
	c-score	71.64 ± 1.54	71.76 ± 1.87	71.89 ± 1.38	71.64 ± 1.54	0.81 ± 0.01	3.67 ± 0.24
	tf-idf	70.23 ± 0.40	70.01 ± 0.00	70.01 ± 0.01	70.01 ± 0.01	0.79 ± 0.01	3.83 ± 0.05
	tf-rf	72.07 ± 1.86	72.01 ± 0.02	74.01 ± 0.01	72.01 ± 0.02	0.82 ± 0.02	3.29 ± 0.12
	bm25-tf-idf	65.52 ± 0.52	65.97 ± 0.52	66.44 ± 0.56	65.52 ± 0.52	0.78 ± 0.01	4.48 ± 0.08
	bm25-tf-rf	*74.15 ± 0.51*	*74.48 ± 0.56*	*75.46 ± 0.56*	*74.15 ± 0.51*	*0.83 ± 0.01*	*3.07 ± 0.10*
Feed-ForwardNN	binary	69.00 ± 0.76	69.52 ± 0.70	71.00 ± 0.52	69.00 ± 0.81	0.79 ± 0.02	3.65 ± 0.17
	c-score	73.74 ± 0.88	74.07 ± 0.73	74.41 ± 0.67	73.74 ± 0.88	0.84 ± 0.02	3.27 ± 0.24
	tf-idf	62.91 ± 0.79	63.32 ± 0.70	65.04 ± 0.52	62.91 ± 0.83	0.73 ± 0.02	4.00 ± 0.10
	tf-rf	74.13 ± 1.33	74.17 ± 1.47	75.43 ± 1.07	74.13 ± 1.40	0.85 ± 0.02	3.07 ± 0.24
	bm25-tf-idf	68.18 ± 1.83	68.79 ± 1.28	69.42 ± 0.76	68.18 ± 1.83	0.82 ± 0.02	4.07 ± 0.54
	bm25-tf-rf	**76.44 ± 0.66**	**76.95 ± 0.68**	**77.48 ± 0.78**	**76.44 ± 0.66**	**0.86 ± 0.02**	**2.75 ± 0.13**

Table 3.10. Class based experiment test results with NN on full gene BM25-tf-rf dataset.

Cancer Type	**F-Score**	**Precision**	**Recall**	**FPR**
Lung	85.47 ± 1.20	88.03 ± 2.00	83.16 ± 1.94	2.42 ± 0.48
Breast	95.92 ± 1.81	94.23 ± 2.41	97.69 ± 1.44	1.09 ± 0.47
Brain	69.80 ± 1.23	64.19 ± 3.47	77.13 ± 1.46	2.61 ± 2.32
Kidney	68.51 ± 1.14	73.59 ± 3.48	64.23 ± 2.22	2.72 ± 0.48
Colorectal	88.89 ± 1.92	88.21 ± 3.45	89.93 ± 2.89	1.28 ± 0.66
Thyroid	51.40 ± 3.35	47.86 ± 4.29	56.54 ± 4.43	14.79 ± 1.25
Prostate	39.80 ± 2.28	37.97 ± 4.03	42.32 ± 2.41	15.36 ± 1.03
Skin	89.56 ± 1.21	95.66 ± 3.49	84.38 ± 2.09	1.29 ± 0.26
Stomach	60.30 ± 2.33	74.45 ± 4.41	51.86 ± 4.50	10.31 ± 0.84
Liver	71.51 ± 2.20	83.98 ± 3.67	63.28 ± 4.00	7.75 ± 0.47

Table 3.11. Machine learning experiment test results on the separated exonic and intronic mutations.

Mutation Set	**Accuracy**	**F-Score**	**Precision**	**Recall**	**Roc-Auc**	**FPR**
exonic	54.56 ± 1.18	55.52 ± 0.96	56.52 ± 0.83	54.56 ± 1.18	0.67 ± 0.01	5.44 ± 0.17
intronic	74.39 ± 1.58	75.54 ± 1.30	76.74 ± 1.10	74.39 ± 1.58	0.83 ± 0.01	2.91 ± 0.33
all	76.44 ± 0.66	76.95 ± 0.68	77.48 ± 0.78	76.44 ± 0.66	0.86 ± 0.02	2.75 ± 0.13

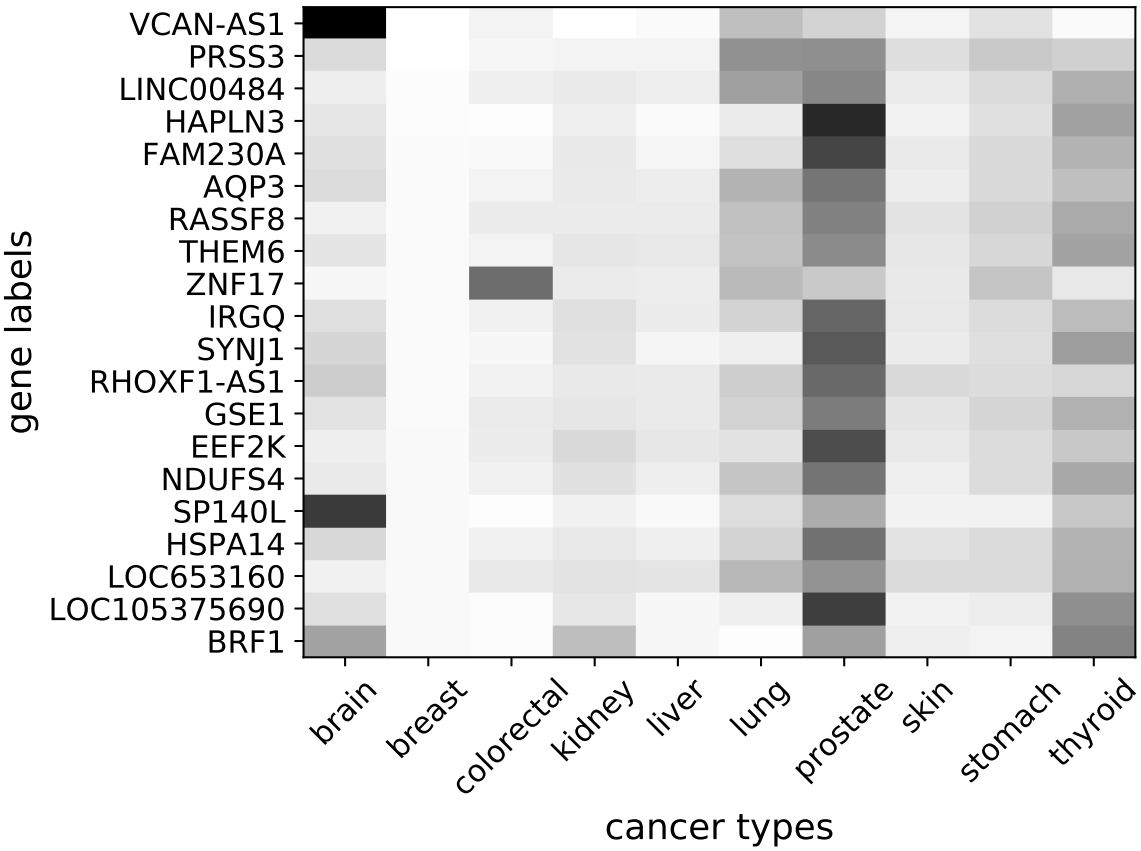


Figure 3.3. The heat map of the most effective genes in NN with BM25-tf-rf model for breast cancer. A light colored region for a gene and a cancer type can be interpreted as the gene is more effective in the decision of the cancer type. A dark colored region corresponds to less effective state.

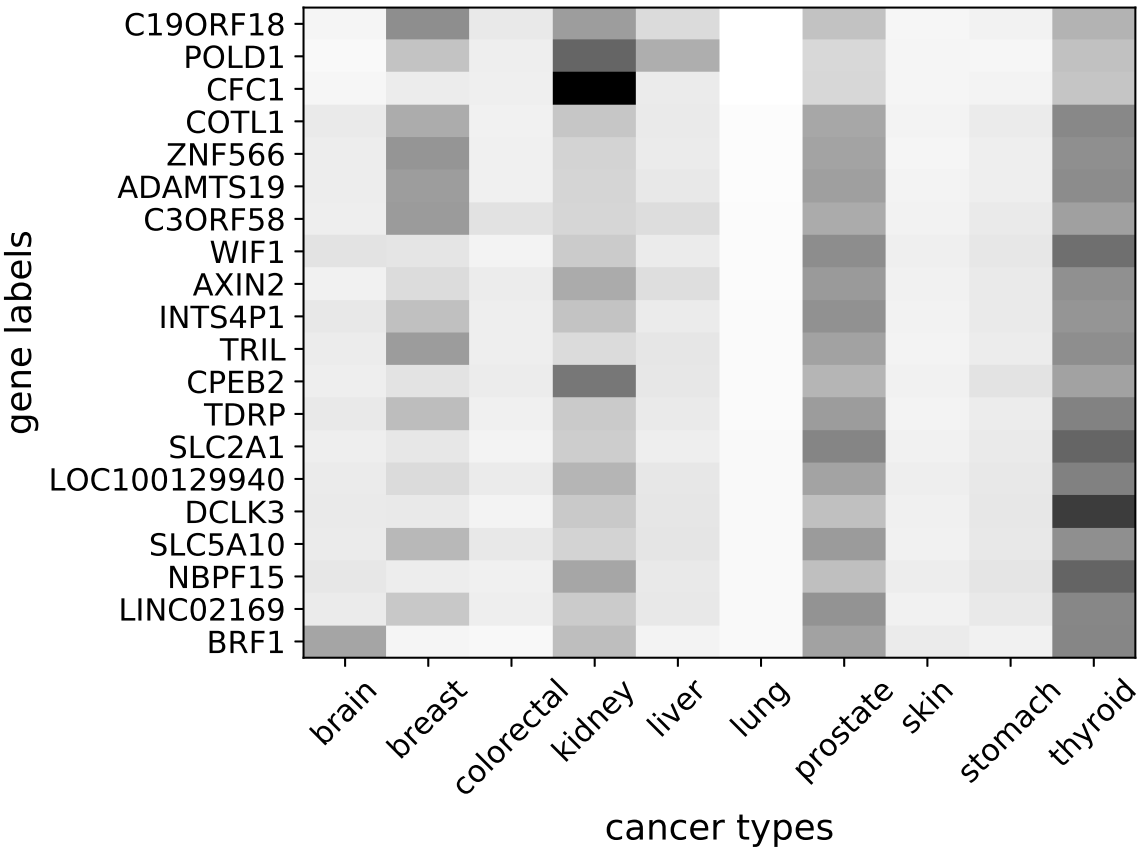


Figure 3.4. The heat map of the most effective genes in NN with BM25-tf-rf model for lung cancer. A light colored region for a gene and a cancer type can be interpreted as the gene is more effective in the decision of the cancer type. A dark colored region corresponds to less effective state.

4. GENE SELECTION METHOD FOR GENE EXPRESSION DATA

Cancer is a poligenetic disease with each cancer type having a different mutation profile. Genomic data can be utilized to detect these profiles and to diagnose and differentiate cancer types. Variant calling provide mutation information. Gene expression data reveal the altered cell behaviour. The combination of the mutation and expression information can lead to accurate discrimination of different cancer types. In this study, we transferred the information of the most effective mutations from our previous study (Chapter 3) as a novel gene selection method for gene expression data. We tested the proposed method in order to diagnose and differentiate cancer types. It is a disease specific method as both the mutations and expressions are filtered according to the selected cancer types. Our experiment results show that the proposed gene selection method leads to similar or improved performance metrics compared to classical feature selection methods and curated gene sets. The details of this study are presented in the following sections.

4.1. Background

Cancer is among the leading causes of death worldwide [6]. It is a group of diseases and each cancer type is labeled by the primary area of the body where the cancer cells arise. A different set of causal genes leads to each cancer type and the disease emerges from the combination of various mutations of these genes [6]. The cancer treatment is planned according to the driving mutations. The unknown or wrong analysis of these mutations lead to incorrect treatments and this is one of the major problems for cancer patients. Genomic data can be utilized for diagnosis of the disease and for recognizing different types. Genomic tests reveal the gene mutations that may be driving a cancer's behavior. This information helps doctors while deciding on the patient's personal treatment [66].

Driving mutations are located by detailed analysis on genomic data. Whole genome sequences and variant calling are utilized for mutation analysis [67–69]. Both coding and non-coding regions of the DNA are analyzed for the discovery of mutational signatures of cancer types.

Besides comprehensive statistical analysis, machine learning algorithms may help to detect the driving mutations. A commonly used data type for cancer classification is gene expression data. A number of studies have utilized gene expression data and addressed the classification of cancer types [27–31]. A major challenge of using gene expression data is the small sample size with high dimensionality. There may be thousands of genes in each sample but only a few of them are effective on the target disease, and most of them are irrelevant [70]. Gene selection methods are commonly applied prior to classification to overcome the high dimensionality problem [33, 34]. Hovewer, the feature selection step may eliminate genes that in general have minor effects on disease generation while still being significant for the diagnosis of particular cancer types for some patients. In addition, the irrelevant genes add noise and reduce the classifier performance for machine learning approaches [71, 72].

In this study, we propose a novel gene selection method targeting gene expression data for the task of cancer type classification. In a previous study [8], we utilized the mutation information in variant call format (VCF) [5] files. The most effective genes in the discrimination of cancer types are identified. In this study, these most effective genes within VCF data are employed for gene selection on gene expression data. The proposed method is compared to computational based and manually curated gene signature lists. The most important aspect of this method is that every step is disease specific and can be adapted to any genomic disease.

Our work brings the following contributions:

(i) A novel and disease/trait specific gene selection method is proposed and tested.
(ii) The valuable information in DNA mutations are transferred and used with gene

expression data.

(iii) Similar and slightly improved classification results are achieved compared to computational based and manually curated methods.

(iv) This system can be applied to any genomic disease or trait.

4.2. Methods

4.2.1. Dataset

We utilized the gene expression files in FPKM (fragments per kilobase million) format and the VCF files for the samples listed in CAMDA 2019 Hi-Res Cancer Data Integration Challenge [73]. All files are downloaded from The Cancer Genome Atlas Project (TCGA) [11]. The challenge includes samples for three cancer types: Breast, Lung Adenocarcinoma (Lung) and Kidney Renal Clear Cell Carcinoma (Kidney). We selected samples both having FPKM and VCF files. The list of cancer types and the sample counts for each cancer type are provided in Table 4.1.

