University of Glasgow | School of Computing Science

Honours Individual Project Dissertation

The Assessing of Mutation Gene Status Prediction Driven by Cancer Related Single Base Substitution Signatures

Zhouyang Shen

April 09, 2021

Abstract

With the evolution of human cancer research, many works have been proposed to understand the function of tumours. They are based on large-scale analysis of gene expression for molecular profiling of human cancer. Those methods have revealed the tremendous complexity of the alternations across the genome, the proteome of tumours, and the transcriptome. Therefore, traditional statistical modelling has been substantially restricted to make accurate predictions on those characteristics and further understand the role genes play in cancers.

Instead of exploring the limited information provided by standard features, more and more research has authenticated that the deciphering of the mutational signatures in cancers could present more insights into the biological mechanisms involved in carcinogenesis and normal somatic mutagenesis. The mutational signatures have manifested their applicability in cancer treatment and cancer prevention.

Some of them later proposed a binary classifier to detect specific mutation type of tumours within the breast, ovarian and pancreatic cancers and achieved the excellent result of prediction.

This project ranks the importance of mutation signatures by understanding the operating of each tumour. The significant signatures are used to determine the mutation status of genes within those tumours to study the relationship between them. Similar to previous work, machine learning models are built to achieve the goals. However, there are two challenges. Firstly, a binary classifier containing full connected layers is hard to distinguish 32 cancer types and recognize the complex SBS signature patterns to derive the various driver gene mutation status. Secondly, the sample distribution is under imbalance, and it could potentially cause the biased learning of the built model.

In this project, several technologies are developed to address these issues. The stratified sampling has been deployed to address the issue of the biased classifier. The softmax regression model has been proposed to deal with the multi-class classification as well as extracting patterns. The convolutional neural network is constructed to resolve the problem brought by data sparsity and locating the local spatial critical features from the patterns in the identification of mutation status.

The experimental results show the AUC scores in most cancer classification tasks could be around 1.0. The models could give a reasonable prediction on the top frequently mutated driver gene's mutation status within certain cancer types as the AUC could achieve around 0.80 to 1.0. These results have established a fundamental thesis for the later detection of multiple cancer-related gene mutation statuses in the future.

Acknowledgements

I would like to thank my supervisor, Dr.Ke for providing a great deal of support throughout this project. Without his guidance, this work would not have been possible. Dedicated to my parents, who supported me through my four years of studying at Glasgow University and who always encouraged me to achieve the best I can.

Education Use Consent

I hereby grant my permission for this project to be stored, distributed and shown to other University of Glasgow students and staff for educational purposes. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Signature: Zhouyang Shen Date: 6 April 2021

Contents

1 Introduction

1.1 Overview

With the development of human diseases research, many works about understanding the function of tumours have been proposed based on the enormous scale endeavours of molecular profiling of human cancer. The data used in those experiments have reflected the massive complexity of the changes across the genome, proteome, and transcriptome of tumours. Moreover, more researches have been done. The fact that the cancer is not likely to be related to one specific gene is identified and further complicated cancer understanding. Consequently, conventional statistical modelling has been considerably limited to make valuable predictions on those characteristics.

Therefore, the new method of using the data-driving method has been developed to analyze the previous challenges to address these issues with the information perceived. in Hoadley et al. (2018), The idea is identified to use the features from 5 platforms and mutational signatures from the TCGA platform for classifying different cancers and their driver gene's mutation status.

This project aims to investigate tumours and build a data-driven system to test the feasibility of predicting different gene mutation status within those tumours to further understand tumour functioning by examining its correlation with the corresponding genes and the workability of later unknown gene prediction. The concept is that this system does not rely on the complicated molecular features of tumours, which is proved to be exceedingly costly in traditional methods. Instead, we extract competent characteristic information from the collected somatic mutation data based on cancer types before applying a machine learning model to predict gene mutation status. This project consists of the following three steps:

1. Data processing,
2. Extracting cancer-related features,
3. Predicting gene mutation status.

To be more specific, the base section is to extract mutational signatures from collected data samples and re-sampling them to ensure the appropriate distribution. The second part is to find the significant features based on the cancer type. Some capable models, like logistic regression or softmax regression, could be exploited as a cancer-types classifier, and the filter to extract weights of those valid signatures that could determine the significance of those features for the identification of gene mutation status in those cancers since we assume that those significant features are the patterns for specific cancer. The third step of the work is to predict the most frequented driver gene mutation status in each cancer based on the samples, which only shared the top-K most cancer-type related features. In this process, two convolutional neural networks are applied for comparison and determined to improve the accuracy of prediction and reduce the potential problems elicited from the data sparsity.

1.2 Motivation

Over the last few decades, the large scale of molecular profiling cancer has resulted in an exceptional understanding of the tumours' inner operating. However, the immense complexity

of the alternation across genome, transcriptome and proteome of the tumour has limited the prediction of traditional mathematical modelling. To further understand the tumour's inner working and define them in a molecular perspective and assist with minimizing the costly tumour sequencing process in the future. One of the approaches is to analyze the most common gene's mutation status within individual cancers to understand the relationship between genes and each cancer type and learn how to obtain better performance on predicting the gene mutation status within those tumours.

In order to be able to predict the gene mutation status within each tumour and deduce more mutation status of the related gene in the future, the first step of the process would be searching for a way to effectively distinguishing the tumour and its corresponding driver gene's mutation status. However, with current methods, it is generally difficult to achieve this goal. The general information perceived from the somatic mutation is not enough for any models to give an accurate prediction on tumour types as well as providing convincing clinical explanations of them. The limited structure of the mutation information has also proved to be infeasible in predicting the gene mutation status in an efficient way.

In that case, the suggested alternative direction is to perform cancer type analyses based on the newly discovered features denoted as mutational signatures. Moreover, as previous works have received an unexpected breach by deploying a machine learning-based approach to detect the samples' gene mutation status in the limited cancer types using the mutational signature. For instance, Davies et al. (2017) has proposed a model called HRDetect to detect BRCA1/BRCA2-deficient tumors within breast, ovarian and pancreatic cancers and has achieved the excellent result of prediction(area under the curve (AUC) = 0.98) ,The theory that the prediction of more gene's mutation status across more cancers (32 of them) might be applicable via the mutational signatures is established. To be more precisely, for a given unknown samples with knowing its mutational signature exposures, the proposed model should be able to derive their cancer types, as well as their mutation status of the gene related to those cancer types.

Though, there are several challenges/issues:

1. **How to process the raw cancer data set and elicit all mutational signature exposures from different samples?**
2. **How to process the feature-extracted samples effectively to avoid the absence of data in training and testing?**
3. **How to address the challenge brought by various cancer types to the previous classification models, e.g., the linear regression model and to build an adaptive machine learning model to remove the noisy mutational signatures from each sample by knowing its cancer type?**
4. **How to predict the gene mutation status, which can minimize the loss to gain accurate results?**
5. **How to effectively evaluate the results we obtained from the built model and analyze and validate them by combining them with the pathological phenomenons?**

1.3 Aim and Objective

In this project, in order to verify the possibility of gaining high performance of classifying various cancer types, the top frequently mutated driver gene's mutation status of the samples given mutational signature, there are several works to be contemplated. At the first stage, the prioritized focus would be deriving the mutational signature from the raw data file and searching for the method to effectively re-sample the data to support the later process.

After the preparation process is done, a machine learning-based approach is explored to classify single base substitution (SBS) signature exposure in each sample to identify the corresponding

cancer type. Afterwards, with the presence of the extracted SBS signatures from data preprocessing, simultaneously, the previous machine learning model should, therefore, not only be trained to identify the cancer type but also be provided as a useful tool in discovering the corresponding SBS signature weights, which present the correlation between a given cancer type and total extracted SBS signatures in each cancer.

Subsequently, to reduce the noisy data that might affect our gene mutation status prediction, a pruning step is processed to remove the lesser informative SBS signatures in each cancer. All samples then will be pruned, and their selected SBS signatures in their corresponding cancer type are inputted as the feature data of predicting gene mutation status.

Finally, an indirect relationship between genes and cancer types could be built through the shared mutational signatures. The prediction of the driver gene in cancer can then be performed and evaluated. To achieve the procedures described above, there are several objectives to be reached:

- **Feature extraction:** The samples' single-base substitution signature exposures, the genetic mutation status, and its cancer type need to be extracted in each sample within the original raw data file.
- **Softmax regression:** A classification model is to be developed based on SoftMax regression to support the classification of 32 cancer types based on the extracted mutational signatures.
- **Feature selection and CNN implementation:** Based on each sample's informative single base substitution signature exposures, two predictors based on a convolutional neural network are to be deployed to identify its genetic mutation status.
- **Evaluation:** K-fold cross-validation is also to be evaluated on the data set. The results could be examined and validated with the provided clinical information.

1.4 Dissertation Organization

The rest of this dissertation is organized as follow:

- Chapter 2: This chapter mainly provides the background information that is required to understand the processes of project, including cancers, genes, their relationships, and so on. Moreover, Several related works have also been discussed to motivate this research.
- Chapter 3: This chapter mainly provides the analysis on the data set collected.Few issues are identified here.Several observations are also made to facilitate and motivate the approaches for the defined problem and provide explanations for the hypothesis that has been formulated.
- Chapter 4: This chapter demonstrates the detailed design, e.g., the system workflow and the machine learning models based on the analysis and approaches given in the former chapter. Therefore, the design is at the methodology level and could be exploited and implemented under different platforms.
- Chapter 5: This chapter offers overall implementations of the proposed design, which is mentioned in the previous chapter. Many important details are revealed, and some preliminary results, such as tallying components and extractions of SBS signatures, are shown here.
- Chapter 6: This chapter mainly discussed the experimental result and concentrated on evaluating the implementations. These results are in line with our expectations and confirms the majority of the pathological phenomenon.
- Chapter 7: This chapter concludes the entire project and illustrates future improvements and ideas to extend from the work.