4.2.2. Feature Selection for Gene Expression

For baseline, we used the whole gene list in gene expression files. There are $60,483$ distinct genes in the whole gene expression set. For feature selection, we applied SelectKBest from scikit-learn library [48], minimum Redundancy Maximum Relevance (mRMR) with pymrmr library [74] and Relief with Weka [75, 76]. SelectKBest is a feature selection method which selects features according to the highest scores of a selected scoring function. mutual_info_classif from scikit-learn library is used as scoring function. It scores features according to their mutual dependencies. mRMR is a feature selection method which chooses a feature subset by considering a trade-off between relevance and redundancy. Relief is a feature selection method which returns a feature subset based on relevance. We applied these methods with different numbers of features; $1,000$, $5,000$ and $10,000$ for SelectKBest and Relief and $1,000$ and $5,000$ for mRMR.

Another feature selection method is to use a gene signature list. The Hallmark gene set collection [77] is generated by a hybrid approach that combines an automated computational procedure with manual expert curation. It consists of multiple gene sets and displays the discriminating behavior across a number of test datasets. In order to benefit information from all of the source datasets, we used the union of all Hallmark gene sets as feature list. This results in $4,266$ gene features.

The last feature selection method of this study is the one that we propose. An overview of the proposed system is shown in Figure 4.1. In our previous study [8], we employed VCF files for cancer type classification. An impressive output of that study was the list of most effective genes in decision making. Most of these genes were found to be proposed as target genes in the literature. For a novel feature selection method, we combined the most effective genes from our previous study for the three cancer types. We selected $3,000$ and $3,500$ most effective genes for each cancer type and combine them to curate a feature set that represents all three cancer types. For the $3,000$ most effective genes, the final gene list has $6,752$ genes. For the $3,500$ most effective genes, the final gene list has $7,741$ genes.

4.2.3. Implementation of Machine Learning Methods and Experiment Design

All experiments are implemented with Python and Weka. For the machine learning algorithms, the scikit-learn and pymrmr libraries are used. We applied Logistic Regression (LR) on the curated datasets. Each test is applied with 5-fold cross validation. The reported results are the micro-averaged scores and standard deviations on the applied cross-validation folds. Accuracy, f-score, false positive rate (FPR), area under the receiver operating curve (roc-auc) and Matthews correlation coefficient (MCC) are used as the performance measures.

4.3. Results and Discussion

4.3.1. Comparison with a Previous Study

For baseline, the whole gene set in FPKM files are used in classification task. In order to compare with another data type, we utilized our previous study that operates on VCF data in order to classify cancer types. We applied and compared a number of statistical representation methods in that study. The best performing representation method was BM25-tf-rf. Therefore, we applied this model for this problem. The LR experiment results for FPKM and VCF datasets are presented in Table 4.2.

The number of features in FPKM dataset is four times that of VCF dataset. Despite the extra cost these features cause for the classification model, the performance metrics are improved with this dataset. The accuracy result for FPKM dataset is 99.46% whereas it is 93.70% for VCF dataset. The f-score result for FPKM dataset is also 99.46% whereas it is 93.62% for VCF dataset. When we consider FPR and MCC results, the difference between two datasets are more clearly observed. The FPR value for FPKM dataset is 0.40% whereas it is 3.60% for VCF dataset. The MCC result for FPKM dataset is 99.09% whereas it is 89.31% for VCF dataset. According to these results, we will utilize FPKM files for further experiments.

4.3.2. Gene Selection Results

The use of all genes leads to good results in classification task. But it also increases the computational cost. Therefore, we applied a number of gene selection methods in order to create more FPKM based datasets. The LR experiment results using these datasets are presented in Table 4.3.

When we applied SelectKBest with $10,000$ features, the accuracy and f-score values increase slightly to 99.57% compared to all features. But when we applied the same feature selection method with less features, the classification performance decreases as

the number of features decreases. mRMR algorithm produces less accurate results with accuracy and f-score values as 98.98% with 5,000 features. With an opposite tendency compared to SelectKBest, Relief leads to improved classification performance with less features. Accuracy and f-score values as 99.46% are achieved with Relief method with 1,000 features. When we compare these three feature selection algorithms, SelectKBest with 10,000 features outperforms other two methods.

We employed another feature selection method in order to reduce the feature count even more. When we consider a combination of all Hallmark gene sets for our feature list, the experiment results show that a similar performance can be achieved with SelectKBest with 10,000. But the number of features is less than the half.

Hallmark gene sets are known and used for years now. They depend on the previously curated gene sets. We further attempt to create a gene selection method that only depends on the data itself. Therefore, we selected most effective genes for the three cancer types from our previous study which employs VCF files. By this method, the information hidden in mutations are transferred to gene expression data. When we selected 3,000 genes for each cancer type and utilized the union of them, the resulting dataset leads to similar performance in classification with the Hallmark gene set. When we selected 3,500 genes for each cancer type and used the union of them, there is a slight improvement in the performance results. The resulting accuracy and f-score is 99.68% and MCC value is 99.46%. The FPR also decreases slightly to 0.24%. According to these results, the proposed VCF based gene selection method leads to similar or improved performance as the number of most effective genes is adjusted. As the disease is caused by mutations in DNA, it is reasonable to use these mutations in order to select effective genes and further analyze their expression levels. Our experiment results support this idea.

The comparison of f-score values and feature counts can be observed in more detail in Figure 4.2. The most effective methods can be taken as the ones with f-score values above 99.5%. These are SelectKBest with 10,000 features, Hallmark and VCF

based methods. Although VCF based gene selection method with $3,500$ genes doesn't provide the least feature count, it produces a slight improvement in f-score compared to the most successful methods in this study.

4.4. Final Remarks

Changes in DNA alter cell behavior and cause genomic diseases. Each genomic disease occurs in consequence of a different mutation profile. Besides the existence of a gene mutation, its effect can also be analyzed by the expression levels of genes. In this study, we provide the union of the existence and expression level information of mutated genes and propose a novel gene selection method. We utilized the mutation information in DNA for the selection of relevant genes in the gene expression data.

Based on our previous study [8], we selected gene features in expression data with the help of the most effective gene mutations for each cancer type. By this method, the valuable information in variant calling files are transferred and used with a different genomic data type. Although the number of samples is very limited in this study, this novel gene selection method leads to similar and slightly improved classification results compared to classic feature selection methods as SelectKBest, mRMR, Relief and curated gene sets as Hallmark. The proposed feature selection method is specific to the target disease as the effective genes are decided accordingly. Therefore, this system can be adapted and applied to any genomic disease or trait.

Table 4.1. The list of cancer types and sample counts in our dataset.

Cancer Type	**Sample Count**
Breast	1020
Lung	507
Kidney	330

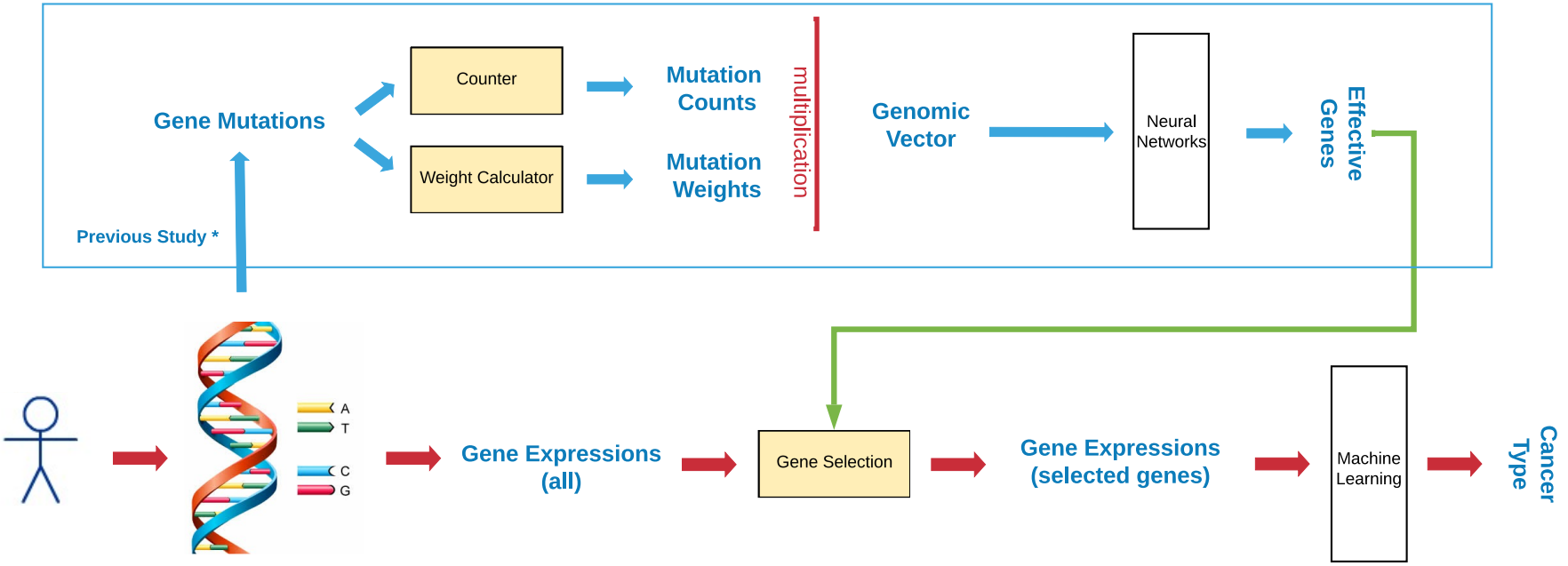


Figure 4.1. Overview of the proposed system.

Table 4.2. Machine learning experiment test results of gene expression and VCF data.

Dataset	**Feature Count**	**Accuracy**	**F-Score**	**Roc-Auc**	**FPR**	**MCC**
FPKM all features	60483	99.46 ± 0.42	99.46 ± 0.42	99.45 ± 0.50	0.40 ± 0.31	99.09 ± 0.71
VCF BM25-tf-rf	16383	93.70 ± 1.02	93.62 ± 1.07	93.26 ± 1.27	3.60 ± 0.71	89.31 ± 1.76

Table 4.3. Machine learning experiment test results of gene expression data.

Feature Sel.	**Feature Count**	**Accuracy**	**F-Score**	**Roc-Auc**	**FPR**	**MCC**
FPKM all features	60483	99.46 ± 0.42	99.46 ± 0.42	99.45 ± 0.50	0.40 ± 0.31	99.09 ± 0.71
SelectKBest	10000	99.57 ± 0.13	99.57 ± 0.13	99.61 ± 0.18	0.29 ± 0.13	99.27 ± 0.22
SelectKBest	5000	99.30 ± 0.27	99.30 ± 0.27	99.47 ± 0.22	0.46 ± 0.24	98.82 ± 0.46
SelectKBest	1000	98.92 ± 0.38	98.92 ± 0.38	99.13 ± 0.41	0.71 ± 0.32	98.18 ± 0.64
mRMR	5000	98.98 ± 0.43	98.98 ± 0.43	99.08 ± 0.33	0.73 ± 0.23	98.28 ± 0.72
mRMR	1000	98.44 ± 0.52	98.45 ± 0.52	98.63 ± 0.56	0.82 ± 0.32	97.37 ± 0.88
Relief	10000	98.44 ± 0.31	98.45 ± 0.31	98.51 ± 0.45	0.82 ± 0.28	97.37 ± 0.53
Relief	5000	99.30 ± 0.36	99.30 ± 0.36	99.34 ± 0.37	0.27 ± 0.14	99.34 ± 0.37
Relief	1000	99.46 ± 0.17	99.46 ± 0.17	99.54 ± 0.16	0.39 ± 0.14	99.09 ± 0.29
Hallmark	4266	99.57 ± 0.22	99.57 ± 0.21	99.57 ± 0.23	0.29 ± 0.14	99.27 ± 0.36
VCF 3000 Effective Genes	6752	99.57 ± 0.47	99.57 ± 0.47	99.64 ± 0.37	0.31 ± 0.34	99.27 ± 0.79
VCF 3500 Effective Genes	7741	99.68 ± 0.40	99.68 ± 0.39	99.72 ± 0.30	0.24 ± 0.26	99.46 ± 0.67

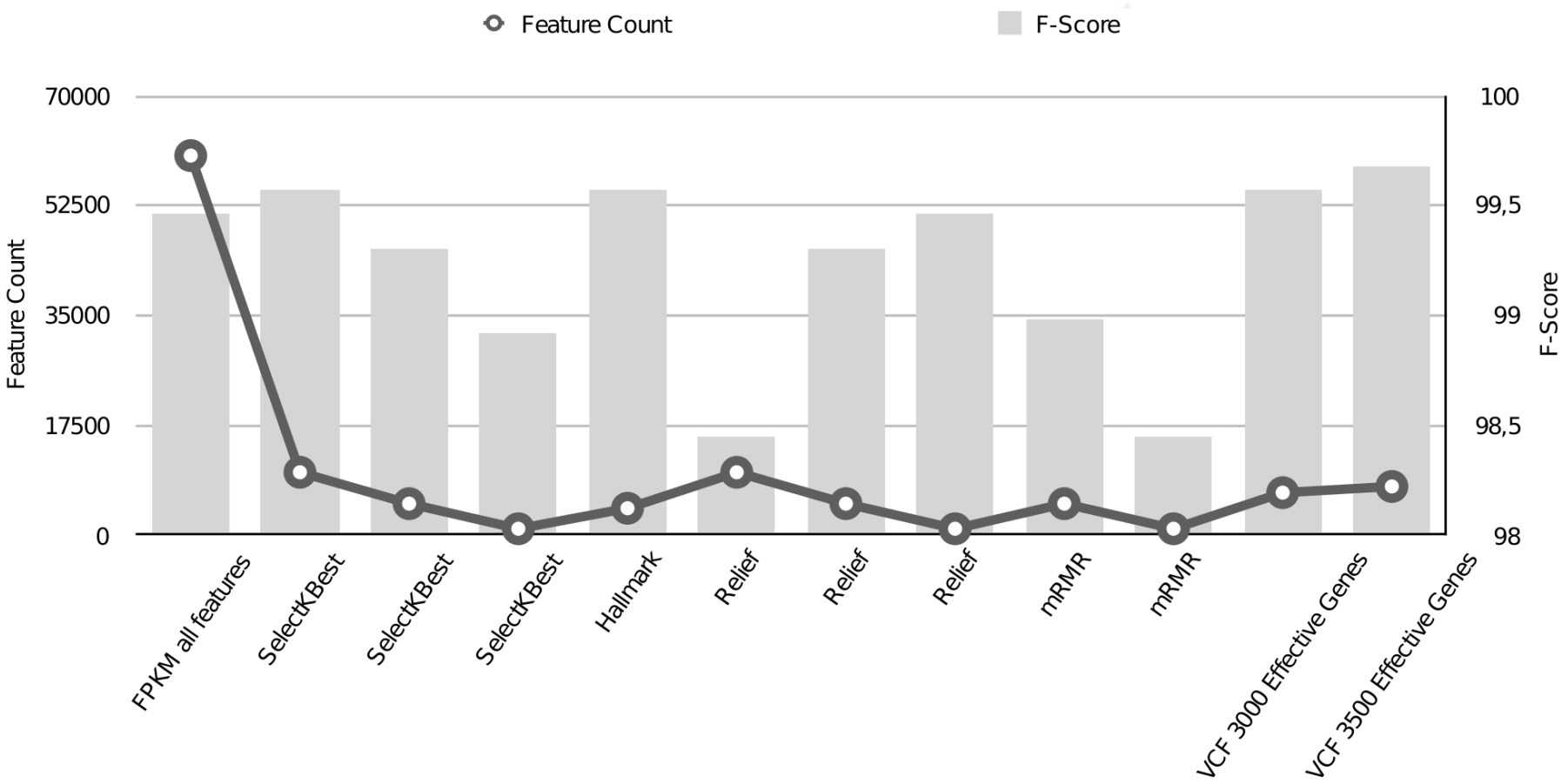


Figure 4.2. F-score and feature count comparison of experiment results.