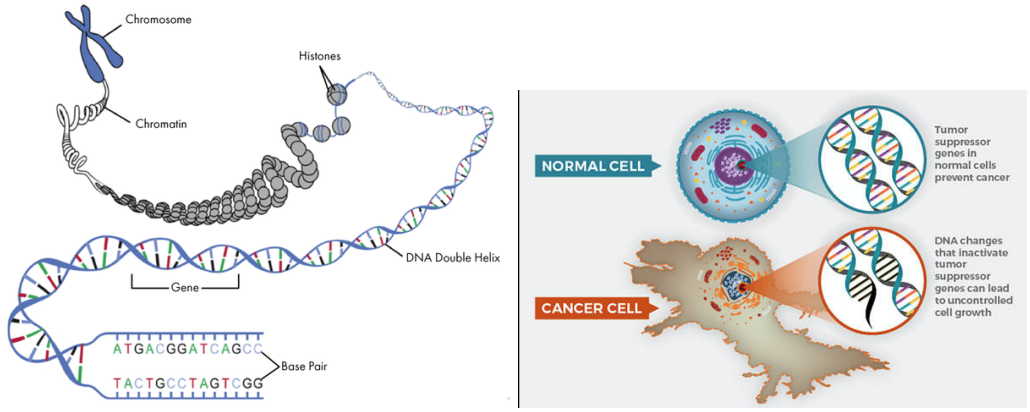
2 Background

In this chapter, we first introduce the basic concepts of cancers and genes in Section 2.1. Then, Section 2.2 shows the relationship between cancers and genes. It reveals that the cancer driver genes play crucial roles in identifying different cancers. In Section 2.3, we illustrated different mutation types and found several previous works that can evince the theory of cancer-type classification based on the Single-Base Substitution (SBS) signatures.Finally, we present several related works in detail and conclude this chapter.

2.1 Cancers and Genes

Gene is an inheritance unit that passes a trait from parent to child. In biology, a gene is the basic physical and functional unit of heredity. Genes are made up of DNA. Some genes act as genetic instructions to make molecules called proteins. Sequences of DNA that were converted into strands of messenger RNA could be used as the basis for building associated protein piece by piece. The DNA molecules coiled around chromosomes are seen as long strings on which the sequences of genes are like discrete beads. (see in Figure 2.1a).

Cancer is a type of diseases that involves abnormal cell growth with the potential to invade or spread to other parts of the body. Cancer can destroy recruited human cells and turn them into pathological organisms or the building blocks of tumours, posing significant threats to human health. It is caused by certain changes to genes, the basic physical and functional units of inheritance. (see in Figure 2.1b).Cancer is a global disease, and the number of people who are suffering from cancer is growing. In 2019, about 18 million new cases occur annually, it caused about 8.8 million deaths, and the death rate reaches up to 15.7% according to World Cancer Report 2014 Stewart BW (2014). It is predicated by Bray et al. (2012) that 24.6 million number of people would suffer cancer and the death number will reach to 13 million by 2030. For male,



(a) *Genes arranged in chromosomes Jung (2019)* ***(b)*** *Normal cell vs cancer cellInstitute (2015)*

Figure 2.1: *Genes and cancers*

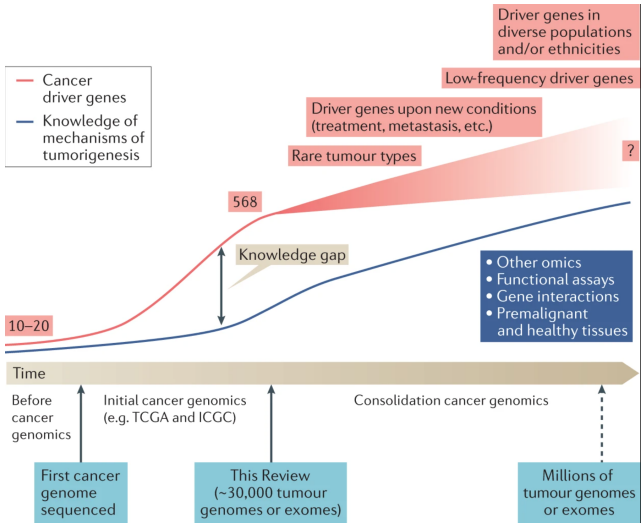


Figure 2.2: *An overview of history of cancer genomics Martinez-Jiménez et al. (2020)*

stomach cancer, prostate cancer, colorectal cancer and lung cancer are one of the top threats to them. As for females, the most common types are breast cancer, cervical cancer, colorectal cancer and lung cancer. Importantly, as reported by Vineis and Wild (2013), around 33%-50% of all cancer deaths are preventable. According to this study, a particularly effective approach to fighting cancer is primary prevention based on current knowledge of risk factors. In other words, we may predict cancer before having cancer, and the method of predicting is highly associated with genes.

2.2 Correlation between cancers and genes

In the previous study, we have shown a strong association between cancer and gene. Take breast cancer as an example; microarray-based gene expression profiles can separate different tumour classes, thus having a crucial impact on our understanding of breast cancer. From a study on 99 breast cancer patients by Sotiriou et al. (2003), the gene patterns generated by cDNA microarrays were found to be highly linked with estrogen receptor (ER) status, moderately associated with grade, and are related with detailed clinic-pathological characteristics and clinical outcomes in both node-negative and node-positive breast cancer patients. As there are more studies on breast cancer classification based on gene expression profiling, Reis-Filho and Pusztai (2011) Others researched the potential clinical use of the molecular classification system and looked into possibilities of deploying prognosis and predictors of breast cancer. The classification of gene expressions is of great significance in cancer diagnosis and drug discovery. Few questions the correlation between cancer and gene, yet some expressed concern about the limited diagnostic ability of cancer classification studies. Besides that, some questions that we are still far away from technique of integrating primary gathered molecular data into electronic health records. We may still far away from using it optimally as part of a clinical workflow. However, methods and techniques of cancer classification are being explored and researched. According to Tan and Gilbert (2003), identification of genes and genomes expressed and not expressed in tumour cells is of great difficulty in microarray analysis of cancer gene expression profiles. As a result, they explored other techniques in cancer classification, such as the C4.5 decision tree and bagged and boosted decision trees. After that, more techniques are being explored in order for analyzing cancer gene relations. For instance, the modified K-Nearest Neighbors technique and cervical cancer cell detection and classification system were proposed by Ayyad et al. (2019) and Ghoneim

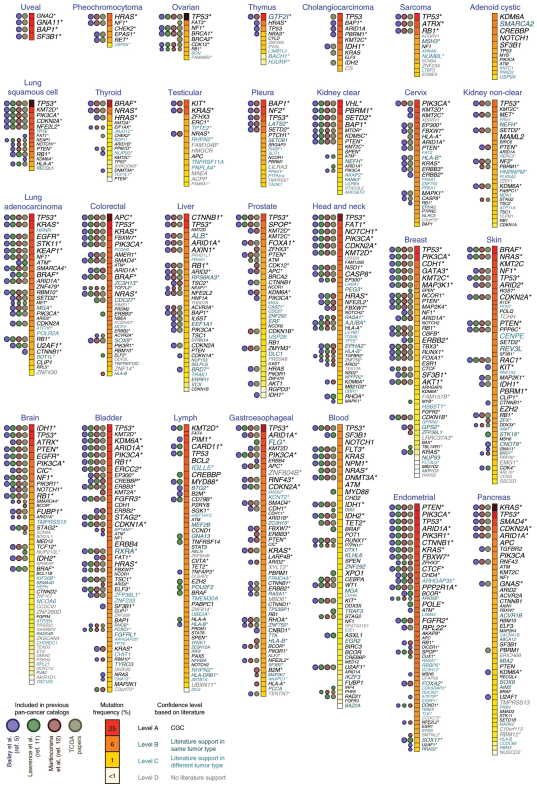


Figure 2.3: *A catalog of driver genes in human cancer Dietlein et al. (2020)*

et al. (2019) respectively.

2.2.1 Cancer driver genes

The driver genes of the cancer are genes that give cells an opportunity to grow when they are mutated. The pattern of somatic mutations they observed in the tumours in the cohort deviated from what would be expected of neutral mutations, and this is the reason why identification of driver genes can be discovered. The fundamental goal of cancer research is to understand cellular transformation mechanisms, which includes cancer detection and cancer treatment. One milestone towards this objective is to identify all the genes that can drive tumours.

According to Martinez-Jiménez et al. (2020), the research on mutational driver genes started from the identification of the first cancer genes through sequencing of the first tumours. This happens even before the start of the cancer genomics era (see in Figure 2.2). Many projects were carried on in identifying the sets of genes that drive malignancies so as to provide a road map for the systematic and comprehensive identification of mutational driver genes. These cancer driver mutations are catalogued and could be used as reference points for diagnosing and treating the disease.

According to a newly published paper by Dietlein et al. (2020), a new computational tool is used to detect cancer-causing genes at unprecedented sensitivity, which would high-potentially be a useful resource to the community of cancer researchers. They identified 460 genes that are important for the development of cancer, uncovering tumour-gene associations that had not previously been identified, which is comprehensive resources of cancer driver genes to date (see in Figure 2.3).

2.3 Mutational Signatures and SBS Signatures

In the past few decades, there have been several waves of technological advancement in the characterization of mutations in cancer genomes. A significant number of large-scale studies have revealed many mutational signatures of a variety of human cancer types. The idea of using the mutation signatures to classify the different cancer types has been researched and developed in the past few years as well.

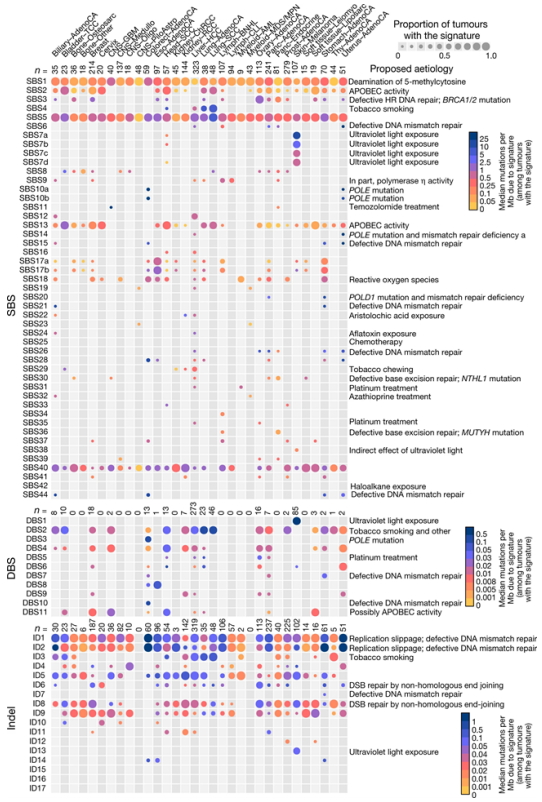


Figure 2.4: The number of mutations contributed by each mutational signature to the PCAWG tumours Alexandrov et al. (2020).

2.3.1 Mutational Signatures

Somatic mutations occur in all sections of the human body and run through the entire life cycle. They are the result of multiple mutational processes involving the slight intrinsic infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA and defective DNA repair.

Different mutational processes do produce distinctive combinations of mutation types, known as "Mutational Signatures". Typically, The most of the mutational signatures that have been utilized in the recent research are single base substitution (SBS) signatures, Doublet base substitution (DBS) signatures and small insertions, deletions (ID) signatures and rearrangement signatures.

In a research conducted by Nik-Zainal et al. (2016), the whole genomes of 560 breast cancers and non-neoplastic tissue from each individual were sequenced. Other than using a mathematical approach to extract mutational signatures, they also applied rearrangement mutational processes. This process is the first formal investigation in classifying signatures. Rearrangement signatures

followed an unsupervised hierarchical clustering based on proportions of rearrangement signatures in each cancer (see in Figure 2.5). As a result, at least 12 base substitution mutational signatures and 6 rearrangement signatures are utilized to help to find somatic mutations, and thus paving the way to a comprehensive understanding of the origins and consequences of somatic mutations in breast cancer.

2.3.2 Single Base Substitution Signatures

Currently, three different variant classes are considered, resulting in the following sets of mutational signatures. Single based substitution is one of the three types of base mutational signatures. There are different types of intermediate values that can be used. Generally known, there are six classes of base substitution, C>A, C>G, C>T, T>A, T>C, T>G (see in Figure 2.6). However, from the information appended from 5' and 3' adjacent bases, the mutations are therefore could be expanded to 96 possible types. In our experiment, in order to achieve simplicity, we also chose a single-based substitution signature, which is defined as a single-nucleotide variation marker, as the classification feature. Recently, more and more realized the enormous information that could be retrieved from those molecular characteristics and the possible reasons could be:

1. It is the mutational signature that described the fundamental somatic mutation process of replacement of single nucleotide substitution. Therefore, it contains useful information in recording the basic mutation process.
2. It provides clear distinguishing of various different signatures collected and the discovery of different signatures is reliable,comprehensive and it is at the stage of technology maturity.

Moreover, with the researches that have been done, using SBS Signature has been revealed to be a promising method to classify cancers. In the past few years, there are many large-scale scrutinises have revealed the indication that mutational signatures such as Single base substitution signatures have spanned the spectrum of human cancer types, including the latest effort by the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes. PCAWG Network Alexandrov et al. (2020) uses data from more than 23,000 cancer patients.

Nevertheless, with the recent studying of the SBS signatures, Alexandrov et al. (2020) has found that the specific type of mutational signature does provide more scrutinized plausible underlying aetiologies for explaining the occurrence of the cancers compared to 2 other common base mutation signatures. As we can observe from Figure 2.4, the base mutational signatures extracted in each cancer by the sigProfiler tool from the graph below Alexandrov et al. (2020).

2.4 Related Works

2.4.1 HRDetect

In those years, many works have been proposed to do cancer-types classification and other analyses on gene expression data Lu and Han (2003); Tan and Gilbert (2003), including the sequencing of gene Kamps et al. (2017); Loo et al. (2012), mutational signatures Davies et al. (2017); Alexandrov et al. (2013), and so on. In this section, we particularly pay attention to HRDetect, which gives first inspiration of this paper. Followed by the introduction of CNN, which is an method used in gene classification.

At Present, the most remarkable improvements in mutational signature analysis-based diagnosis are in the category of breast cancer. The fact that those tumours with mutations in BRCA1/2 are defective in the HRR process is realized and enhanced in performing the HRDetect model. With the HRR-deficiency features from the complete mutation catalogue of base substitutions, indels, and structural rearrangements utilized by the computational tool of the model, HRDetect could predict BRCAness. (i.e., a BRCA1/2-associated phenotype) with a sensitivity of almost 100%,

which is an improvement on the sensitivity of 60% obtained by most traditional copy number-based tests uses. Hoeck et al. (2019).The use of this tool revealed that microhomology-mediated indels, two COSMIC signatures (further referred to as CS) and two rearrangement signatures (further referred to as RS) are correlated with HRR deficiency .

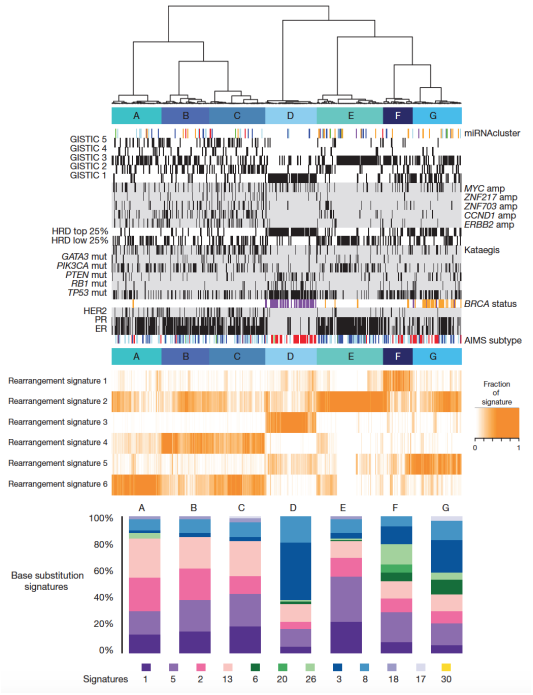


Figure 2.5: *Integrative analysis of rearrangement signatures. Nik-Zainal et al. (2016)*

In the implementation perspective, HRDetect (Davies et al. (2017)) designed a binary classifier to detect BRCA1/BRCA2-deficient samples.HRDetect is based on a lasso logistic regression model (see Equation 2.1). In their work, the BRCA1/BRCA2 deficiency is reasoned by mutational signatures.Based on these signatures, HRDetect could identify whether the breast cancer suffers BRCA1/BRCA2 deficiency or not. On a data set of 371 breast cancer samples, the area under the ROC curve could achieve 0.98.

$$P(C_i = BRCA) = \frac{1}{1 + e^{-(\beta_0 + x_i^T \beta)}} \tag{2.1}$$

This approach is based on the lasso logistic regression model, which can only be used to do binary classification. And in this scenario, the problem of classifying various cancers can not be achieved. Also, Their approach has the limitation: The research has been constrained within two typical genes, The BRCA1 and BRAC2, it has not been adapted to classify multiple labels such as the different gene's mutation status in different cancer types. For example, this approach is not functional when it comes to predict the different diver gene's mutation status within different samples in various cancer types,where we already acknowledged that there are distinctive driver genes in different cancers

2.4.2 Softmax Regression

The Softmax Regression is also called Multinomial logistic regression. It is the category of logistic regression that could be used for multi-class classification under the specification that the classes

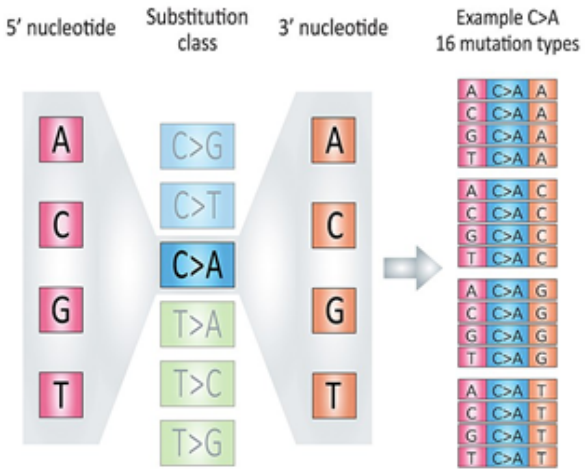


Figure 2.6: *The graph illustrating the 96-mutation type as component wiki (2021).*

are mutually exclusive. It is commonly acknowledged that the standard Logistic Regression model can only be deployed to perform binary class classification. In this project, the softmax function could be exploited to replace the sigmoid function deployed by the traditional logistic regression models to expand the prediction to the multinomial probability distribution. One characteristic to be noticed here is that as the output of the implied model always depends on the summation of the inputs and parameters, it is also considered as a "non-linear" version of linear regression. Therefore, it is efficacious when the goal is to extract the contribution of the features after the calculation.

2.4.3 CNN

A convolutional neural network(CNN) is a class of deep neural networks which has been successfully applied in the fields of visual imagery and speech recognition. The network uses a mathematical operation called convolution, which is performed by extracting information with the convolutional kernel and helping the model learn local and global structures from the input data. For image classification problems, these structures extracted information such as edges, curves, corners, etc. Similarly, CNN uses this strategy to learn the topology of similar networks.

Luo et al. (2019) firstly combined CNN with similarity network to predict driver genes. In their research, the convolution is achieved by combining mutation-based features with gene similarity networks. Specifically, given the feature vectors of gi and its k nearest neighbors gs1, gs2, …, gsk, a feature matrix Φ_i is constructed by arranging the 2k vectors into a 2k × nf matrix, which is then used in the convolution process(see in Figure 2.7). In addition, Ghoneim et al. (2019) propose a CNNs-based cervical cancer detection and classification system inspired by image processing as well. One of three CNN models they investigated has a shallow architecture. Figure 2.8 illustrates that the model has two convolutional layers that are fully connected with each other followed by a softmax (output) layer. After the training was finished, it was fine-tuned by a training subset of the target database. The proposed CNN-ELM-based system achieved 99.5% accuracy in the detection problem and 91.2% in the classification problem, demonstrating the great impact of CNN in gene classification and prediction.

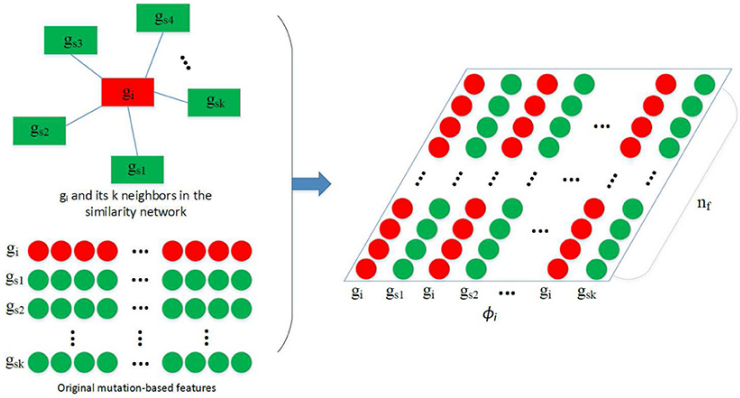


Figure 2.7: *The construction of Φ_i used in the convolution Luo et al. (2019)*

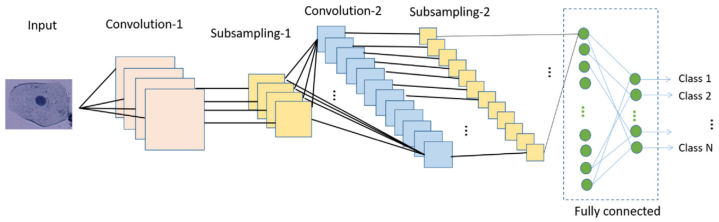


Figure 2.8: *Architecture of the shallow CNN model. Ghoneim et al. (2019).*

2.5 Chapter Summary

In this chapter, we have introduced several key terms and surveyed many previous studies to help with understanding the project. We first learned the background of this research and understood the correlation between cancer and genes, of which cancer driver genes has been specifically addressed. Apart from that, mutation signatures and SBS signatures have been introduced and demonstrated by previous studies to classify cancers successfully. Then, the simple introduction of the softmax regression model is also appended to briefly summarize its utility . Finally, we introduced HRDetect and CNN as examples of successful application in cancer-type and gene classification.