5. GNNMUTATION: A HETEROGENEOUS GRAPH-BASED FRAMEWORK FOR CANCER DETECTION BASED ON DNA MUTATIONS

In this chapter, we introduce a novel approach based on a graph neural networks that jointly considers genetic mutations and protein interactions for cancer prediction. We use DNA mutations in whole exome sequencing data and create a heterogeneous graph in which patients and proteins are represented as nodes and protein-protein interactions as edges. Furthermore, the patient nodes are connected to the protein nodes based on the mutations in the patient's DNA. With our new heterogeneous graph structure, the patient nodes are updated by both mutations and protein interactions. Since the effects of each mutation on disease development are different, we processed the input graph with attention-based graph neural networks.

We compiled a dataset from the UKBiobank consisting of patients with a cancer diagnosis as a case group and patients without a cancer diagnosis as a control group. We evaluated our approach for breast cancer and prostate cancer, which are among the most common cancers, and showed that the proposed framework effectively discriminates between case and control. The results show that our proposed graph structure and node updating lead to better performance in cancer classification. We have also extended our system with an explainer. This reveals a list of causal genes for recognising a cancer patient. We have found that some of them have already been investigated in cancer studies. This confirms that our system is able to recognise real causal genes for the selected cancers and make predictions based on them. Details of this study are given in the following sections.

5.1. Introduction

Cancer is one of the leading causes of death worldwide [6]. Genetic changes can lead to the development of cancer if they alter cell growth and spread. Most cancer-

causing DNA changes occur in the genes [6]. The mutations in the genes affect the proteins, as they carry the instructions for the production of proteins. Impaired protein function can lead to changes in protein-protein interactions and can cause the formation of cancer cells. In this study, we investigate the effects of DNA mutations and protein interactions on cancer development. We selected breast cancer and prostate cancer as our case study. Breast cancer is the most common type of cancer in women and prostate cancer in men [40].

Biological interactions can be modeled as graph structure in computerized systems. Protein-protein interaction (PPI) networks are utilized as the base graph structure for most genetic disease related studies [78]. In [79], a graph-based framework for novel cancer gene discovery is presented by using graph neural networks (GNNs) for node level classification. They labeled genes of a PPI network according to their cancer relevance (either cancer-related or not). Then they applied graph convolution networks (GCN) to predict whether unlabeled genes are related to cancer or not. In [80], they also use PPI as graph input but applied graph level classification to detect disease subnetworks. They constructed a PPI network for each patient and classify using a graph isomorphism network (GIN) architecture. GNN is used for the identification of cancer driver genes [81], and for the analysis of cancer subtypes [82]. Attention mechanisms are also applied for analyzing cancer driver genes, predicting the disease and classifying disease subtypes [83–85]. Since GNN is a complex method, explainers are applied to reveal the most effective features for network prediction [79, 80, 86].

In the studies mentioned above, homogeneous graphs are used for data modeling. In real life, different types of objects interact with each other. Heterogeneous graphs can model these variations in both nodes (objects) and edges (interaction types). Heterogeneous graph structures are also utilized for data modeling with GNNs [87–92].

In this study, we propose an attention-based graph neural network, called GNNMutation, to recognize cancer patients. We used the gene mutation information in UKBiobank [16] for a corpus of patients with and without cancer diagnosis. GNN-

Mutation operates in a heterogeneous environment that merges patient mutation data with protein-protein interaction data. To the best of our knowledge, this heterogeneous graph structure combining patients and proteins is being tried for the first time. Previous studies with graph input map patient features to proteins or other node types and analyze the interaction of biological objects. Here, we define the patient as an independent object and observe the result of mutation updates. We define edges from proteins to patients based on the corresponding gene mutations. These edges allow the protein nodes to update the patient nodes to achieve broader information transfer and better classification performance. GNNMutation precisely differentiates between case and control groups. The explanatory part of GNNMutation specifies the causal genes that lead to the development of the disease.

The key contributions of our work are summarized as follows:

(i) A novel heterogeneous graph structure has been proposed for the representation of patients and proteins for disease classification.
(ii) The protein interaction information and mutation information have been transferred into the decision mechanism by defining corresponding edges from protein nodes to patient nodes.
(iii) The proposed framework has achieved a leap in classification accuracy.
(iv) The parameters of the classification model have been examined to identify a list of causal genes that reflect the effectiveness of the proposed heterogeneous graph modeling approach.

5.2. Materials and Methods

In this study, we propose a heterogeneous graph-based framework for cancer detection. The input of the system is a novel heterogeneous graph that models patients and proteins. This graph represents the protein-protein interactions and DNA mutations by defining edges between the corresponding patient and protein nodes. A Graph

Attention Network (GAT) classifier is applied on this graph to decide whether an unlabeled patient has cancer or not. An explainer module in the framework examines the GAT model parameters and suggests a list of causal genes for each cancer type. Figure 5.1 shows an overview of the proposed framework.

5.2.1. Dataset

We utilized the variant calling files [5] of whole exome sequencing data from UKBiobank [16]. The sequencing data is obtained from blood samples. Therefore, no biopsy or tumor sample is required for our framework. From the data to which we have access rights, we randomly selected 1000 patients diagnosed with the selected cancers. For control group, we selected a random set of same number of patients without a cancer diagnosis. For our gene set, we created a union of the mutated genes in UKBiobank, the genes that produce proteins in the physical protein-protein interaction (PPI) network of StringDB [93] and the genes in the Hallmark gene sets [94]. Based on this list of genes, we selected a subset of the physical PPI network of StringDB.

5.2.2. GNNMutation

5.2.2.1. Heterogeneous Graph. We formulate our problem as a node classification problem on a global graph for patients and proteins. Inspired from [95], we defined our heterogeneous graph model. Instead of documents and words, we used patients and proteins. To our knowledge, this is the first study that constructs a heterogeneous graph with patient, mutation and protein information and defines patients as separate nodes. We used proteins as nodes and defined edges between proteins from the StringDB physical PPI network. The protein-protein edges in the graph were defined as undirected. We also used patients as nodes and defined edges from proteins to patients if there is a mutation in the gene that produces the protein for that patient. The protein-patient edges in the graph were defined as directed. In this way, the patient nodes are updated by mutation information and indirectly by protein interactions. Based on these updates, the network predicts the class label. An example input

graph can be found in Figure 5.2. In the figure, PPI network is in the center. The patient nodes are scattered over the PPI. There are multiple edges from the proteins in the center to each patient corresponding to the DNA mutations. The labels of the test patients are masked in runs (they are shown as gray nodes). We assigned the protein node features as one-hot vectors. The patient node features were initialized as statistical mutation vectors of the selected gene set.

5.2.2.2. Patient Feature Vectors. We calculated statistical features for patient nodes based on our previous study [8]. We constructed a feature vector for each patient. This vector summarizes the gene mutation information in the VCF files. It is a one dimensional vector with length of total number of genes in our system. Each element in the feature vector provides the mutation gene score for that gene and for that patient. We used the feature type that performed best in the previous study, namely bm25-tf-rf. Calculation details can be found in our previous study [8].

5.2.2.3. GNN Model. After deciding on the heterogeneous graph structure and features, we need to select a graph network. There are different GNN types and implementations. The most commonly used graph network is GCN. But the GCN implementation of the PyG [96] library does not support graphs with edges connecting different node types. Also, we want to distinguish between gene updates for a patient, since not every gene mutation affects disease development equally. We have PPI scores and patient mutation vector features that we can use as edge weights. But these are static weights. Instead, we want the network to learn the weights during training. Therefore, we chose GAT [18] as our graph network.

5.2.2.4. Explainer Model. Classification algorithms used to be black box methods. Nowadays, interpretation libraries allow us to understand how these complex methods make decisions. In this study, we applied Captum [97] explainer to find the roots of the predictions of our heterogeneous graph based model. Our goal is to evaluate and compare the impact of gene mutations on the final class decisions for patients. We com-

puted integrated gradients (IG) [98] on protein-patient edges for cancer patients. The edges with higher attribution values indicate that the gene that produces the protein in the source of the edge has a greater impact on disease detection and development. To obtain attribution scores, we first transformed our heterogeneous model into a Captum heterogeneous model. Second, we converted our inputs from our heterogeneous graph into Captum inputs. Then, we computed the IG attribution algorithm for all protein-patient edges in our Captum inputs using our Captum heterogeneous model for cancer samples. We formulate our deep model as a function F : $R^n \rightarrow [0, 1]$. We have our feature vectors and a set of nodes and edges and our model aims to perform a binary classification. Let $x \in R^n$ be the input at hand, and $x' \in R^n$ be the baseline input. The integrated gradient along the i^{th} dimension for an input x and baseline x' is defined as

$$IG_i(x) ::= (x_i - x_i{}') * \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha. \tag{5.1}$$

5.2.3. Implementation and Experiment Design

GNNMutation model is implemented in Python with the PyG library [96]. The best performing model for our heterogeneous graph is GAT with 2 levels. We tuned the network parameters with several tests. The reported results are with the best parameters, namely 128 for the number of hidden units, 8 for the number of GAT heads and $0,001$ for the learning rate. We performed our tests with transductive learning. All tests were performed with 3 repeats of 10-fold cross-validation by keeping the rate of case-control samples equal in each fold. The average accuracy with standard deviation is used as a performance measure for comparisons. Best parameters were selected according to the validation set accuracy. The reported accuracies were calculated on test set.

To determine the effect of graph data model in classification performance, we selected algorithms without graph input for comparison. We applied logistic regression

(LR) and multilayer perceptron (MLP) with 4 layers. We also applied Model-Half from our previous study [8]. To our knowledge, for VCF data and cancer classification, Model-Half provides the state-of-the-art results. It calculates a statistical feature vector for each patient and applies a neural-network model. For the graph-based comparison, we tested a homogeneous graph classification with our data and called it as HomogeneousGNN. It creates a PPI graph for each patient, assigns patient-related features to the protein nodes in the graph and applies graph-based classification.

5.3. Results and Discussion

5.3.1. Comparison of Different Numbers of Protein-Patient Edges

We defined our protein-patient edges based on the genes that have maximum feature values for each patient. We connected each patient to the k proteins that have the highest mutation scores in the patient's feature vector. For the initial tests, we selected k as 100 genes for each patient. We repeated our tests with different number of genes. The classification test results with different number of protein-patient edges per patient are shown in Table 5.1. The validation set accuracy values are parallel to the test set results. For breast cancer, increasing the number of genes first leads to improved accuracy. However, when the number of edges is beyond a threshold, in our case 300 genes, the accuracy value starts to decrease. This may be due to the fact that, in graph networks nodes may have similar values after certain runs when they are updated by similar edges. This similarity makes it difficult to differentiate between nodes and degrades performance. The highest accuracy for both breast and prostate cancer is achieved by selecting k as 300 genes.

5.3.2. Comparison of Different Protein-Patient Edge Strategies

The input for GNNMutation is a heterogeneous graph of proteins and patients. There are two types of edges in the input graph. The protein-protein edges were defined according to the PPI network. The protein-patient edges were defined based

on the mutations in the genome of the patient. Each patient node is connected to the 300 proteins for which the corresponding genes of the patient have the highest (i.e., maximum) mutation scores. In order to validate our approach and demonstrate the impact of mutation information, we also created baseline where each patient node is connected to randomly chosen 300 protein nodes.

The classification test results for these two protein-patient edge strategies are shown in Table 5.2. The validation set accuracy values are parallel to test set results. From this table, it can be seen that the maximum selection leads to higher accuracy for both cancer types. The random selection yields poor results, demonstrating that incorporating PPI information with mutation information is essential and leads to a powerful approach for cancer detection.

5.3.3. Comparison of Algorithms

Classification test results are presented in Table 5.3. We tested our data with 3 algorithms, LR, MLP and Model-Half, which don't use a graph structure. LR leads to better accuracy compared to MLP. For breast cancer, LR ncreases the accuracy by 2 points and for prostate cancer 3 points compared to MLP results. Model-Half is more accurate compared to LR and MLP, which is consistent with our previous study [8]. When GNNMutation is considered, we find that the classification accuracy is significantly better all 3 algorithms. For breast cancer, GNNMutation achieves an accuracy of 90.3% and for prostate cancer 86.6%. This accuracy improvement is the result of the graph input model. By using the graph representation, the additional biological system information is appended to the features. And by using a GNN model, the messaging between nodes enable the information transfer from proteins to patients. By modeling the GNN in multiple layers, not only mutation information but also the protein interactions are used to update patient nodes.

We consider HomogeneousGNN as another algorithm with graph input. According to our tests, the average accuracy of HomogeneousGNN is listed as the least

performing method for both cancer types. Based on the results, GNNMutation clearly outperforms HomogeneousGNN, for this type of mutation data. The better results with GNNMutation shows that the heterogeneous representation with the addition of mutation related edges besides protein interactions is more accurate for this problem.

If we look at Table 5.1 and Table 5.3 together, we can also see that even the lowest accuracy of the proposed system, GNNMutation, is better than all other methods tested in this study. If we look at the highest score of GNNMutation, it clearly outperforms the other methods and produces a significant leap in performance.

5.3.4. Discussion of Explainer Results

The input of the GNNMutation framework is a heterogeneous graph containing both patient and protein information. With the undirected protein-protein edges, the protein nodes update each other. With the directed protein-patient edges, the protein nodes update the patient nodes with mutation information and their interaction information. Graph-based frameworks make predictions that take these updates into account as well as the feature values. The explainer for the graph model extracts the rationale behind the framework's decision mechanisms. We let the explainer evaluate the protein-patient nodes to see the effects of mutations and protein interactions on the predictions. We then select 20 genes with the highest scores for each cancer type. For visualization, we normalized the gene scores of the top 20 genes between 0 and 1.

The most effective genes and their normalized scores for breast cancer are shown in Figure 5.3. [99] suggests that CASP9 transcriptional regulation is an important factor in the development of breast cancer. [100] proposes CASP9 as a biomarker and a therapeutic target in inflammatory breast cancer. [101] investigates the role of IQGAP2 in tumor angiogenesis in breast cancer. [102] investigates the role of IQGAP2 as a tumor suppressor in breast cancer. [103] examines breast tissue microarrays and shows overexpression of MADD. [104] shows that the PFKP protein is highly expressed in triple-negative breast cancer (TNBC). [105] investigates that silencing NOTCH4 promotes

tumorigenesis and inhibits metastasis in TNBC. [106] suggests that MED13L positively correlates with survival of patients with breast cancer. [107] finds that NEDD4 expression is a predictive factor for the response to hormone therapy in breast cancer patients. [108] indicates upregulated MXRA5 expression in breast cancer tissue.

The most effective genes and their normalized scores for prostate cancer are shown in Figure 5.4. ARHGEF3 is shown to be an oncogene and is proposed as a novel biomarker for the prediction of invasive prostate cancer in [109]. [110] associates high and intermediate levels of CCDC88A with higher prostate cancer survival. [111] finds that overexpression of DOCK4 is observed in primary prostate tumors in the TCGA dataset. [112] finds that CR-1 mRNA and protein are upregulated in prostate cancer and suggests that CR-1 expression may be a novel biological target for personalized therapy. [113] identifies RXRA as a novel target in prostate cancer. PDLIM5 is reported to be abnormally enriched in prostate cancer tissues in [114]. According to [115], studies using various systems biology methods have highlighted NR3C1 as a potential gene in prostate cancer metastasis. SATB1 is shown to be overexpressed in metastatic prostate cancer in [116]. [117] finds that serum concentration of TYRO3 is significantly increased in men with localized or metastatic prostate cancer compared to men without prostate cancer.

We searched the literature if there exist research for top 20 of these proposed genes related with the corresponding cancer type by our explainer. We found that nearly half of them took part in previous studies. This shows that our framework emphasizes on proper genes for the selected cancer types. And also the rest of the genes that are proposed by the explainer can be utilized as noval target genes for new cancer studies.

5.4. Final Remarks

DNA mutations can affect the structure and function of the corresponding proteins. These effects can change the behavior of a cell from normal to cancerous [6]. In

this study, we propose an explainable graph-based framework to recognize the genomic roots of cancer. We selected breast and prostate cancer for our tests. For this purpose, we utilized gene mutations from whole exome sequencing data. We modeled our input graph to contain separate patient and protein nodes. With this heterogeneous graph structure, we can model the effects of gene mutations and protein-protein interactions on patient decisions by defining corresponding edges. We compared our framework, GNNMutation, with both graph-based and non-graph based algorithms. Our test results show that the updates in the heterogeneous graph structure lead to a significantly better discrimination between case and control patients. Furthermore, we employ an explainer module that examines GNNMutation and evaluates the genes according to their influence on cancer decision. Our literature search on these genes shows that many of them have been studied and reported in association with the corresponding cancer types. Therefore, we suggest that the remaining genes in the explainer output list are candidate genes to be studies further for the selected cancer types. GNNMutation is trained and run on two cancers but can be applied to any genomic disease.

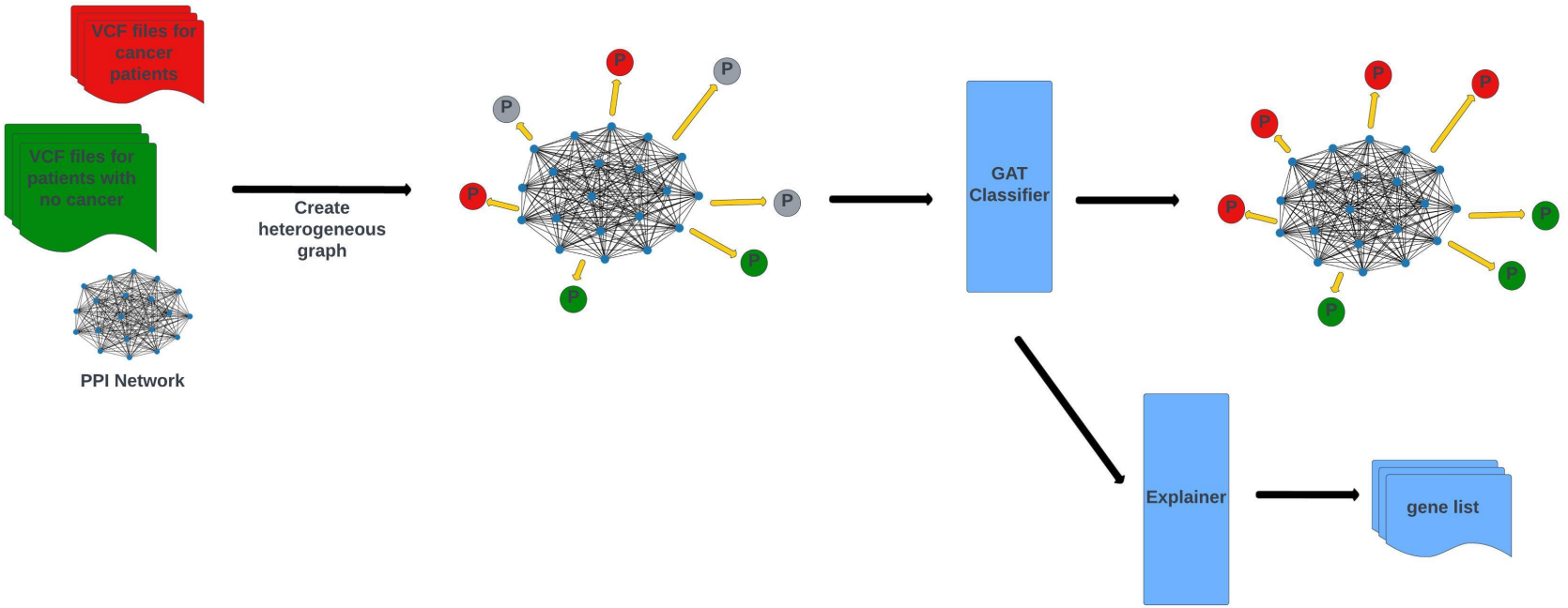


Figure 5.1. Overview of GNNMutation: The input of the framework is a heterogeneous graph modeling DNA mutations and protein interactions. GAT classifier is applied to decide whether the patient has cancer or not. The explainer module examines classifier parameters and outputs a list of causal genes.

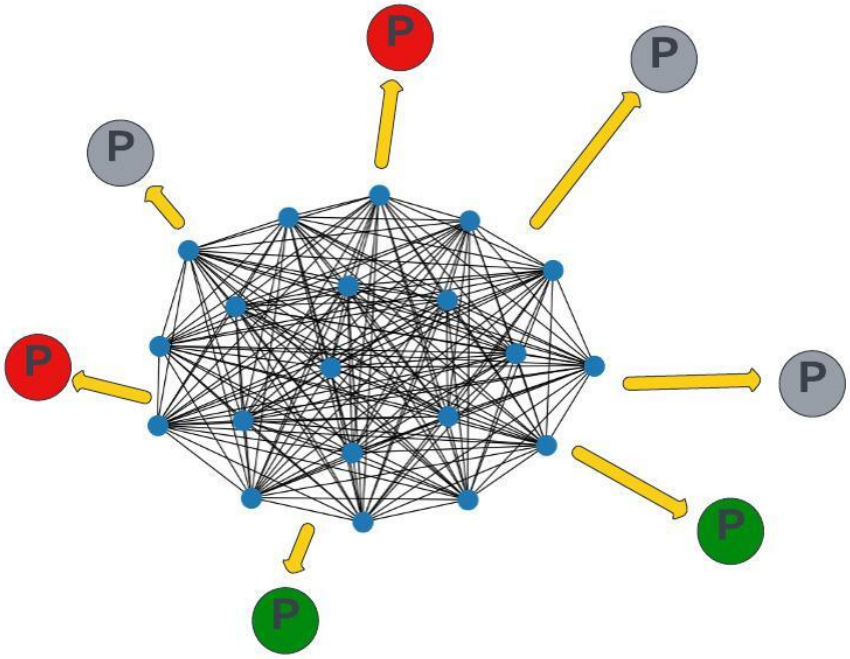


Figure 5.2. Heterogeneous graph of patients and genes: Small blue nodes are the gene nodes and make PPI network among themselves. Big nodes with P letter are the patient nodes. Red represents cancer patients, green represents no cancer patients and grey represents test nodes to be predicted.

Table 5.1. Classification test accuracy results with different number of protein-patient edges per patient.

Edge	**Breast**	**Prostate**
max 100	0.740 (0.008)	0.847 (0.003)
max 200	0.824 (0.006)	0.800 (0.025)
max 300	**0.903 (0.003)**	**0.866 (0.036)**
max 400	0.873 (0.019)	0.841 (0.018)

Table 5.2. Classification test accuracy results with different protein-patient edge strategies. In the first strategy, each patient node is connected to the proteins for which the mutation scores in the patient feature vector are maximum. In the second strategy, each patient node is connected to random protein nodes. Thus, the first strategy incorporates mutation information, whereas the second strategy doesn't.

Edge	**Breast**	**Prostate**
max 300	**0.903 (0.003)**	**0.866 (0.036)**
random 300	0.495 (0.003)	0.495 (0.005)

Table 5.3. Classification test accuracy results.

Model	**Cancer Type**	**Accuracy**
LR	Breast	0.711 (0.000)
	Prostate	0.770 (0.000)
MLP	Breast	0.606 (0.030)
	Prostate	0.735 (0.013)
Model-Half	Breast	0.728 (0.007)
	Prostate	0.791 (0.011)
HomogeneousGNN	Breast	0.499 (0.030)
	Prostate	0.503 (0.022)
GNNMutation	Breast	**0.903 (0.003)**
	Prostate	**0.866 (0.036)**

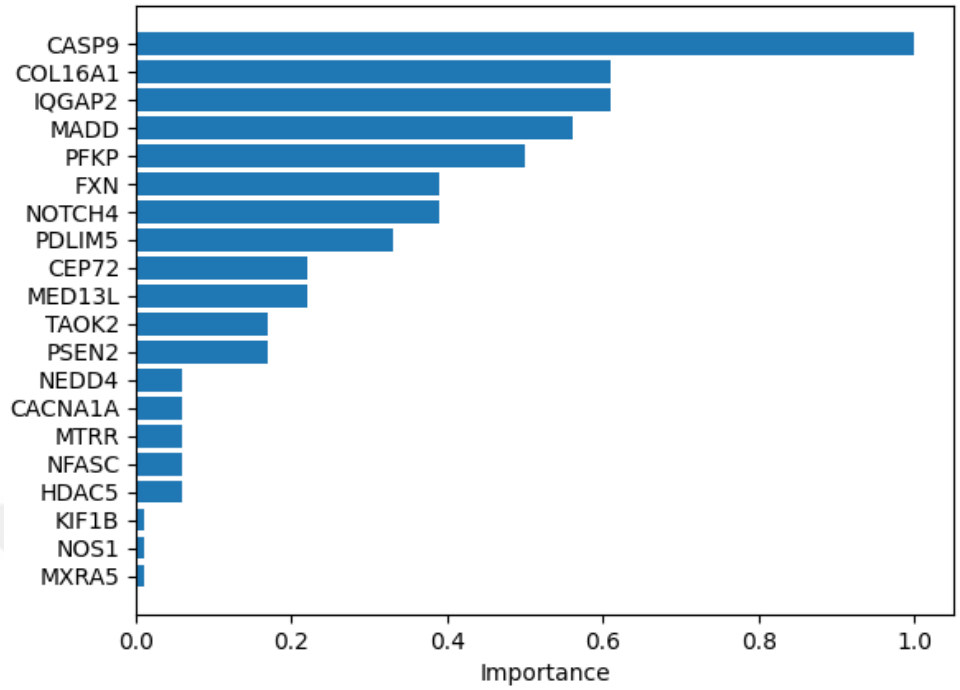


Figure 5.3. Explainer results for breast cancer.