3 Analysis

In this chapter, several observations are described to reveal the relationship between SBS signatures and cancer types and explain the feasibility of classifying them in a practical way. In the Section 3.2, we present the problems to be solved to achieve the objective of this research. Section 3.3 explains why it needs to use a softmax regression model to carry out cancer-types classification. Furthermore, the 1-D convolutional neural networks are also deployed to do gene status prediction for gaining high accuracy.

3.1 Observations

3.1.1 Observation of the processed data set

Firstly,for finding the single base substitution signature, the raw data file (MAF file, Mutation Annotation Format) only contains the variant type of simple nucleotide variation that is obtained from the TCGA Data Portal is employed.The inclusive data set contains 32 cancers , see (table 3.1). Secondly,the genome reference of GRCh38 is also certified to ensure reliable genomics analysis results than the previous versions as it could provide corrected misassembled regions compare to the obsoleted reference version.

Moreover, the research has performed on an overall of 9637 samples. For each cancer classes, the number of samples is recorded and shown in Fig 3.1. In the data, most of the samples are labeled with breast cancer, The second disease share by most of the samples is Lung adeno-carcinoma and the least sample's label we have is Lymphoid Neoplasm Diffuse Large B-cell Lymphoma(DLBC)(37 samples).

3.1.2 Observation of the features

As mentioned in Section 2.4.1, a binary classifier, namely HRDetect, could be used to detect the BRCA1/BRCA2-deficient samples based on the mutational signatures. Motivated by this work, a new model could be developed to support multiple cancer-types classification with the SBS signatures which is one of the fundamental features that used by HRdetect model as the input features. To practically assess if the utilization of SBS signature as feature to classify on the cancer is applicable.The observation has been proceeded on the part of processed data collected. For example, in Figure 3.3a, two different samples has very similar SBS signature exposures, while Figure 3.3b shows the samples from different samples have different subsets of SBS signatures.This is useful information as we could confirm the idea again practically that the mutational signature, specifically, the single-base substitution signatures are varied significantly in different cancers and therefore, it could theoretically be utilized as valuable information in identifying different tumours from given samples.

Moreover, the observation from the above analysis has also elicited whether we could use those subsets of the important SBS signatures listed within each cancer to help facilitate the higher classification performance of the gene mutation status. Therefore, the concept is constructed with the significant SBS signatures are extracted for each cancer types and inputted as features to

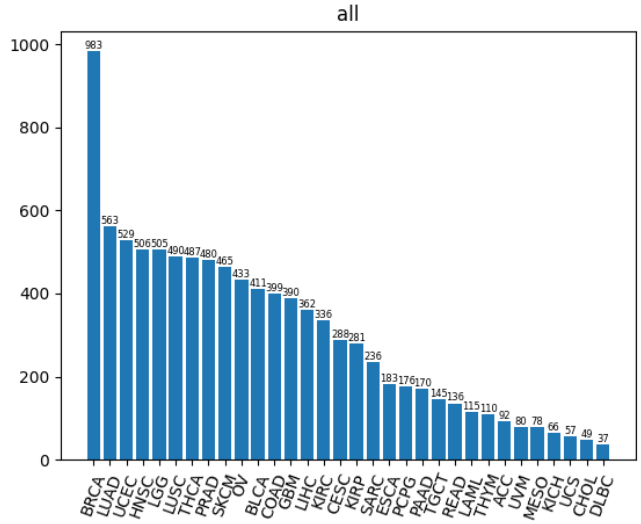


Figure 3.1: *The sample distribution of different cancer types on the whole data set.*

predict the gene mutation states based on a machine learning model. Compared with directly using the cancer type to make the prediction, the extracted SBS signatures could contain more information and be expected to predict accurately.

3.1.3 Observation of top frequently mutated gene in each cancer

As the introduction has been made, the strong relationship between cancer and their driver gene is often a trending topic, and it is the domain where most of the recent research has been directed on. In this project, For gaining reliable result and avoid verbosity,we only deployed the analysis and prediction of the top frequently mutated driver gene in each cancer,those driver genes are identified in each cancer based on statistical analysis and the prior knowledge contributed by the driver gene supplementary table by Bailey et al. (2018). The result of analysis is shown in Fig 3.2. The observation that has been made is that most of the cancers do have TP53 as their top frequently mutated driver gene in the data obtained. Moreover, another crucial finding that is also made is that although the gene is the top frequently mutated driver gene in cancers, the mutated samples of those genes are not balanced with the non-mutated samples in some of those cancers, which might cause the problem of the class imbalance or the crisis of missing labels during the later training or testing process.

3.2 Problem Analyzing

During the early stage, The initial data set used in the previous experiments is obtained and combined from different papers. The SBS signature exposure values are approximated integer value and, therefore, may lose the meaningful threshold, after gaining the inferior results from the data file. Like other papers, which start their research by generating their data from the data downloaded directly from the TCGA platform,the initial problem to be solved is to process data cautiously and rigorously.

From the observations made above, we identified the number of cancer classes (32) that needed to be distinct. It is essential as the number of types of the output of the built model is determined by this finding. The sample numbers of different classes are also recognized and reported. With the

Table 3.1: *The 32 kinds of cancer types from the data.*

Cancer Type Name	Abbreviation
Adrenocortical carcinoma	ACC
Bladder Urothelial Carcinoma	BLCA
Breast invasive carcinoma	BRCA
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC
Cholangiocarcinoma	CHOL
Colon adenocarcinoma	COAD
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC
Esophageal carcinoma	ESCA
Glioblastoma multiforme	GBM
Head and Neck squamous cell carcinoma	HNSC
Kidney Chromophobe	KICH
Kidney renal clear cell carcinoma	KIRC
Kidney renal papillary cell carcinoma	KIRP
Acute Myeloid Leukemia	LAML
Brain Lower Grade Glioma	LGG
Liver hepatocellular carcinoma	LIHC
Lung adenocarcinoma	LUAD
Lung squamous cell carcinoma	LUSC
Mesothelioma	MESO
Ovarian serous cystadenocarcinoma	OV
Pancreatic adenocarcinoma	PAAD
Pheochromocytoma and Paraganglioma	PCPG
Prostate adenocarcinoma	PRAD
Rectum adenocarcinoma	READ
Sarcoma	SARC
Skin Cutaneous Melanoma	SKCM
Testicular Germ Cell Tumors	TGCT
Thyroid carcinoma	THCA
Thymoma	THYM
Uterine Corpus Endometrial Carcinoma	UCEC
Uterine Carcinosarcoma	UCS
Uveal Melanoma	UVM

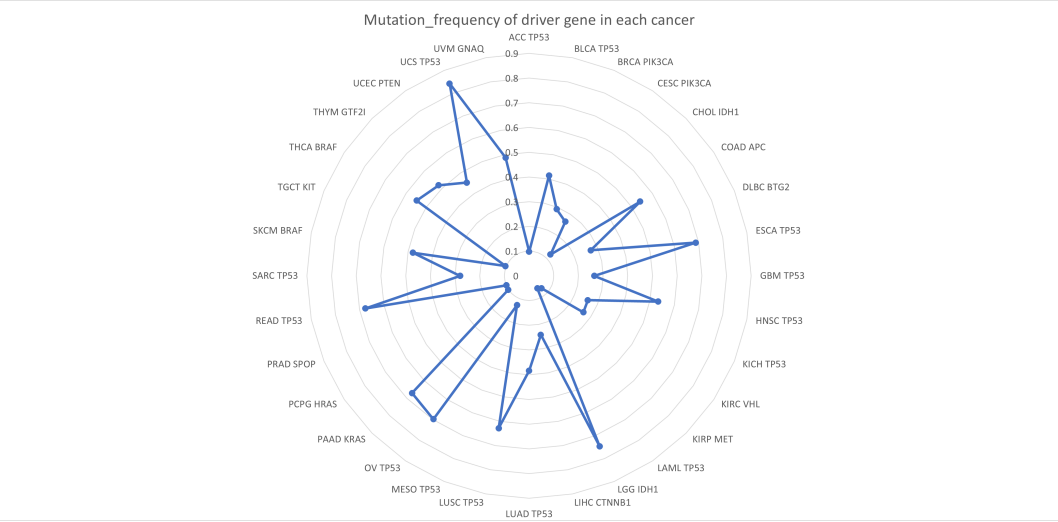
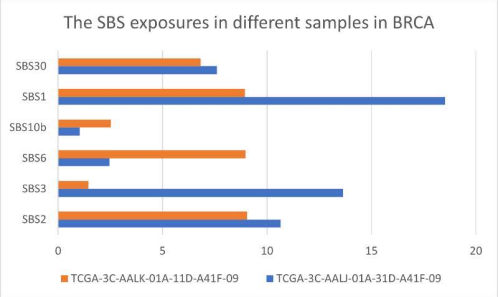
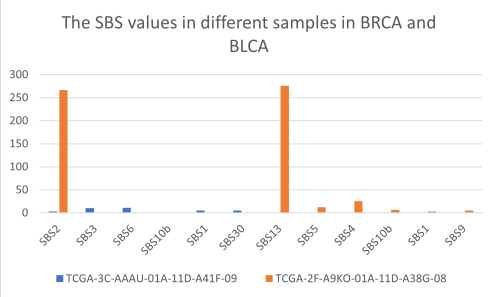


Figure 3.2: *The mutation frequency of top frequently mutated driver gene in each cancer.*



(a) SBS signatures of different samples in Breast invasive carcinoma.



(b) SBS signatures in Breast invasive carcinoma and Bladder Urothelial carcinoma.

Figure 3.3: *SBS signature exposures of some samples in the data set.*

information provided, we diagnosed the imbalanced distribution problem of the cancer classes. Therefore, a new method called stratified sampling is elicited and suggested to overcome the problem. Also, the distinguishing characteristic of the SBS signatures in each cancer type is realized and again verified the fundamental concept of deploying them as classification features for separating cancers given the samples. Moreover, related to the findings above, another hypothesis of using the unique set of determinable SBS signatures in each cancer as the features to predict gene mutation status is proposed. The conjecture is that we could find those determinable SBS signatures in each cancer types by taking their weights forged in the cancer classification model as the measuring metric. Furthermore, with the inspection been made on the topmost frequently mutated gene in each cancer, we established the categories of the labels of the driver genes that are sighted to be utilized in the later classification process. Nevertheless, with the scrutiny, the imbalanced gene mutation frequency for some of the driver genes in certain cancers is discerned and concerned as the vital problem for the validation of the model that needs to be attended to.

3.3 Approach

3.3.1 Data handling

With the identified problem of imbalanced classes, the proposed algorithm should be deployed to address the problem. In this scenario, In order to secure the prediction and the training of the model could be performed on all classes or labels to reduce the possibility of misclassification. The idea to search for the strategy to represent both training data and the testing data in a certain way to imitate the actual distribution in the overall dataset is constructed and strengthen.

3.3.2 Cancer types classification

Cancer types classification is a relatively easy job and some simple models, like logistic regression Davies et al. (2017), are guaranteed to achieve good accuracy. To support multi-class classification defined above, the *sigmoid* logistic function in HRDetect is replaced with the *softmax* function $\Phi_{softmax}$.

Another problem to be identified is that, like most of the deep learning models, For instance, in the CNN, it is hard to collect the weights of every SBS signature in each cancer. The convolutional neural network, which uses shared kernel weights to process features is only for its receptive field. Therefore, a softmax model, which can help identify the related weights of each SBS signatures in each cancer type, is deployed to address this problem and act as filter to extract SBS signatures' contributions in cancer identification for the later prediction of gene mutation status.

The cancer-types classification can hence be modelled as a K multi-class classification, :

$$\hat{y} = argmax_{k \in \{1,2,...,K\}} F_k(x)$$

, where the function F is to be searched and the label $\hat{y}$ is the predicted label of the cancer type which can be encoded into one-hot vectors.

3.3.3 Gene mutation status prediction

Apart from cancer classification task, Gene mutation status prediction is modelled as a multiple binary classification problem, i.e., one binary classifier is trained for each label independently. It is tricky for several challenges:

- It needs to identify the significant SBS signatures of each cancer types. These SBS signatures information contains much more information than the cancer type itself. They are proved to be helpful in predicting the gene mutation status.
- Compared with the whole space of SBS signatures, the information that can be extracted from total SBS signatures of each cancer types could be very limited. This information is usually located at a few small ranges. therefore, we need to select an adaptive model, which can help with identifying the spatially local feature patterns.

To address the first challenge, the weights extracted from the cancer-types classification can be used to identify the important SBS signatures. The SBS signatures with the top K highest weights could be used to represent the precise molecular features of each cancer. Moreover, two 1-D convolutional neural networks are developed to learn the spatially local feature patterns. The benefits of convolutional neural networks in this senario are:

1. They enable deep neural network structures and could be fitted to more complicated functions.
2. Their convolutional structure could improve the computation efficiency and learn the local feature patterns fast. At the end of the convolutional neural networks, the sigmoid function should ideally be used as the output of a binary classifier to provide the predicted mutation status probability.

Unlike the classification problem, the gene mutation status prediction can be modelled as a K multi-label classification because the gene's mutation status is assumed to be independent. The approach is supposed to be finding a model that maps inputs x to binary vectors y, where the content of the labels (y) are settled as 0(the gene is not mutated) and 1(the gene is mutated).

3.3.4 Validation

To judge whether our model has performed decent job, certain evaluation strategies are supposed to be delivered. At this project, The K-fold cross-validation is nominated, and the training, testing, and validation of all models are to be based on it. In k-fold cross-validation, the whole data set is divides into k partitions. $k-1$ out of k data partitions are used as the training data, while the remaining data partition is used as the testing data. The whole process could be repeated for k times and it gives k validation result on different data partitions.

The benefit of k-fold cross validation is to reduce bias and increase the confidence of validation, because it could evaluate the validation of the machine learning model on the whole dataset, instead of a sub-partition and therefore produce generalized report.

Moreover, four of the performance measures are applied to evaluate the models,they are: Accuracy, Precision, Recall, F1 Score (see Equations 3.1). In a validation stage, there are four kinds of station

Table 3.2: *Four performance measures of a prediction model.*

	Predicted class		
Actual class		class = True	class = False
	class = True	True Positive (TP)	False Positive (FP)
	class = False	False Negative (FN)	True Negative (TN)

of measurement for a prediction result (see in Table 3.2).

$$\begin{aligned} accuracy &= \frac{TP + TN}{TP + FP + FN + TN} \\ precision &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + TN} \\ F1 &= 2 * \frac{recall * precision}{recall + precision} \end{aligned} \tag{3.1}$$

Furthermore, To comprehensively validate our model, the receiver operating characteristic curve(ROC) and the Area Under Curve (AUC) are also added as part of the evaluation metrics, where the two axis of the ROC graph is defined as True positive rate against False positive rate of the prediction, the formula of defining those terms is provided below (see in Equations 3.2), the area under curve is representing the degree or quantification of separability of the model based on the ROC.

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} \\ FPR &= \frac{FP}{FP + TN} \end{aligned} \tag{3.2}$$

3.4 Chapter Summary

This chapter first concluded the problems of this research via some observations, The overall data distribution, the sample size of the data, the different type of cancer types are illustrated here, other questions such as the usability of the SBS signatures is also analyzed and answered here and used to form the design concept. Several hypotheses are also made as to the foundation concepts of the later work. Then, it analyzed the requirements and showed very abstract design points with the approaches to arrive at them. Afterwards, it also concisely showed the evaluation metrics that are potentially indispensable for the project assessments and provided the detailed computation. The next chapter will give an overview of the designs and methods for this project and show the detailed design step by step.

4 Design

In this chapter, the approaches are introduced in detail, including the data set preprocessing, stratified sampling and machine learning algorithm used for cancer-types classification and gene mutation status prediction. Section 4.1 gives a workflow overview of our system. In Section 4.2, it deploys a signatures extraction tool to handle the raw data set and generate the the SBS signature exposure for each sample. A softmax regression based model is used to support multiple cancer-types classification in Section 4.4. In addition,Section 4.5 fulfilled the design structure of training multiple binary classification models, which are based on 1-D convolutional neural network for predicting the probability of genes mutation.

4.1 Overview

As shown in Figure 4.1, the overall workflow of our system consists of: (1) Data processing, (2) Stratified sampling, (3) Cancer-types classification, (4) The analyses of weights of the SBS signature in each cancer type, and (5) Gene mutation status prediction.

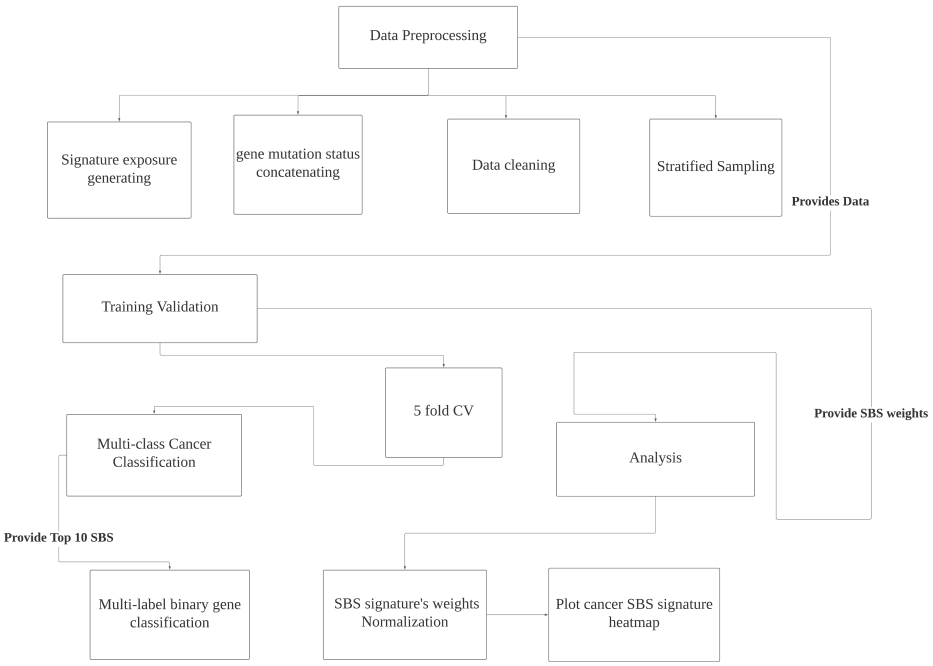


Figure 4.1: *The overall workflow of our design.*

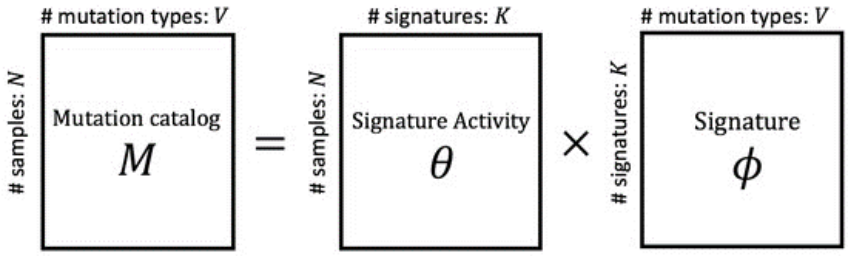


Figure 4.3: *An illustration of non-negative matrix factorization of mutation catalogues generated by tallying components to predict mutation signatures Matsutani and Hamada (2020).*

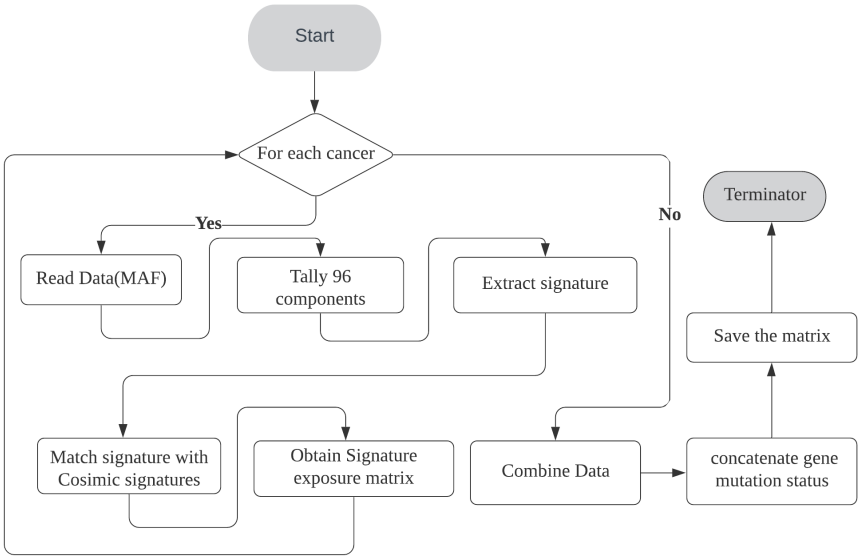


Figure 4.2: *The workflow of data preprocessing.*

4.2 Data Preprocessing

This section introduces how to extract the SBS signature exposure from the data set provided by Institute (2021). As discussed in Section 3.2, the SBS signature exposure are important features to classify the cancers and predict gene mutation status.