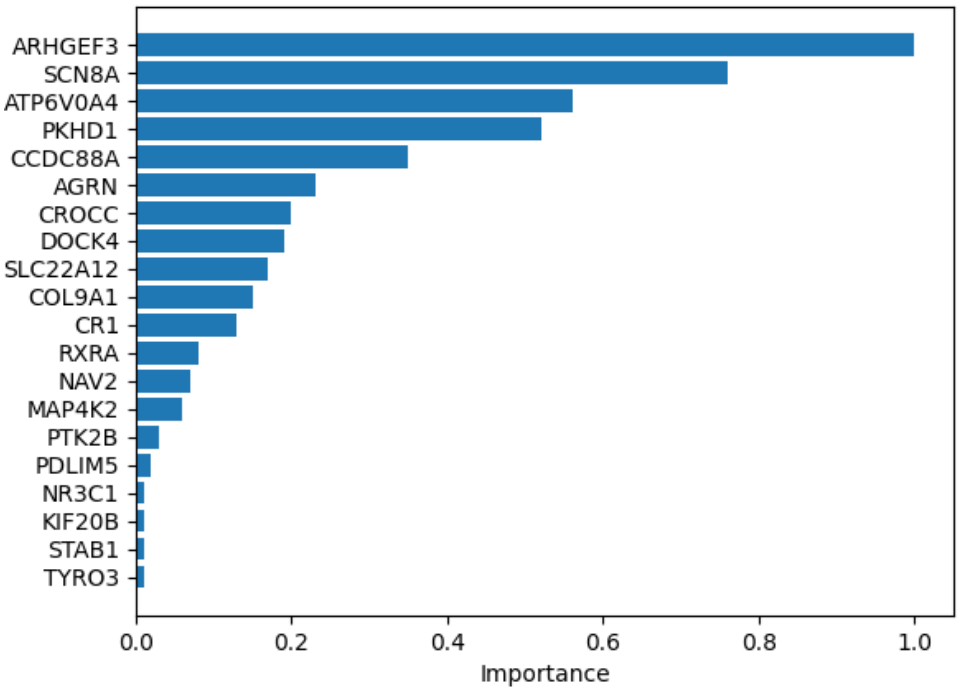


Figure 5.4. Explainer results for prostate cancer.

6. CONCLUSION

6.1. Discussion

In this thesis, we focused on two research problems. The first one is to predict whether or not a patient has a certain disease based on DNA mutation information. The second one is to identify disease causing genes or target genes for genomic testing or drug development. We proposed frameworks for solving these problems and tested them on cancer specific data. If a different dataset for a different genomic disease is utilized, these systems can easily be trained and adapted for that disease.

We proposed three different frameworks for the problem of disease prediction. Chapter 3 presents our initial study. In this study, we define the VCF file as a document and the DNA mutations as words in that document. We then computed term weighting metrics in this context and used them as our features. We computed several such metrics and compared them for the task of cancer type classification. In addition, we proposed an explainable neural network as a classification algorithm. Our test results show that the proposed method achieves significantly better performance compared to classical classification algorithms. Chapter 4 presents our second study. In this study, we transformed our knowledge of VCF data into a novel feature selection method for gene expression data. The expression levels of the resulting genes are used to classify cancer types. Our test results show that the proposed gene selection method leads to equal or better classification performance compared to classical feature selection algorithms. Chapter 5 presents our third study. In this study, we modelled the VCF data as a graph of patients and genes/proteins. Our novel heterogeneous data representation model with attention based GNN leads to a leap in classification performance compared to other classical or graph based methods. In these studies, we utilized VCF data and also ensembled the mutation information with gene expression and protein interaction data. Therefore, we believe that we have performed a detailed and versatile analysis of disease related DNA mutations using other genomic data.

We examined two of the proposed methods on the detection of a disease for causal or target gene prediction. In Chapter 3, we applied an explainable neural network to VCF data. After the network was trained for classification, we examined the network parameters and extracted the importance degree of each gene for predicting the type of cancer. In Chapter 5, we applied an attention based GNN, which is also explainable. After completing the training phase, we analysed the network parameters for the importance of the genes in distinguishing patients with or without cancer diagnosis. For the two explainer studies, we examined the proposed list of the most effective genes in cancer prediction in the cancer literature. We found that almost half of the resulting genes have been previously studied as causal or target genes in published cancer studies. This finding leads us to conclude that the proposed frameworks make their predictions on true causal or target genes. Moreover, the remaining proposed genes can be studied as novel genes in cancer diagnosis or treatments.

In this thesis, we tackled disease classification and disease-gene association problems. We encountered data, memory and computational limitations and defined and chose ways to overcome them. For example, we wanted to work with raw DNA sequencing data but it was huge in size for both downloading and storing. We chose to work with VCF data as a representation of the altering parts of the DNA sequence to overcome this handicap.

Every thesis is a step forward to the solution of the targeted problems. In this thesis, we explored genomic roots of diseases using mutation information and proposed representation and classification methods. Our studies are evaluated as original and published in highly rated journals. We know that the problems are not completely solved yet, but we strongly believe and are confident that we accomplished a step forward for these problems.

6.2. Future Work

Our main future direction for the research studies reported in this thesis is to analyse the raw sequencing data for these problems. But as mentioned in the challenges section, the raw sequencing data size is very large. As more studies are conducted to understand the DNA as a language, more models can be defined to represent the raw sequencing data. This will give us the opportunity to represent the raw data in a much smaller way and utilize for more research problems.

Before that stage, some additional mutation information can be used besides our input vectors to improve the prediction of diseases. The VCF file contains an INFO field for each mutation. This field can provide additional statistical or categorical information about the mutation when selected in the annotation phase. Some examples of additional information include mutation frequency, various pathogenicity scoring values and the type of mutation. Each of these fields represents information provided based on previous research studies on these mutations. We believe that if we can combine all these values in the representation of a gene for a patient, we can improve the classification performance.

In addition, multi-omics data can be used to better represent the genomic or biological condition of a patient. This can improve both the classification performance and our understanding of the disease development processes.

REFERENCES

1. National Human Genome Research Institute, `https://www.genome.gov/`, accessed on March 01, 2024.

2. The Human Genome Project, `https://www.genome.gov/human-genome-project`, accessed on January 01, 2024.

3. The Genome Reference Consortium, `https://www.ncbi.nlm.nih.gov/grc`, accessed on March 01, 2018.

4. World Health Organisation, `https://www.who.int/`, accessed on February 01, 2024.

5. "The Variant Call Format Version 4.2 Specification", Global Alliance for Genomics and Health, Toronto, Canada, 2022, `https://samtools.github.io/hts-specs/VCFv4.2.pdf`, accessed on February 01, 2024.

6. The National Cancer Institute, `https://www.cancer.gov/`, accessed on January 01, 2024.

7. Qiagen, `https://www.qiagen.com/us`, accessed on February 01, 2024.

8. Özcan Şimşek, N., A. Özgür and F. Gürgen, "Statistical Representation Models for Mutation Information within Genomic Data", *BMC Bioinformatics*, Vol. 20, No. 324, pp. 1–13, 2019.

9. Özcan Şimşek, N., A. Özgür and F. Gürgen, "A Novel Gene Selection Method for Gene Expression Data for the Task of Cancer Type Classification", *Biology Direct*, Vol. 16, No. 7, pp. 1–5, 2021.

10. Piovesan, A., M. C. Pelleri, F. Antonaros, P. Strippoli, M. Caracausi and L. Vitale, "On the Length, Weight and GC Content of the Human Genome", *BMC Research Notes*, Vol. 12, No. 106, pp. 1–7, 2019.

11. The Cancer Genome Atlas, `https://www.cancergenome.nih.gov/`, accessed on March 01, 2024.

12. National Institutes of Health, `https://www.nih.gov/`, accessed on January 01, 2024.

13. Weinstein, J. N., E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander and J. M. Stuart, "The Cancer Genome Atlas Pan-Cancer Analysis Project", *Nature Genetics*, Vol. 45, No. 1, pp. 1113–1120, 2013.

14. Hoadley, K. A., C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A. M. Taylor, A. D. Cherniack, V. Thorsson, R. Akbani, R. Bowlby, C. K. Wong, M. Wiznerowicz, F. Sánchez-Vega, A. G. Robertson, B. G. Schneider, M. S. Lawrence, H. Noushmehr, T. M. Malta, J. M. Stuart, C. C. Benz and P. W. Laird, "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer", *Cell*, Vol. 173, No. 1, pp. 291–304, 2018.

15. Genomic Data Commons Data Portal, `https://portal.gdc.cancer.gov/`, accessed on January 01, 2024.

16. Uk Biobank, `https://www.ukbiobank.ac.uk/`, accessed on February 01, 2024.

17. Wellcome Trust, `https://wellcome.org/`, accessed on March 01, 2024.

18. Velickovic, P., G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, "Graph Attention Networks", ArXiv:1710.10903, 2018.

19. "Health, United States, 2017: with Special Feature on Mor-

tality", National Center for Health Statistics, Hyattsville, 2018, https://www.ncbi.nlm.nih.gov/books/NBK532685/, accessed on January 01, 2024.

20. Zacharaki, E., S. Wang, S. Chawla, D. Yoo, R. Wolf, E. Melhem and C. Davatzikos, "Classification of Brain Tumor Type and Grade using MRI Texture and Shape in a Machine Learning Scheme", *Magnetic Resonance in Medicine*, Vol. 62, No. 1, pp. 1609–1618, 2009.

21. Joshi, D., D. Rana and V. Misra, "Classification of Brain Cancer using Artificial Neural Network", *International Conference on Electronic Computer Technology*, Kuala Lumpur, Malaysia, 2010.

22. Zulpe, N. and V. Pawar, "GLCM Textural Features for Brain Tumor Classification", *International Journal of Computer Science*, Vol. 9, No. 3, pp. 354–359, 2012.

23. Patil, S. and M. Kuchanur, "Lung Cancer Classification using Image Processing", *International Journal of Engineering and Innovative Technology*, Vol. 2, No. 3, pp. 37–42, 2012.

24. Kuruvilla, J. and K. Gunavathi, "Lung Cancer Classification using Neural Networks for CT Images", *Elsevier Computer Methods and Programs in Biomedicine*, Vol. 113, No. 1, pp. 202–209, 2014.

25. Xu, Y., J.-Y. Zhu, E. Chang and Z. Tu, "Multiple Clustered Instance Learning for Histopathology Cancer Image Classification, Segmentation and Clustering", *Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition*, Rhode Island, America, 2012.

26. Wang, H., F. Xing, H. Su, A. Stromberg and L. Yang, "Novel Image Markers for Non-Small Cell Lung Cancer Classification and Survival Prediction", *BMC*

Bioinformatics, Vol. 15, No. 310, pp. 1–12, 2014.

27. Nguyen, D. and D. Rocke, "Multi-Class Cancer Classification via Partial Least Squares with Gene Expression Profiles", *Bioinformatics*, Vol. 18, No. 1, p. 1216, 2002.

28. Tan, A. C. and D. Gilbert, "Ensemble Machine Learning on Gene Expression Data for Cancer Classification", *Applied Bioinformatics*, Vol. 2, No. 1, pp. 75–83, 2003.

29. Statnikov, A., L. Wang and C. Aliferis, "A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray-Based Cancer Classification", *BMC Bioinformatics*, Vol. 9, No. 319, pp. 1–10, 2008.

30. Liu, J., X. Wang, Y. Cheng and L. Zhang, "Tumor Gene Expression Data Classification via Sample Expansion-Based Deep Learning", *Oncotarget*, Vol. 8, No. 3, pp. 109646–109660, 2017.

31. Xiaoa, Y., J. Wub, Z. Linc and X. Zhao, "A Deep Learning-Based Multi-Model Ensemble Method for Cancer Prediction", *Elsevier Computer Methods and Programs in Biomedicine*, Vol. 153, No. 3, pp. 1–9, 2018.

32. Draghici, S., P. Khatri, A. C. Eklund and Z. Szallasi, "Reliability and Reproducibility Issues in DNA Microarray Measurements", *Trends in Genetics*, Vol. 22, No. 2, pp. 101–109, 2006.

33. Wang, Z., "Multi-Class HingeBoost Method and Application to the Classification of Cancer Types using Gene Expression Data", *Methods of Information in Medicine*, Vol. 51, No. 2, pp. 162–167, 2012.

34. Gao, L., M. Ye, X. Lu and D. Huang, "Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification", *Elsevier Genomics Proteomics Bioinformatics*, Vol. 15, No. 2, pp. 389–395, 2017.

35. Amar, D., S. Izraeli and R. Shamir, "Utilizing Somatic Mutation Data from Numerous Studies for Cancer Research: Proof of Concept and Applications", *Oncogene*, Vol. 36, No. 2, pp. 3375–3383, 2017.

36. He, Z., J. Zhang, X. Yuan, Z. Liu, B. Liu, S. Tuo and Y. Liu, "Network Based Stratification of Major Cancers by Integrating Somatic Mutation and Gene Expression Data", *Plos One*, Vol. 12, No. 2, pp. 1–12, 2017.

37. Kircher, M., D. M. Witten, B. J. O. P. Jain, G. M. Cooper and J. Shendure, "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants", *Nature Genetics*, Vol. 46, No. 2, pp. 310–315, 2015.

38. Rentzsch, P., D. Witten, G. M. Cooper, J. Shendure and M. Kircher, "CADD: Predicting the Deleteriousness of Variants Throughout the Human Genome", *Nucleic Acids Research*, Vol. 47, No. 3, pp. 886–894, 2018.

39. Vural, S., X. Wang and C. Guda, "Classification of Breast Cancer Patients using Somatic Mutation Profiles and Machine Learning Approaches", *BMC Systems Biology*, Vol. 10, No. 62, pp. 1–14, 2016.

40. American Cancer Society, `https://www.cancer.org/`, accessed on March 01, 2024.

41. Wang, K., M. Li, and H. Hakonarson, "Annovar: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data", *Nucleic Acids Research*, Vol. 38, No. 16, pp. 164–171, 2010.

42. The Online Mendelian Inheritance in Man, `https://omim.org/`, accessed on February 01, 2018.

43. The Multi Symbol Checker, `https://www.genenames.org/cgi-bin/symbol_checker`, accessed on March 01, 2018.

44. The Hugo Gene Nomenclature Committee, `https://genenames.org/`, accessed on January 01, 2018.

45. Jones, K., "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", *Journal of Documentation*, Vol. 28, No. 3, pp. 11–21, 1972.

46. Lan, M., C. L. Tan, J. Su and Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization", *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 3, pp. 721–735, 2009.

47. Jones, K. S., S. Walker and S. E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments - Part 1", *Information Processing and Management*, Vol. 36, No. 4, pp. 779–808, 2000.

48. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-Learn: Machine Learning in Python", *Journal of Machine Learning Research*, Vol. 12, No. 4, pp. 2825–2830, 2011.

49. Chollet, F., "Keras", 2015, `https://github.com/keras-team/keras`, accessed on February 01, 2015.

50. Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zhang, "TensorFlow: A System for Large-Scale Machine Learning", *USENIX Symposium on Operating Systems Design and Implementation*, Carlsbad, California, America, 2016.