In order to classify the gene mutation status and its corresponding cancer type with the the SBS signature exposure. The matrix contains sample id, the SBS signature exposure , and the relative gene mutation status of each sample as columns is essentially required. The process of obtaining the gene mutation status of each sample is trial ,the extracting of the gene mutation status for each sample is performed by concatenate the tumour sample barcode with its Hugo symbol listed in each of the different cancer type in MAF file.

The signature identification is separated into 4 stages, they are:

1. Reading the MAF file downloaded from the GDC Data Portal and create MAF object for each cancer with the R package maftools provided.
2. Tallying the 96 components in each sample listed in MAF object and generate sample-by-mutation types matrix.

Figure 4.4: *The comparison between the Cosmic SBS signature and signatures extracted for BRCA.*

3. Signatures are extracted by performing Bayesian non-negative matrix factorization, as they could be observed from the graph. The assumption that has been made is that the mutation in each sample are signified by combination of multiple signatures. Therefore, the mutation catalogues generated above are to indicate the number of mutations that is represented in the genome in each sample. The mutation catalogues are factorized into the signature activity matrix θ and signature distribution matrix Φ.(see in Figure 4.3) The signature active matrix θ is to demonstrate the mutation signatures distribution in each samples and the signature distribution matrix Φ is to indicate the extracted mutation signatures of values spread in each mutation types. This project mainly uses the signature activity matrix as the input data of the cancer-types classification and the gene mutation status prediction.
4. To match the mutation signatures extracted with the the SBS signature exposure provided on the COSMIC and understand their etiologies, the comparison between identified signatures and reference signatures from COSMIC database is explored. Then, the SBS signature exposure extracted for each cancer types are visualized. The similarities between the extracted signatures and the reference signatures in BRCA as an example could be observed in the format of heatmap (see in Figure 4.4).This comparison is critical as the SBS signatures utilized in the later process is the most similar signature which has identified by the automatic similarity identification.Later, the SBS signature exposure and gene mutation status for every sample in different classes is concatenated.

After data processing, in each column of the output matrix, the information contains sample id and the corresponding SBS signature exposure and tumour types as well as the gene mutation status for each samples can therefore be acquired and perceived.

4.3 Stratified Sampling

At this section, we provide the strategy to deal with the problem that might be elicited from the data distribution problem.As introduced before, the evaluations of the models are based on the method called K-fold cross validation. However, if the traditional data separation process is performed, there might be the situation that certain classes of the cancer samples could not be separated and distributed well in each of the cross validation folds, and therefore could cause the bias or the missing training opportunities to the classifier and loss the accuracy. Subsequently, we proposed the strategy that could minimizes selection bias and ensures that the entire population group is represented in each individual fold. The overall process of the stratified sampling is shown in the Figure 4.5. To be concise, the whole process could be summarized as extracting the percentage of samples that having the top frequently mutated driver gene in their cancer

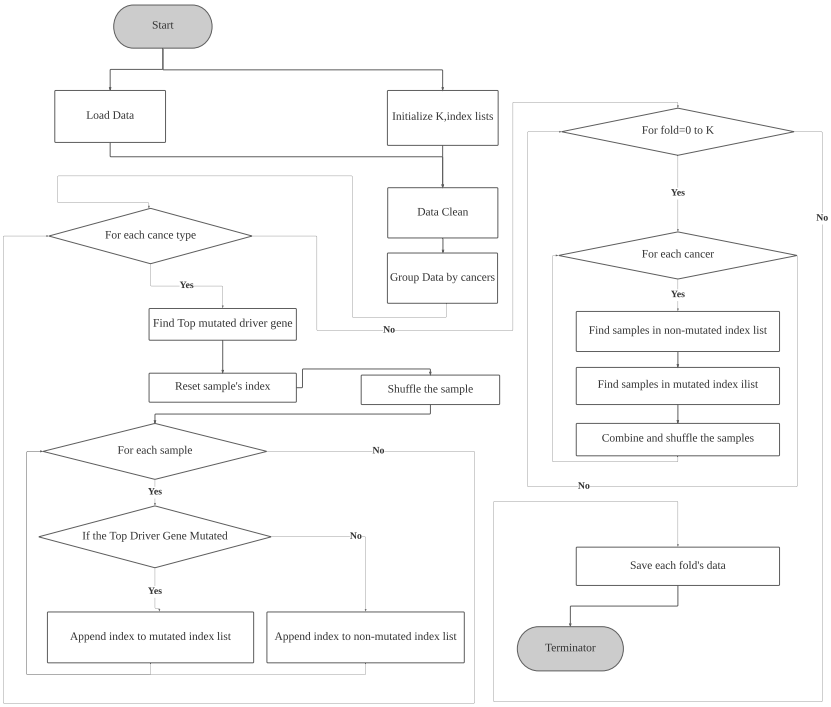


Figure 4.5: *Stratified sampling process.*

type mutated and the percentage of samples that having the driver gene not mutated from the whole data set in that cancer and assign to each of the fold to ensure the complete distribution of the classes and mutation labels in each individual partition is satisfied.

4.4 Cancer-types Classification

Goodfellow et al. (2016) shows that the softmax regression model is widely used in single-labelled, multi-class classification tasks. Unlike the convolutional neural network, which uses shared kernel weights to process features only for its receptive field. The shared kernel filters make it hard to collect the weight of every SBS signature exposure, the softmax regression as introduced before act as linear regression and therefore can help us directly extract the contribution of the features from the model. As shown in Figure 4.6 softmax regression model. It applies a linear transform to convert the input features of each sample to a output vector at a length of 32 (the number of cancer types). Then, the softmax function (a non-linear transform)is used to output a vector at the length of the class number and form the probability of being classified to each cancer type. The softmax regression model is formatted by Figure 4.1 for k types of cancers:

$$P(y = j|z^i) = \Phi_{softmax}(z^i) = \frac{e^{z^i}}{\sum_{j=0}^{k} e^{z_k^i}}, z^i = W^T x^i + b \tag{4.1}$$

, where x_i is the the SBS signature exposure of the i-th sample.

After constructing the abstracted learning model, Figure 4.7 shows the training, testing, and validation workflow based on this model. The whole data set is divided into 6 folds. five folds for cross validation and one validation data set. The system trains the softmax regression model based on the randome data of 4 cross validation folds for 1000 epochs with a batch size of 16.

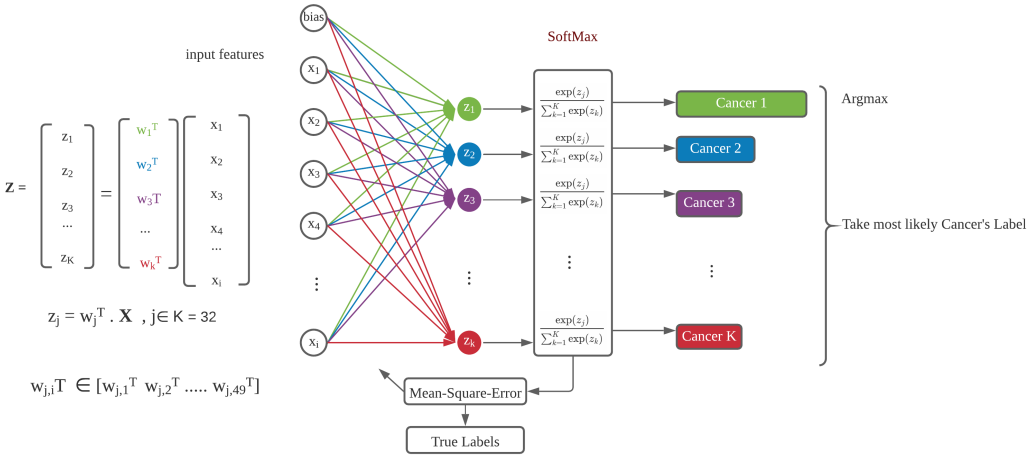


Figure 4.6: *Our softmax regression model.*

On each epochs, it first fits the parameters on the training data and use the accuracy tested on the training data to update the optimized accuracy and test on the testing set for examining our model in the cross validation. After the training and testing process, it evaluates the model on the validation data set which has never taken part in training or testing procedures to assess on the generalization ability of the model and extract weights from the learnt model for later procedures.

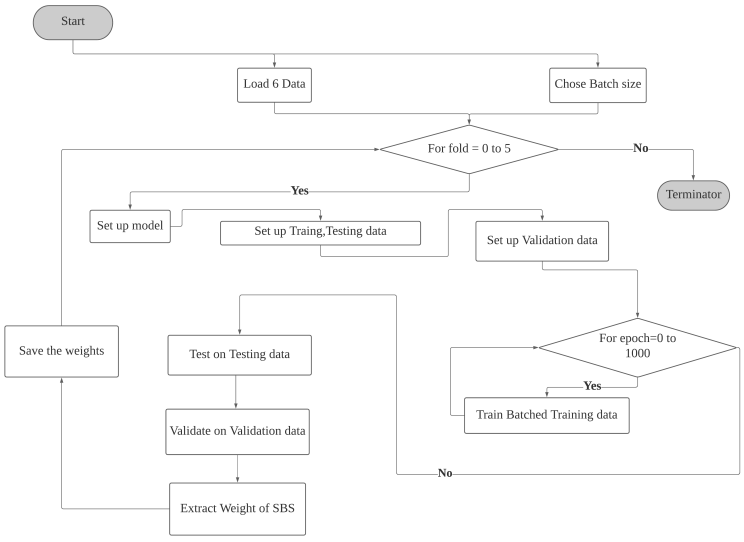
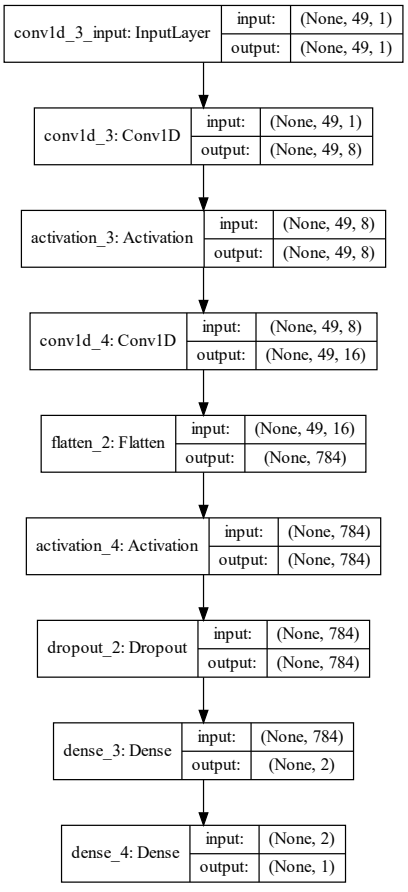
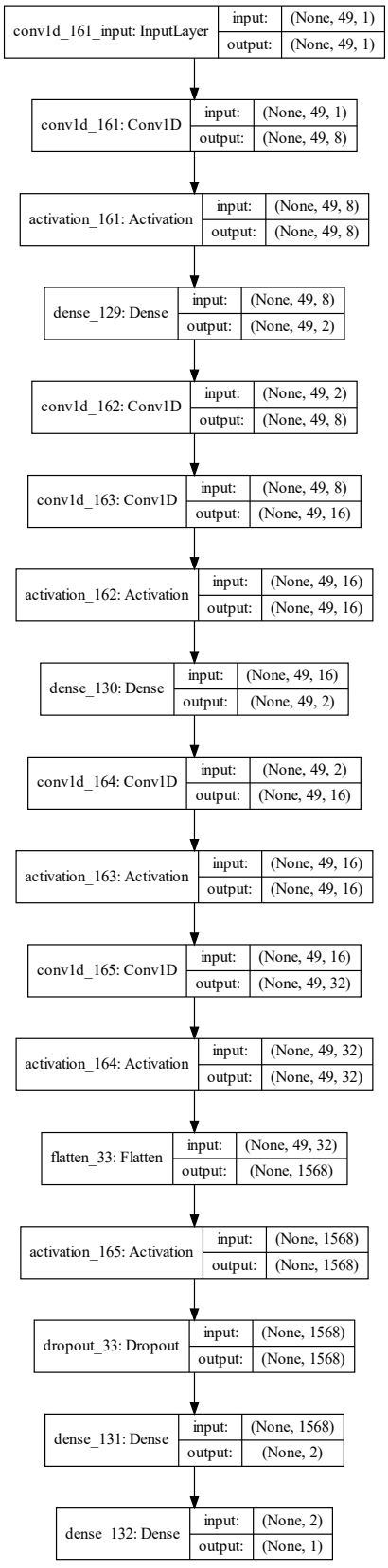