51. Burke, L. J., J. evík, G. Gambino, E. Tudini, E. J. Mucaki, B. C. Shirley, P. J.

Whiley, M. T. Parsons, K. D. Leeneer, S. Gutiérrez-Enríquez, M. Santamariña, S. M. Caputo, E. S. dos Santos, J. Soukupová, M. Janatová, P. Zemánková, K. Lhotová, L. Stolarova, M. Borecka, A. Moles-Fernández, S. Manoukian, B. Bonanni, S. L. Edwards, M. J. Blok, T. van Overeem Hansen, M. Rossing, O. Díez, A. Vega, K. B. M. Claes, D. E. Goldgar, E. Rouleau, P. Radice, P. Peterlongo, P. K. Rogan, M. A. Caligo, A. B. Spurdle and M. A. Brown, "BRCA1 and BRCA2 5 Noncoding Region Variants Identified in Breast Cancer Patients Alter Promoter Activity and Protein Binding", *Human Mutation*, Vol. 39, No. 5, pp. 2025–2039, 2018.

52. dos Santos, E. S., F. Lallemand, L. J. Burke, D. Stoppa-Lyonnet, M. A. Brown, S. M. Caputo and E. Rouleau, "Non-Coding Variants in BRCA1 and BRCA2 Genes: Potential Impact on Breast and Ovarian Cancer Predisposition", *Cancers*, Vol. 10, No. 453, pp. 1–21, 2018.

53. Hockla, A., D. C. Radisky and E. S. Radisky, "Mesotrypsin Promotes Malignant Growth of Breast Cancer Cells Through Shedding of CD109", *Breast Cancer Research and Treatment*, Vol. 124, No. 4, pp. 27–38, 2009.

54. Satooka, H. and M. Hara-chikuma, "Aquaporin-3 Controls Breast Cancer Cell Migration by Regulating Hydrogen Peroxide Transport and Its Downstream Cell Signaling", *Molecular and Cellular Biology*, Vol. 36, No. 7, pp. 1206–1218, 2016.

55. Santuario-Facio, S. K., S. Cardona-Huerta, Y. X. Perez-Paramo, V. Trevino, F. Hernandez-Cabrera, A. Rojas-Martinez, G. Uscanga-Perales, J. L. Martínez-Rodríguez, L. A. Martínez-Jacobo, G. R. Padilla-Rivas, G. E. Muñoz-Maldonado, J. F. González-Guerrero, J. Valero-Gomez, A. L. Vazquez-Guerrero, H. G. Martínez-Rodríguez, Á. Barboza-Quintana, O. Barboza-Quintana, R. Garza-Guajardo and R. Ortiz-Lopez, "A New Gene Expression Signature for Triple Negative Breast Cancer Using Frozen Fresh Tissue Before Neoadjuvant Chemotherapy", *Molecular Medicine*, Vol. 23, No. 1, pp. 101–111, 2017.

56. Kuo, S.-J., S.-Y. Chien, C. Lin, S.-E. Chan, H.-T. Tsai and D.-R. Chen, "Significant Elevation of CLDN16 and HAPLN3 Gene Expression in Human Breast Cancer", *Oncology Reports*, Vol. 24, No. 3, pp. 759–766, 2010.

57. Rykova, E. Y., T. E. Skvortsova, A. Hoffmann, S. N. Tamkovich, S. N. Tamkovich, A. V. Starikov, O. E. Bryzgunova, V. I. Permjakova, J. M. Warnecke, G. Sczakiel, V. V. Vlassov and P. P. Laktionov, "Breast Cancer Diagnostics Based on Extracellular DNA and RNA Circulating in Blood", *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*, Vol. 2, No. 3, pp. 208–213, 2008.

58. Fang, Z., Y. Yi, G. Shi, S. Li, S. Chen, Y. Lin, Z. Li, Z. He, W. Li and S. Zhong, "Role of Brf1 Interaction with ER, and Significance of Its Overexpression, in Human Breast Cancer", *Molecular Oncology*, Vol. 11, No. 3, pp. 1752 – 1767, 2017.

59. Chai, P., J. Tian, D. Zhao, H. Zhang, J. Cui, K. Ding and B. Liu, "GSE1 Negative Regulation by miR-489-5p Promotes Breast Cancer Cell Proliferation and Invasion", *Biochemical and Biophysical Research Communications*, Vol. 471, No. 1, pp. 123–128, 2016.

60. Nyqvist, J., F. Persson, T. Z. Parris, K. Helou, E. K. Sarenmalm, Z. Einbeigi, A. Borg, P. Karlsson and A. Z. Kovacs, "Metachronous and Synchronous Occurrence of 5 Primary Malignancies in a Female Patient between 1997 and 2013: A Case Report with Germline and Somatic Genetic Analysis", *Case Reports in Oncology*, Vol. 10, No. 1, pp. 1006 – 1012, 2017.

61. Jeong, H.-C., G.-I. Kim, S.-H. Cho, K.-H. Lee, J.-J. Ko, J.-H. Yang and K. H. Chung, "Proteomic Analysis of Human Small Cell Lung Cancer Tissues: Up-Regulation of Coactosin-Like Protein-1", *Journal of Proteome Research*, Vol. 10, No. 1, pp. 269–276, 2011.

62. Guo, S., P. Yang, X. Jiang, X. Li, Y. Wang, X. Zhang, B. Sun, Y. Zhang and

Y. Jia, "Genetic and Epigenetic Silencing of mircoRNA-506-3p Enhances COTL1 Oncogene Expression to Foster Non-Small Lung Cancer Progression", *Oncotarget*, Vol. 8, No. 1, pp. 644 – 657, 2017.

63. Bahl, C. R. H., S. Sharma, N. Singh and D. K. Behera, "Association Study between Genetic Variations in Axin2 Gene and Lung Cancer Risk in North Indian Population: A Multiple Interaction Analysis", *Tumour Biology*, Vol. 39, No. 4, pp. 1–18, 2017.

64. Liu, S., X. Chen, R. Chen, J. Wang, G. Zhu, J. Jiang, H. Wang, S. Duan and J. Huang, "Diagnostic Role of Wnt Pathway Gene Promoter Methylation in Non Small Cell Lung Cancer", *Oncotarget*, Vol. 8, No. 4, pp. 36354 – 36367, 2017.

65. Goodwin, J., M. L. Neugent and J. whan Kim, "Lung Squamous Cell Carcinoma Exhibits a Targetable Glucose Dependency Unique among Non-Small Cell Lung Cancers", *Molecular and Cellular Oncology*, Vol. 4, No. 5, pp. 1364211 – 1364212, 2017.

66. Genetic and Genomic Testing by Cancer Treatment Centers of America, `https://www.cancercenter.com/diagnosing-cancer/genetic-and-genomic-testing`, accessed on February 01, 2020.

67. Alexandrov, L. B., J. Kim, N. J. Haradhvala, M. N. Huang, A. S. Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. López-Bigas, L. J. Klimczak, J. R. Mcpherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen and Stratton, "The Repertoire of Mutational Signatures in Human Cancer", *Nature*, Vol. 578, No. 1, pp. 94 – 101, 2018.

68. Rheinbay, E. *et al.*, "Analyses of Non-Coding Somatic Drivers in 2658 Cancer Whole Genomes", *Nature*, Vol. 578, No. 1, pp. 102 – 111, 2020.

69. Priestley, P., J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, P. Roepman, M. Voda, H. Bloemendal, V. Tjan-Heijnen, C. V. van Herpen, M. Labots, P. Witteveen, E. Smit, S. Sleijfer, E. Voest and E. Cuppen, "Pan-Cancer Whole-Genome Analyses of Metastatic Solid Tumours", *Nature*, Vol. 575, No. 1, pp. 210 – 216, 2019.

70. Wang, Y., D. J. Miller and R. Clarke, "Approaches to Working in High-Dimensional Data Spaces: Gene Expression Microarrays", *British Journal of Cancer*, Vol. 98, No. 1, pp. 1023 – 1028, 2008.

71. Liang, Y., C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z. Xu and H. H. Zhang, "Sparse Logistic Regression With a $L^{1/2}$ Penalty for Gene Selection in Cancer Classification", *BMC Bioinformatics*, Vol. 14, No. 1, p. 198, 2013.

72. Yang, Z.-Y., Y. Liang, H. B. Zhang, H. Chai, B. Zhang and C. Peng, "Robust Sparse Logistic Regression With the $L_q(0 < q < 1)$ Regularization for Feature Selection using Gene Expression Data", *Institute of Electrical and Electronics Engineers Access*, Vol. 6, No. 1, pp. 68586–68595, 2018.

73. CAMDA 2019 Hi-Res CancerData Integration Challenge, `https://camda2019.bioinf.jku.at/doku.php`, accessed on January 01, 2019.

74. Peng, H., F. Long and C. Ding, "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 1, pp. 1226–1238, 2005.

75. Kira, K. and L. Rendell, "A Practical Approach to Feature Selection", *ML Workshop*, San Francisco, California, America, 1992.

76. Witten, I. H., E. Frank, M. A. Hall and C. J. Pal, *Data Mining, Fourth Edi-*

tion: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco, California, America, 2016.

77. Liberzon, A., C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov and P. Tamayo, "The Molecular Signatures Database Hallmark Gene Set Collection", *Cell systems*, Vol. 1, No. 6, pp. 417–425, 2015.

78. Ramirez, R., Y.-C. Chiu, A. Hererra, M. Mostavi, J. Ramirez, Y. Chen, Y. Huang and Y.-F. Jin, "Classification of Cancer Types using Graph Convolutional Neural Networks", *Frontiers of Physics*, Vol. 8, No. 203, pp. 1–14, 2020.

79. Schulte-Sasse, R., S. Budach, D. Hnisz and A. Marsico, "Integration of Multiomics Data with Graph Convolutional Networks to Identify New Cancer Genes and Their Associated Molecular Mechanisms", *Nature Machine Intelligence*, Vol. 3, No. 1, pp. 513 – 526, 2021.

80. Pfeifer, B., A. Secic, A. Saranti and A. Holzinger, "GNN-SubNet: Disease Subnetwork Detection with Explainable Graph Neural Networks", *Bioinformatics*, Vol. 38, No. 1, pp. 120–126, 2022.

81. Zhang, S., J. Xu and T. Zhang, "DGMP: Identifying Cancer Driver Genes by Jointing DGCN and MLP from Multi-Omics Genomic Data", *Genomics, Proteomics & Bioinformatics*, Vol. 20, No. 1, pp. 928 – 938, 2022.

82. Li, X., J. Ma, L. Leng, M. Han, M. Li, F. He and Y. ping Zhu, "MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis", *Frontiers in Genetics*, Vol. 13, No. 1, pp. 1–12, 2022.

83. Peng, W., R. Wu, W. Dai and N. Yu, "Identifying Cancer Driver Genes Based on Multi-View Heterogeneous Graph Convolutional Network and Self-Attention Mechanism", *BMC Bioinformatics*, Vol. 24, No. 16, pp. 1–22, 2023.

84. Zhong, Y., Y. Peng, Y. Lin, D. Chen, H. Zhang, W. Zheng, Y. Chen and C. Wu,

"MODILM: Towards Better Complex Diseases Classification using a Novel Multi-Omics Data Integration Learning Model", *BMC Medical Informatics and Decision Making*, Vol. 23, No. 82, pp. 1–18, 2023.

85. Li, B. and S. Nabavi, "A Multimodal Graph Neural Network Framework for Cancer Molecular Subtype Classification", *BMC Bioinformatics*, Vol. 25, No. 27, pp. 1–19, 2023.

86. Chatzianastasis, M., M. Vazirgiannis and Z. Zhang, "Explainable Multilayer Graph Neural Network for Cancer Gene Prediction", *Bioinformatics*, Vol. 39, No. 11, pp. 1–10, 2023.

87. Shao, L., C. Fu and X. Chen, "A Heterogeneous Graph Convolutional Attention Network Method for Classification of Autism Spectrum Disorder", *BMC Bioinformatics*, Vol. 24, No. 363, pp. 1–17, 2023.

88. Shi, H., X. Zhang, L. Tang and L. Liu, "Heterogeneous Graph Neural Network for lncRNA-Disease Association Prediction", *Scientific Reports*, Vol. 12, No. 17519, pp. 1–11, 2022.

89. Jagodnik, K. M., Y. Shvili and A. Bartal, "HetIG-PreDiG: A Heterogeneous Integrated Graph Model for Predicting Human Disease Genes Based on Gene Expression", *Plos One*, Vol. 18, No. 1, pp. 1–27, 2023.

90. Mccombe, N., J. Bamrah, J. M. Sanchez-Bornot, D. P. Finn, P. L. McClean and K. Wong-Lin, "Alzheimer's Disease Classification using Cluster-Based Labelling for Graph Neural Network on Heterogeneous Data", *Healthcare Technology Letters*, Vol. 9, No. 1, pp. 102 – 109, 2022.

91. Zhu, R., C. Ji, Y. Wang, Y. Cai and H. Wu, "Heterogeneous Graph Convolutional Networks and Matrix Completion for miRNA-Disease Association Prediction", *Frontiers in Bioengineering and Biotechnology*, Vol. 8, No. 901, pp. 1–9, 2020.

92. Kim, S., N. Lee, J. Lee, D. Hyun and C. Park, "Heterogeneous Graph Learning for Multi-Modal Medical Data Analysis", *Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, America, 2022.

93. StringDB, `https://www.string-db.org/`, accessed on January 01, 2024.

94. Hallmark, `https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp/`, accessed on February 01, 2024.

95. Yao, L., C. Mao and Y. Luo, "Graph Convolutional Networks for Text Classification", *Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, America, 2019.

96. PyG, `https://pyg.org/`, accessed on March 01, 2024.

97. Captum, `https://captum.ai/`, accessed on January 01, 2024.

98. Sundararajan, M., A. Taly and Q. Yan, "Axiomatic Attribution for Deep Networks", *International Conference on Machine Learning*, Sydney, Australia, 2017.

99. Theodoropoulos, G. E., N. V. Michalopoulos, M. P. Pantou, P. Kontogianni, M. Gazouli, T. Karantanos, M. Lymperi and G. C. Zografos, "Caspase 9 Promoter Polymorphisms Confer Increased Susceptibility to Breast Cancer", *Cancer Genetics*, Vol. 205, No. 10, pp. 508–512, 2012.