Figure 4.7: *The workflow of cancer-types classification.*



(a) *A simple CNN with 9 layers.*



(b) *A complex CNN with 17 layers.*

Figure 4.8: *Two convolutional neural networks (CNNs) for the prediction of gene mutation status*

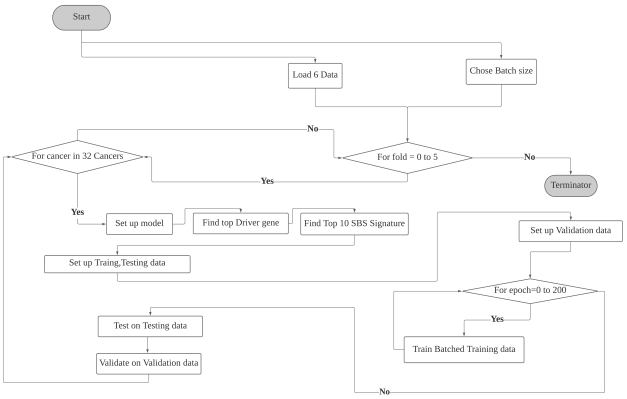


Figure 4.9: *The workflow of the gene mutation status prediction.*

4.5 Gene Mutation Status Prediction

As mentioned in Section 3.3.3, different from the single-label multi-class problems, the prediction of gene mutation status could be modelled as a multiple binary classification problem. The 1-D CNN models are developed for the prediction of gene mutation status within the 32 cancers. In Figure 4.8, it shows two CNNs used in this part: a simple CNN structure with 9 layers and a deeper CNN with 17 layers. For a specified gene, the model predicts whether the gene is mutated or not based on the subsets of SBS signature exposure, which is extracted based on the weights given by the cancer-types classification model in the formal design.

Figure 4.9 shows the detail workflow of the gene mutation status prediction. Given a cancer, if the absolute value of the weight of one SBS signature is relatively higher than the rest, it will be regarded as a significant SBS signature to that cancer. For every cancer type, it finds the top related SBS features (10 in our case) based on the weights trained in the softmax regression model in Section 4.4.

Similar to what have been done in Section 4.4, the system uses 5-fold cross validation to train and test our prediction model. It trains 32 CNN models at each fold for 200 epochs with a specified batch size of 1280 which could give us optimized performance and use the performance on the training data partition in each batch to update the better model accuracy. Similarly, the same process is also performed on the complex CNN to formulate the comparison of the performance and form the discussion.

4.6 Chapter Summary

In this chapter, the overall design of our system is presented in the beginning: (1) The data processing strategy that extract the require features as well as generating the data format needed, (2) The stratified sampling applied to ensure the reasonable distribution of the overall classes and mutation labels across the different folds (3) A softmax regression model for cancer-types classification as well as extracting the essential features, (4) A multiple binary model based on CNN for gene mutation status prediction. Furthermore, each sub-section explained the design by workflow diagram for each step.

In the next Chapter, We will focus on how the design was implemented. It would be structured to show the challenges when implementing the design, such as specific languages, technical choices, libraries, and the algorithms used.

5 Implementation

This chapter introduces how to implement our proposed design in real system. Section 5.1 and Section 5.2 present the environment setup and libraries used in this project. Section 5.3, 5.4, and 5.5 respectively give the implementations of data processing, cancer-type classification, and gene mutation status prediction.

5.1 Environment Setup

This project is done on a personal computer, equipped with Intel i9-8950HK CPU (2.9 GHz and six cores) and 32 GB DRAM. The operating system is windows 10 and installed a package management system and environment management system, namely Conda. Conda Anaconda (2021) can easily creates, saves, loads and switches between environments.

5.2 Software and Packages

The programming language used is based on Python 3.7 and R 4.0.3. Table 5.1 summarize the major libraries used in this project:

Table 5.1: *The libraries and software used in this project.*

name	version	function
conda	4.3.30	Package and environment management system
matplotlib	3.34	Plot the result data
maftools	0.9.30	Summarize, Analyze and Visualize MAF files from TCGA
numpy	1.20.1	Data process and supoort matrix operation
pandas	1.1.5	Data loader and process
Sigminer	1.2.5	Extraction of Genomic alteration signature
scikitlearn	0.24.1	Evaluation metric, such as roc, auc
scipy	1.5.4	Deploy cosine distance for finding similarities
seaborn	0.11.1	Draw heatmap
torch	1.7.1	build softmax regression models
tensorflow	1.6.0	severs as the backend of Keras
Keras	2.1.5	build convolutional neural network models

5.3 Data Processing

5.3.1 Data Preprocessing

The generation of the necessary matrix used in later process is based on Sigminer Shixiang et al. (2020), which is an accessible toolkit to extract, analyse and visualize mutational signatures. The simplified R code for extracting SBS signature exposure is shown in List 5.1.

```r
library(sigminer)
# concatenate the gene mutation status with each sample
b<-organ@data
sampleid_gene<-as.data.frame.array(t(table(b$Hugo_Symbol,b$Tumor_Sample_
    Barcode)))
# Tally a Genomic Alteration Object,organ--cancer type
mt_tally <- sig_tally(
    organ,
    ref_genome = "BSgenome.Hsapiens.UCSC.hg38",
    useSyn = TRUE
)
# Bayesian Non-negative Matrix Factorization, nrun:runtimes, stable--
    robust strategy
mt_sig2 <- sig_auto_extract(mt_tally$nmf_matrix,
    K0 = 10, nrun = 10,
    strategy = "stable"
)
# get signature exposure
matrix<-get_sig_exposure(mt_sig2)
)
```

Listing 5.1: *Data preprocessing*

```python
# load the data, organ is cancer type
overall = overall.fillna(0)  # handling the NaN
types = list(train.groupby("organ"))  # we group the data using cancer types for
    future sampling
indexs1 = [] # the list to append the all of mutated samples in each cancer
indexs0 = [] # the list to append the all of non-mutated samples in each cancer
for cancer, data in types:
    index = list(range(len(data))) # rest the index in each cancer
    np.random.shuffle(index) # shuffle the index
    index_cancer_0 = [] ,index_cancer_1 = [] # store index not mutated and mutated
    for indx in index:
        # getting the top frequently mutated driver gene in cancer
        gene = tool.find_top_gene(cancer_type, gene_prob_in_cancer)
        if data.iloc[indx][gene][0] == 0: # if the driver gene is not mutated
            index_cancer_0.append(indx)
        if data.iloc[indx][gene][0] == 1: # if the driver gene is mutated
            index_cancer_1.append(indx)
    # we only want to take (total/6) samples from each cancer and put into folds
    indexs0.append((index_cancer_0, int(len(index_cancer_0) / 6)))
    indexs1.append((index_cancer_1, int(len(index_cancer_1) / 6)))
for fold in range(6):
    tmp_result = []
    for i, type_ in enumerate(types):
        cancer, data = type_
        num_0 = indexs0[i][1],num_1 = indexs1[i][1] # find samples we want to
    assign to each fold
        if fold + 1 < 6:
            index0 = indexs0[i][0][fold * num_0:(fold + 1) * num_0]
            index1 = indexs1[i][0][fold * num_1:(fold + 1) * num_1]
        else:
            # take the rest and put into the last fold
# combine the index and shuffle it and find data and split them to save to each
    fold
```

Listing 5.2: *The Stratified Sampling*

5.3.2 Stratified Sampling

In this part, the previous designed stratified sampling strategy is deployed and the algorithm of taking percentage of samples from the overall data as well as appending them to each of the individual fold is provided.As it is shown in List 5.2. The main idea is to first find the percentage of samples that are having the top frequently mutated gene in that cancer mutated, and that of it is not mutated from the total amount of the samples from that cancer type. Afterwards, we append those shuffled percentages of data from the overall dataset to each of the individual folds for gaining the guaranteed distribution and show the real distribution of the overall data status in each partition.

5.4 Cancer Types Classification

This part is implemented in PyTorch, which is widely used in both academics and industry. In PyTorch, it uses tensors to store the data (input, outputs, and intermediate features). It can flexibly implement the neural networks and define the forward and backward process, where the forward process is to calculate the probability of the sample being a specific cancer type, and the backward propagation process will minimize the error between the predicted label and true label by performing gradient descent which in our case is performed by Adam optimizer

In this project, we mainly deployed a customized softmax regression model and its forward process in PyTorch and used its built-in Mean-Squared Error (MSE) loss function and backward process for updating the parameters of the model. note here, the reason why the mean-square error is utilized as the loss function in our case is because

1. We deployed the cancer label in one hot encoding format, therefore, the deployment of mean-squared error could minimize the loss to an adequate limit.