100. Zhang, M., K. Wu, M. Wang, F. Bai and H. Chen, "CASP9 As a Prognostic Biomarker and Promising Drug Target Plays a Pivotal Role in Inflammatory Breast Cancer", *International Journal of Analytical Chemistry*, Vol. 2022, No. 1043445, pp. 1–12, 2021.

101. Kumar, D., S. A. Patel, R. Khan, S. Chawla, N. Mohapatra and M. Dixit, "IQ Motif-Containing GTPase-Activating Protein 2 Inhibits Breast Cancer Angiogenesis by Suppressing VEGFR2–AKT Signaling", *Molecular Cancer Research*,

Vol. 20, No. 1, pp. 77 – 91, 2021.

102. Kumar, D., S. A. Patel, M. K. Hassan, N. Mohapatra, N. Pattanaik and M. Dixit, "Reduced IQGAP2 Expression Promotes EMT and Inhibits Apoptosis by Modulating the MEK-ERK and p38 Signaling in Breast Cancer Irrespective of ER Status", *Cell Death & Disease*, Vol. 12, No. 389, pp. 1–14, 2021.

103. Turner, A., L. cheng Li, T. Pilli, L. Qian, E. L. Wiley, S. Setty, K. T. Christov, L. Ganesh, A. V. Maker, P. Li, P. Kanteti, T. K. D. Gupta and B. S. Prabhakar, "MADD Knock-Down Enhances Doxorubicin and TRAIL Induced Apoptosis in Breast Cancer Cells", *Plos One*, Vol. 8, No. 2, pp. 56817–56825, 2013.

104. Peng, Z.-M., X. Han, T. Wang, J.-J. Li, C.-X. Yang, F.-F. Tou and Z. Zhang, "PFKP deubiquitination and stabilization by USP5 activate aerobic glycolysis to promote triple-negative breast cancer progression", *Breast Cancer Research*, Vol. 26, No. 10, pp. 1–14, 2024.

105. Tian, Y., P. Zhang, Y. Mou, W. Yang, J. Zhang, Q. Li and X. Dou, "Silencing Notch4 Promotes Tumorigenesis and Inhibits Metastasis of Triple-Negative Breast Cancer via Nanog and Cdc42", *Cell Death Discovery*, Vol. 9, No. 148, pp. 1–10, 2023.

106. Eini, M., S. Parsi, M. Barati, G. Bahramali, M. A. Zarei, J. Kiani, A. Azarnezhad and A. Hosseini, "Bioinformatic Investigation of Micro RNA-802 Target Genes, Protein Networks, and Its Potential Prognostic Value in Breast Cancer", *Avicenna Journal of Medical Biotechnology*, Vol. 14, No. 1, pp. 154 – 164, 2022.

107. Wan, L., T. Liu, Z. Hong, Y. Pan, S. T. Sizemore, J. Zhang and Z. Ma, "NEDD4 Expression is Associated with Breast Cancer Progression and is Predictive of a Poor Prognosis", *Breast Cancer Research*, Vol. 21, No. 148, pp. 1–16, 2019.

108. Deng, H., C. Wen, S. Jiang, Y. Yu, J. Zhao and B. Zhang, "Single-Cell Analy-

sis Reveals One Cancer-Associated Fibroblasts Subtype Linked to Metastasis in Breast Cancer: MXRA5 as a Potential Novel Marker for Prognosis", *American Journal of Cancer Research*, Vol. 14, No. 2, pp. 526–544, 2024.

109. Liu, Q.-W., J.-Y. Li, X. Zhang, Y. Liu, Q. yu Liu, L. Xiao, W.-J. Zhang, H.-Y. Wu, K.-Y. Deng and H.-B. Xin, "Human Amniotic Mesenchymal Stem Cells Inhibit Hepatocellular Carcinoma in Tumour-Bearing Mice", *Journal of Cellular and Molecular Medicine*, Vol. 24, No. 2, pp. 10525 – 10541, 2020.

110. Suetens, A., M. Moreels, R. Quintens, S. Chiriotti, K. Tabury, A. Michaux, V. Grégoire and S. Baatout, "Carbon Ion Irradiation of the Human Prostate Cancer Cell Line PC3: A Whole Genome Microarray Study", *International Journal of Oncology*, Vol. 44, No. 2, pp. 1056 – 1072, 2014.

111. Mei, Y., K. Li, Z. Zhang, M. Li, H. Yang, H. Wang, X. Huang, X. Li, S. Shi and H. Yang, "miR-33b-3p Acts as a Tumor Suppressor by Targeting DOCK4 in Prostate Cancer", *Frontiers in Oncology*, Vol. 11, No. 740452, pp. 1–12, 2021.

112. Liu, Y., J. Wang, T. Yang, R. Liu and Y. Xu, "Overexpression Levels of Cripto-1 Predict Poor Prognosis in Patients with Prostate Cancer Following Radical Prostatectomy", *Oncology Letters*, Vol. 18, No. 3, pp. 2584–2591, 2019.

113. Ray, J., C. Haughey, C. Hoey, J. Jeon, R. G. Murphy, L. Dura-Perez, N. Mccabe, M. R. Downes, S. Jain, P. C. Boutros, I. G. Mills and S. K. Liu, "miR-191 Promotes Radiation Resistance of Prostate Cancer Through Interaction with RXRA", *Cancer Letters*, Vol. 473, No. 3, pp. 107–117, 2019.

114. Liu, X., L. Chen, H. Huang, J. Lv, M. Chen, F. jun Qu, X. wu Pan, L. Li, L. Yin, X. gang Cui, Y. Gao and D. Xu, "High Expression of PDLIM5 Facilitates Cell Tumorigenesis and Migration by Maintaining AMPK Activation in Prostate Cancer", *Oncotarget*, Vol. 8, No. 3, pp. 98117 – 98134, 2017.

115. Shinawi, T., K. K. Nasser, F. A. Moradi, A. Mujalli, W. F. Albaqami, H. S. Almukadi, R. Elango, N. A. Shaik and B. Banaganapalli, "A Comparative mRNA- and miRNA Transcriptomics Reveals Novel Molecular Signatures Associated with Metastatic Prostate Cancers", *Frontiers in Genetics*, Vol. 13, No. 2, pp. 1–18, 2022.

116. Mao, L., C. hua Yang, J. qi Wang, W. Li, R. Wen, J. cun Chen and J. Zheng, "SATB1 is Overexpressed in Metastatic Prostate Cancer and Promotes Prostate Cancer Cell Growth and Invasion", *Journal of Translational Medicine*, Vol. 11, No. 2, p. 111, 2013.

117. Jansen, F. H., A. L. van Rijswijk, W. J. Teubel, W. M. van Weerden, S. Reneman, G.-J. van den Bemd, M. J. Roobol, C. H. Bangma, F. J. T. Staal and G. W. Jenster, "Profiling of Antibody Production Against Xenograft-Released Proteins by Protein Microarrays Discovers Prostate Cancer Markers", *Journal of Proteome Research*, Vol. 11, No. 2, pp. 728–735, 2012.

118. Cancer Research UK, `https://www.cancerresearchuk.org/`, accessed on February 01, 2024.

119. The Catalogue Of Somatic Mutations In Cancer, `https://cancer.sanger.ac.uk/cosmic`, accessed on March 01, 2018.

120. "Health, United States, 2016: with Chartbook on Long-Term Trends in Health", National Center for Health Statistics, Hyattsville, 2016, `https://www.ncbi.nlm.nih.gov/books/NBK453378/`, accessed on March 01, 2024.

121. Lin, Y., Y. Meng, X. Sun, Q. Han, K. Kuang, J. Li and F. Wu, "BertGCN: Transductive Text Classification by Combining GNN and BERT", ArXiv:2105.05727 [cs.CL], 2021.

122. Zhao, Y. and X. Song, "TextGCL: Graph Contrastive Learning for Transductive Text Classification", *International Joint Conference on Neural Networks*, Queensland, Australia, 2023.

123. Zhou, J., G. Cui, Z. Zhang, C. Yang, Z. Liu and M. Sun, "Graph Neural Networks: A Review of Methods and Applications", ArXiv:1812.08434 [cs.LG], 2018.

124. Zhang, X.-M., L. Liang, L. Liu and M. Tang, "Graph Neural Networks and Their Current Applications in Bioinformatics", *Frontiers in Genetics*, Vol. 12, No. 690049, pp. 1–22, 2021.

125. Nakamura, Y., H. Yasuoka, M. Tsujimoto, Q. Yang, S. Imabun, M. Nakahara, K. Nakao, M. Nakamura, I. Mori and K. Kakudo, "Flt-4-positive Vessel Density Correlates with Vascular Endothelial Growth factor-d Expression, Nodal Status, and Prognosis in Breast Cancer", *Clinical Cancer Research*, Vol. 9, No. 14, pp. 5313–5320, 2003.

126. Motaparthi, K., S. R. Lauer, R. M. Patel, C. I. Vidal and K. D. Linos, "MYC Gene Amplification by Fluorescence in Situ Hybridization and MYC Protein Expression by Immunohistochemistry in the Diagnosis of Cutaneous Angiosarcoma: Systematic Review and Appropriate Use Criteria", *Journal of Cutaneous Pathology*, Vol. 48, No. 1, pp. 578 – 586, 2020.

127. Du, L., B. Dai, X. Liu, D. Zhou, H. Yan, T. Shen, D. Wang and X. Tan, "KDM6B Regulates M2 Polarization of Macrophages by Modulating the Stability of Nuclear β-catenin", *Biochimica et Biophysica Acta*, Vol. 1869, No. 2, p. 166611, 2022.

128. Xun, J., R. Gao, B. Wang, Y. Li, Y. Ma, J. Guan and Q. Zhang, "Histone Demethylase KDM6B Inhibits Breast Cancer Metastasis by Regulating Wnt/β-catenin Signaling", *FEBS Open Bio*, Vol. 11, No. 8, pp. 2075–2383, 2021.

129. Wang, W., K. G. Lim, M. Feng, Y. Bao, P. L. Lee, Y. Cai, Y. Chen, H. Zhang,

D. M. Marzese, D. S. B. Hoon and Q. Yu, "KDM6B Counteracts EZH2-Mediated Suppression of IGFBP5 to Confer Resistance to PI3K/AKT Inhibitor Treatment in Breast Cancer", *Molecular Cancer Therapeutics*, Vol. 17, No. 2, pp. 1973 – 1983, 2018.

130. Akaogi, K., W. Ono, Y. Hayashi, H. Kishimoto and J. Yanagisawa, "MYBBP1A Suppresses Breast Cancer Tumorigenesis by Enhancing the p53 Dependent Anoikis", *BMC Cancer*, Vol. 13, No. 2, p. 65, 2013.

131. Zekri, A.-R. N., A. A. Bahnassy, M. Mourad, I. A. Malash, O. S. Ahmed and M. S. Abdellateif, "Genetic Profiling of Different Phenotypic Subsets of Breast Cancer Stem Cells (BCSCs) in Breast Cancer Patients", *Cancer Cell International*, Vol. 22, No. 423, pp. 1–15, 2022.

132. Moamer, A., I. Y. Hachim, N. Binothman, N. Wang, J.-J. Lebrun and S. Ali, "A Role for kinesin-1 Subunits KIF5B/KLC1 in Regulating Epithelial Mesenchymal Plasticity in Breast Tumorigenesis", *EBioMedicine*, Vol. 45, No. 1, pp. 92 – 107, 2019.

133. Kitz, J., C. Lefebvre, J. Carlos, L. E. Lowes and A. L. Allan, "Reduced Zeb1 Expression in Prostate Cancer Cells Leads to an Aggressive Partial-EMT Phenotype Associated with Altered Global Methylation Patterns", *International Journal of Molecular Sciences*, Vol. 22, No. 12840, pp. 1–16, 2021.

134. de la Rosa, J., J. Weber, M. J. Friedrich, Y. Li, L. Rad, H. Ponstingl, Q. Liang, S. B. de Quirós, I. Noorani, E. Metzakopian, A. J. Strong, M. A. Li, A. Astudillo, M. T. Fernández-García, M. S. Fernández-García, G. J. Hoffman, R. Fuente, G. S. Vassiliou, R. Rad, C. López-Otín, A. Bradley and J. Cadiñanos, "A Single-Copy Sleeping Beauty Transposon Mutagenesis Screen Identifies New PTEN-Cooperating Tumor Suppressor Genes", *Nature Genetics*, Vol. 49, No. 1, pp. 730 – 741, 2017.

135. Liu, Y., S. Zhao, J. Wang, Z. Zhu, L. Luo, Q. Xiang, M. Zhou, Y. Ma, Z. Wang and Z. Zhao, "MiR-629-5p Promotes Prostate Cancer Development and Metastasis by Targeting AKAP13", *Frontiers in Oncology*, Vol. 11, No. 754353, pp. 1–12, 2021.

136. Du, M., L. S. Tillmans, J. Gao, P. Gao, T. Yuan, R. L. Dittmar, W. Song, Y. Yang, N. A. Sahr, T. Wang, G.-H. Wei, S. N. Thibodeau and L. Wang, "Chromatin Interactions and Candidate Genes at Ten Prostate Cancer Risk Loci", *Scientific Reports*, Vol. 6, No. 23202, pp. 1–13, 2016.

137. Shiota, M., T. Tanegashima, S. Tatarano, T. Kamoto, H. Matsuyama, H. Sakai, T. Igawa, T. Kamba, N. Fujimoto, A. Yokomizo, S. Naito and M. Eto, "The Effect of Human Leukocyte Antigen Genotype on Survival in Advanced Prostate Cancer Treated with Primary Androgen Deprivation Therapy", *Prostate Cancer and Prostatic Diseases*, Vol. 1, No. 1, pp. 1–9, 2024.

138. Tuminello, S., Y. Zhang, L. Yang, N. Durmus, M. Snuderl, A. Heguy, A. Zeleniuch-Jacquotte, Y. Chen, Y. Shao, J. Reibman and A. A. Arslan, "Global DNA Methylation Profiles in Peripheral Blood of WTC-Exposed Community Members with Breast Cancer", *International Journal of Environmental Research and Public Health*, Vol. 19, No. 5104, pp. 1–15, 2022.

139. Chen, W.-J., H.-T. Zhong, H.-T. Wu, Y.-Y. Hou, Z.-J. Wu, Z. Fang and J. Liu, "NOTCH3 Inhibits Transcription Factor ZEB1 Expression and Metastasis of Breast Cancer Cells via Transcriptionally Upregulating miR-223", *Journal of Cancer*, Vol. 15, No. 2, pp. 192 – 203, 2024.

140. Chen, M.-N., Z. Fang, Z.-J. Wu, J.-W. Bai, R. Li, X.-F. Wen, G.-J. Zhang and J. Liu, "Notch3 Restricts Metastasis of Breast Cancers Through Regulation of the JAK/STAT5A Signaling pPathway", *BMC Cancer*, Vol. 23, No. 1257, pp. 1–15, 2023.

141. Hall, A., W.-T. Lu, J. D. Godfrey, A. V. Antonov, C. Paicu, C. Paicu, S. Moxon,

T. Dalmay, A. Wilczynska, P. A. J. Muller and M. Bushell, "The Cytoskeleton Adaptor Protein ankyrin-1 is Upregulated by p53 Following DNA Damage and Alters Cell Migration", *Cell Death & Disease*, Vol. 7, No. 1, pp. 2184–2198, 2016.

142. Fukushiro-Lopes, D. F., A. D. Hegel, V. R. Rao, D. Wyatt, A. Baker, E.-K. Breuer, C. Osipo, J. J. Zartman, M. Burnette, S. Kaja, D. E. Kouzoukas, S. Burris, W. K. Jones and S. Gentile, "Preclinical Study of a Kv11.1 Potassium Channel Activator as Antineoplastic Approach for Breast Cancer", *Oncotarget*, Vol. 9, No. 1, pp. 3321 – 3337, 2017.

143. Zhang, X.-X., J. hua Luo and L. qiang Wu, "FN1 Overexpression is Correlated with Unfavorable Prognosis and Immune Infiltrates in Breast Cancer", *Frontiers in Genetics*, Vol. 13, No. 1, pp. 1–16, 2022.

144. Daniunaite, K., A. Bakaviius, K. Zukauskaite, I. Rauluseviciute, J. R. Lazutka, A. Ulys, F. Jankevicius and S. Jarmalaitė, "Promoter Methylation of PRKCB, ADAMTS12, and NAALAD2 Is Specific to Prostate Cancer and Predicts Biochemical Disease Recurrence", *International Journal of Molecular Sciences*, Vol. 22, No. 6091, pp. 1–13, 2021.

145. Zhao, X., Y. Lei, G. Li, Y. Cheng, H. fan Yang, L. Xie, H. Long and R. Jiang, "Integrative Analysis of Cancer Driver Genes in Prostate Adenocarcinoma", *Molecular Medicine Reports*, Vol. 19, No. 2, pp. 2707 – 2715, 2019.

146. Atala, A., "Re: Inhibition of Ca2+ -Activated Cl- Channel ANO1/TMEM16A Expression Suppresses Tumor Growth and Invasiveness in Human Prostate Carcinoma", *The Journal of Urology*, Vol. 189, No. 6, p. 2393, 2013.

147. Tian, Q., Y.-Y. Lu, B. C. Yan and C. Wu, "Integrative Bioinformatics Analysis Reveals That miR-524-5p/MEF2C Regulates Bone Metastasis in Prostate Cancer and Breast Cancer", *Computational and Mathematical Methods in Medicine*, Vol. 2022, No. 5211329, pp. 1–13, 2022.

148. Peng, J., P. Li, Y. Li, J. Quan, Y. Yao, J. Duan, X. Liu, H. Li, D. Yuan and X. Wang, "PFKP is a Prospective Prognostic, Diagnostic, Immunological and Drug Sensitivity Predictor Across Pan-Cancer", *Scientific Reports*, Vol. 13, No. 17399, pp. 1–21, 2023.

149. Schacht, M. J., J. E. Garnett and J. T. Grayhack, "Biochemical Markers in Prostatic Cancer", *The Urologic Clinics of North America*, Vol. 11, No. 2, pp. 253–267, 1984.

150. Huang, C., R. Xu, X. Zhu and H. Jiang, "m6A-modified circABCC4 Promotes Stemness and Metastasis of Prostate Cancer by Recruiting IGF2BP2 to Increase Stability of CCAR1", *Cancer Gene Therapy*, Vol. 30, No. 2, pp. 1426 – 1440, 2023.

151. Dairo, O. O., L. D. Oliveira, E. Schaffer, T. Vidotto, A. A. A. Mendes, J. Lu, S. V. Huynh, J. L. Hicks, A. G. Sowalsky, A. M. D. Marzo, C. E. Joshu, B. Hanratty, K. S. Sfanos, W. B. Isaacs, M. C. Haffner and T. L. Lotan, "FASN Gene Methylation is Associated with Fatty Acid Synthase Expression and Clinical-genomic Features of Prostate Cancer", *Cancer Research Communications*, Vol. 4, No. 1, pp. 152 – 163, 2023.

152. Natori, Y., J. Suga, E. Tokuda, K. Tachibana, J. Imai, R. Honma, Y. Azami, M. Noda, E. Sasaki, S. Watanabe, T. Ohtake and S. Saji, "E3 Ubiquitin Ligase NEDD4 Affects Estrogen Receptor α Expression and the Prognosis of Patients with Hormone Receptor-Positive Breast Cancer", *Cancers*, Vol. 15, No. 2, p. 539, 2023.

153. Sanger, F., G. Air, B. Barrell, N. Brown, A. Coulson, C. Fiddes, C. Hutchison, P. Slocombe and M. Smith, "Nucleotide Sequence of Bacteriophage phi x174 DNA", *Nature*, Vol. 265, No. 1, pp. 687–695, 1977.

154. Guo, Y., Z. Wei and B. J. Keating, "Machine Learning Derived Risk Prediction

of Anorexia Nervosa", *BMC Medical Genomics*, Vol. 9, No. 1, p. 4, 2016.

155. Jiang, X., B. Cai, D. Xue, X. Lu, G. F. cooper and R. E. Neapolitan, "A Comparative Analysis of Methods for Predicting Clinical Outcomes using High-Dimensional Genomic Datasets", *Journal of the American Medical Informatics Association*, Vol. 21, No. 2, pp. 312–319, 2014.

156. Quang, D., Y. Chen and X. Xie, "Dann: a Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants", *Bioinformatics*, Vol. 31, No. 1, pp. 761–763, 2015.

157. Zhou, J. and O. G. Troyanskaya, "Predicting Effects of Noncoding Variants with Deep Learning Based Sequence Model", *Nature Methods*, Vol. 12, No. 1, pp. 931–934, 2015.

158. Ghandi, M., D. Lee, M. Mohammad-Noori and M. Beer, "Enhanced Regulatory Sequence Prediction using Gapped k-mer Features", *Plos Computational Biology*, Vol. 10, No. 2, p. 12, 2014.

159. Lee, D., D. Gorkin, M. Baker, B. Strober, A. Asoni, A. McCallion and M. Beer, "A Method to Predict the Impact of Regulatory Variants from DNA Sequence", *Nature Genetics*, Vol. 47, No. 1, pp. 955–961, 2015.

160. Ritchie, G., I. Dunham, E. Zeggini and P. Flicek, "Functional Annotation of Noncoding Sequence Variants", *Nature Methods*, Vol. 11, No. 1, pp. 294–296, 2014.

161. Muller, H., R. Jimenez-Heredia, A. Krolo, T. Hirschmugl, J. Dmytrus, K. Boztug and C. Bock, "VCF.Filter: Interactive Prioritization of Disease-Linked Genetic Variants from Sequencing Data", *Nucleic Acids Research*, Vol. 45, No. 2, pp. 567–572, 2017.

162. Fu, Y., Z. Liu, S. Lou, J. Bedford, X. Mu, K. Yip, E. Khurana and M. Ger-

stein, "Funseq2: a Framework for Prioritizing Noncoding Regulatory Variants in Cancer", *Genome Biology*, Vol. 15, No. 2, p. 480, 2014.

163. Yuan, Y., Y. Shi, C. Li, J. Kim, W. Cai, Z. Han and D. D. Feng, "DeepGene: an Advanced Cancer Type Classifier Based on Deep Learning and Somatic Point Mutations", *BMC Bioinformatics*, Vol. 17, No. 476, pp. 1–14, 2016.

164. Liu, X., A. Krishnan and A. Mondry, "An Entropy-Based Gene Selection Method for Cancer Classification using Microarray Data", *BMC Bioinformatics*, Vol. 6, No. 2, p. 76, 2005.

165. Kim, K. and S. Cho, "Ensemble Classifiers Based on Correlation Analysis for DNA Microarray Classification", *Elsevier Neurocomputing*, Vol. 70, No. 1, pp. 187–199, 2006.

166. Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri and C. Bloomfield, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, Vol. 286, No. 1, pp. 531–537, 1999.

167. Guyon, I., J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, Vol. 46, No. 3, pp. 389–422, 2002.

168. Ruiz, R., J. Riquelme and J. Aguilar-Ruiz, "Incremental Wrapper–Based Gene Selection from Microarray Data for Cancer Classification", *Elsevier Pattern Recognition*, Vol. 39, No. 1, pp. 2383–2392, 2005.

169. Duan, K.-B., J. Rajapakse, H. Wang and F. Azuaje, "Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data", *Institute of Electrical and Electronics Engineers Transactions on Nanobioscience*, Vol. 4, No. 4, pp. 228–234, 2005.

170. Cawley, G. C. and N. L. C. Talbot, "Gene Selection in Cancer Classification using Sparse Logistic Regression with Bayesian Regularization", *Bioinformatics*, Vol. 22, No. 2, pp. 2348–2355, 2006.

171. Piao, Y., M. Piao, K. Park and K. H. Ryu, "An Ensemble Correlation-Based Gene Selection Algorithm for Cancer Classification with Gene Expression Data", *Bioinformatics*, Vol. 28, No. 2, pp. 3306–3315, 2012.

172. Zhang, H., H. Wang, Z. Dai, M. shun Chen and Z. Yuan, "Improving Accuracy for Cancer Classification with a New Algorithm for Gene Selection", *BMC Bioinformatics*, Vol. 13, No. 1, p. 298, 2012.

173. Khotanloua, H., O. Colliotb, J. Atifc and I. Bloch, "3D Brain Tumor Segmentation in MRI using Fuzzy Classification, Symmetry Analysis and Spatially Constrained Deformable Models", *Elsevier Fuzzy Sets and Systems*, Vol. 160, No. 3, pp. 1457–1473, 2009.

174. Wang, J., Q. J. Liu, S. Yuan, W. Xie, Y. Liu, Y. Xiang, N. Wu, L. Wu, X. Ma, T. Cai, Y. Zhang, Z. Sun and Y. Li, "Genetic Predisposition to Lung Cancer: Comprehensive Literature Integration, Meta-Analysis, and Multiple Evidence Assessment of Candidate-Gene Association studies", *Scientific Reports*, Vol. 7, No. 8371, pp. 1 – 13, 2017.

175. Cabanillas, R., M. Diñeiro, D. Castillo, P. C. Pruneda, C. Penas, G. A. Cifuentes, Á. de Vicente, N. S. Durán, R. Alvarez, G. R. Ordóñez and J. Cadiñanos, "A Novel Molecular Diagnostics Platform for Somatic and Germline Precision Oncology", *Molecular Genetics & Genomic Medicine*, Vol. 5, No. 2, pp. 336–359, 2017.

176. Chen, Y., J. Sun, L.-C. Huang, H. Xu and Z. Zhao, "Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations", *BioMed Research International*, Vol. 2015, No. 491502, pp. 1–9, 2015.

177. Pan, Z., Y. Chen, J. Liu, Q. Jiang, S. Yang, L. N. Guo and G. He, "Design, Synthesis, and Biological Evaluation of polo-like kinase 1/eukaryotic elongation factor 2 kinase (PLK1/EEF2K) Dual Inhibitors for Regulating Breast Cancer Cells Apoptosis and Autophagy", *European Journal of Medicinal Chemistry*, Vol. 144, No. 1, pp. 517–528, 2018.

178. Penault-Llorca, F., F. Bertucci, J. Adélaïde, P. R. Parc, F. Coulier, J. D. Jacquemier, D. Birnbaum and O. Delapeyrière, "Expression of FGF and FGF Receptor Genes in Human Breast Cancer", *International Journal of Cancer*, Vol. 61, No. 2, pp. 170–176, 1995.

179. Seo, Y.-S., O.-H. Kang, R. Kong, T. Zhou, S.-A. Kim, S. Ryu, H.-R. Kim and D.-Y. Kwon, "Polygalacin D Induces Apoptosis and Cell Cycle Arrest via the PI3K/Akt Pathway in Non-Small Cell Lung Cancer", *Oncology Reports*, Vol. 39, No. 4, pp. 1702–1710, 2018.

180. Pfefferle, A. D., Y. N. Agrawal, D. C. Koboldt, K. L. Kanchi, J. I. Herschkowitz, E. R. Mardis, J. M. Rosen and C. M. Perou, "Genomic Profiling of Murine Mammary Tumors Identifies Potential Personalized Drug Targets for p53-deficient Mammary Cancers", *Disease Models and Mechanisms*, Vol. 9, No. 1, pp. 749–757, 2016.

181. Manning, C. D., P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2008.

182. Zerbino, D. R., P. Achuthan, W. A. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. A. Cummins, A. Gall, C. García-Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. Loveland, T. Maurel, W. M. McLaren, B. Moore, J. M. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish,

S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. D. Yates and P. Flicek, "Ensembl 2018", *Nucleic Acids Research*, Vol. 46, No. 1, pp. 754–761, 2018.

183. Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran and D. Lancet, "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses", *Current Protocols in Bioinformatics*, Vol. 54, No. 1, pp. 1–33, 2016.

184. Peterson, T. A., I. I. M. Gauran, J. Park, D. Park and M. G. Kann, "Oncodomains: A Protein Domain-Centric Framework for Analyzing Rare Variants in Tumor Samples", Vol. 13, pp. 1–24, 2017.

185. Marx, V., "Cancer: Hunting Rare Somatic Mutations", *Nature Methods*, Vol. 13, No. 1, pp. 295–299, 2016.

186. Craig, J., "Complex Diseases: Research and Applications", *Nature Education*, Vol. 1, No. 1, p. 184, 2008.