2. It could obtain better performance than the cross-entropy loss after several experiments conducted.

```
1  import torch
2  from torch import nn
3
4  class SoftMaxBPNet(nn.Module):
5      def __init__(self, feature_num, class_num):
6          super(SoftMaxBPNet, self).__init__()
7          self.feature_num = feature_num   # the number of input features is 49
8          self.cls_num = class_num    # the number of output classes is 32
9          self.layer = nn.Linear(feature_num, class_num) # the linear layer
10     def forward(self, x): # the forward action is performed by softmax
11         x = self.layer(x) # linear transform
12         x = torch.softmax(x, dim=1) # softmax transform
13         return x
```

Listing 5.3: *The softmax regression model*

The softmax regression model is shown in Listing 5.3. The main part of the softmax regression model is the linear transform module (line 9), with an input vector at length 49(49 of SBS signatures extracted) and an output vector at length 32(32 of cancer types). This linear transform matrix stores the full weights between SBS signatures and cancer types. During the forward process, the input SBS signatures will first be involved into computation and transformed via the linear layer (line 11) and then fed into a softmax activation function (non-linear transform module at line 12).

```
# load training data set and validation data set as well as cleaning data
for i in range(5): # make the 5-fold cross validation here
    # sbs num : 49,cancer types num : 32
    model = SoftMaxBPNet(49, 32)
    criterion = nn.MSELoss()
    optimizer = torch.optim.Adam(model.parameters(), lr=cfg.LEARNING_RATE)
    train_x, train_y, test_x, test_y = get_data(train_dataset, i)
    valid_x, valid_y = process_data(valid_dataset)
    for j in range(epochs): train for 1000 epochs
        model.train()
        # batch size is 16
        for input, target in train_x, train_y:
            y_pred = model(input)
            loss = criterion(y_pred, target)
            optimizer.zero_grad()
            loss.backward()
            optimizer.step()
        model.eval()
        y_pred = model(test_x)
        evaluate(y_pred, test_y) # ROC, precision. recall
    y_pred = model(valid_x) # scoring on the validation data set
    evaluate(y_pred, valid_y)  # ROC, precision. recall
    weight = model.layer.weight.detach().numpy()
    # save the weights
```

Listing 5.4: *The training process of cancer-types classification*

Listing 5.4 shows the implementation of the training stage. It divide the training data into 5 folds,shown in (line 4), as discussed in the detailed designs. The 5-folds cross validation is done on the softmax regression model by training (line 10-17), testing (line 18-20), and validating (line 21-22). Moreover, the weights of these trained models are stored on disk (line 24). During the training, the Adam optimizer is adopted as it is computationally efficient and robust to deal with noisy data.

5.5 Gene Mutation Status Prediction

This part mainly deploys tensorflow to implement the convolutional neural network. Tensorflow is one of the most widely used library in machine learning field. It provides a robust ecosystem, which supports many programming languages and platforms. However, considered the constructing of models and programming in tensorflow is time consuming for the beginners to create machine learning models.This project finally chooses Keras, a high-level machine learning programming API,built on top of tensorflow. The main advantage of Keras is that it provides a generally easy-to-use interface and thorough documented API. Listing 5.5 and Listing 5.6 shows how to design two CNNs in Keras.

```
import tensorflow as tf
import keras.backend as K
from keras.models import Sequential
from keras.layers import Dense, Activation, Flatten, Dropout, Conv1D

# num_features : 49 SBS features
def simple_cnn_model(num_features):
    simple_model = Sequential()
    # first conv layer
    simple_model.add(Conv1D(filters=8, kernel_size=3, padding='SAME', input_shape
    =(num_features, 1)))
    simple_model.add(Activation('tanh'))
    # second conv layer
```

```
13      simple_model.add(Conv1D(16, kernel_size=3, strides=1, padding='same'))
14      simple_model.add(Flatten())  # flatten layer
15      simple_model.add(Activation('tanh'))
16      simple_model.add(Dropout(rate=0.5)) # dropout layer
17      simple_model.add(Dense(2)) # fully connected layer
18      simple_model.add(Dense(1, kernel_initializer='normal', activation='sigmoid'))
19      return simple_model
```

Listing 5.5: *The simple CNN model*

```
1  import tensorflow as tf
2  import keras.backend as K
3  from keras.models import Sequential
4  from keras.layers import Dense, Activation, Flatten, Dropout, Conv1D
5
6  # num_features : 49 SBS features
7  def complex_cnn_model(num_features):
8      complex_model = Sequential()
9      # first conv layer
10     complex_model.add(Conv1D(8, kernel_size=3, padding='same', input_shape=(
       num_features, 1)))
11     complex_model.add(Activation('tanh'))
12     complex_model.add(Dense(2))
13     # second conv layer
14     complex_model.add(Conv1D(8, kernel_size=2, strides=1, padding='same'))
15     # third conv layer
16     complex_model.add(Conv1D(16, kernel_size=2, strides=1, padding='same'))
17     complex_model.add(Activation('tanh'))
18     complex_model.add(Dense(2))
19     # fourth conv layer
20     complex_model.add(Conv1D(16, kernel_size=2, strides=1, padding='same'))
21     complex_model.add(Activation('tanh'))
22     # fifth conv layer
23     complex_model.add(Conv1D(32, kernel_size=2, strides=1, padding='same'))
24     complex_model.add(Activation('tanh'))
25     complex_model.add(Flatten()) # flatten layer
26     complex_model.add(Activation('tanh'))
27     complex_model.add(Dropout(0.5))
28     complex_model.add(Dense(2))
29     complex_model.add(Dense(1, kernel_initializer='normal', activation='sigmoid'))
30     return complex_model
```

Listing 5.6: *The complex CNN model*

Listing 5.5 shows our proposed simple CNN structures. It first create a sequential model (line 8), which is a a plain stack of layers. Each layer has one input tensor and one output tensor. In the following, it respectively adds **Conv1D** layer, **tanh** active function layer, **flatten** layer, **dropout** layer, **dense** layer, and so on. The **Conv1D** layer is the most important structure in CNN and it is exploited to focus on extract a local area of feature data every time it is implemented. Compared with full connected layer, it could automatically learn the important local feature patterns and match it with the labels. Furthermore, it can reduce the computation overhead by sharing weights among different local patterns. The **tanh** active function layer is used to provide non-linear transformation, which has similar function to **sigmoid**. The **flatten** layer is used to convert the output of previous convolutional layers to a long 1-D data array. This data array will be used as the input of the fully connected layers. Therefore, the **flatten** layer is usually used just before the **dense** layers. To avoid over-fitting during the training process, the **dropout** layer is usually appended to deactivate some neural units in the CNN. Lastly, some **dense** layer, i.e., the fully connected layers, is appended as the end of the CNN model to actually perform

the classification function. The complex neural network (see in Listing 5.6) shows very similar structure to the simple one.But, adding more Convolutional layer to assist with finding the accurate local features and more dense layer to reduce the over-fitting that could be brought by the convolution process.

```
# define the focal loss to help with class imbalance problem
def focal_loss(gamma, alpha):
    def focal_loss_fixed(y_true, y_pred):
        pt_1 = tf.where(tf.equal(y_true, 1), y_pred, tf.ones_like(y_pred))
        pt_0 = tf.where(tf.equal(y_true, 0), y_pred, tf.zeros_like(y_pred))
        return -K.mean(alpha * K.pow(1. - pt_1, gamma) * K.log(pt_1)) - K.mean(
            (1 - alpha) * K.pow(pt_0, gamma) * K.log(1. - pt_0))
    return focal_loss_fixed
# load training data set and validation data set
gene_prob = pd.read_csv('gene_prob.csv') # load the gene occurrence
# make the 5-fold cross validation here
for i in range(5):
    # weights : extracted the top 10 sbs signatures's weight from previous
    functions
    train_x, train_y, test_x, test_y = get_data(train_dataset, weights)
    valid_x, valid_y = process_data(valid_dataset, weights)
    # set up optimizer
    adam = Adam(lr=0.00098, beta_1=0.9, beta_2=0.999, epsilon=1e-09)
    # complie the model (simple CNN or complex CNN)
    model.compile(loss=focal_loss(gamma=2., alpha=0.1), optimizer=adam, metrics=['
    accuracy'])
    # train the model
    history = model.fit(train_x, train_y, epochs=200, batch_size=1280)
    # test on testing data
    accuracy_test = score(model, test_x, test_y)
    # test on validate data
    valid_y_pred = model.predict(valid_x)
```

Listing 5.7: *The training process of gene mutation status prediction*

Listing 5.7 implements the training, testing and validation of the CNN models for the gene mutation status prediction. The processing is very similar to that of process in Listing 5.4. The CNN model is fitted for 200 epochs (line 21) which is enough for the convergence to be happened on the training data set ,then, the testing and the evaluation is made on test dataset and validate dataset respectively. The batch size is increased to 1280 for reducing the training time. The strategy is to continue to use the Adam optimizer when updating the CNN model, while changes the loss function to (**focal_loss**) (line 2-8).

The implementation of the binary focal loss is done to assigning more weights to hard or easily misclassified samples to deal with the severe imbalance class problem that has been mentioned in the formal problem analysis, In the function , the weighted loss is adopted to minimize the inaccurate loss that could be brought by the training of disproportionate labels of gene mutation status.

5.6 Chapter Summary

In this chapter, the environment setup, system, and installed software are clearly clarified. four following sections respectively describe how to implement data preprocessing, stratified sampling ,cancer-types classification, and gene mutation status prediction as well as provided the solution of addressing the imbalanced gene mutation status label by implementing the robust focal loss function. In the next section, the experimental results are reported, evaluated and discussed.

6 Evaluation And Discussion

This chapter mainly reports and encapsulate the results of our experiments. Firstly, the accuracy of cancer-types classification is outstanding, except for a few cases. It is because that the key SBS signatures among different cancers varies a lot, and the sample quantities in some sample's related cancers that are limited might be suggested for the few inferior exceptions. Secondly, gene mutation status prediction results are compared between two CNN models constructed in the previous chapter. The comparison results and both models' performance are provided to analyse the feasibility of conducting future research based on the work. Finally, we also discovered that the relationship between SBS signatures and cancers proves to be explainable in the pathological phenomenons.

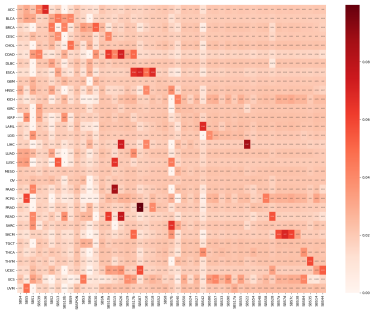
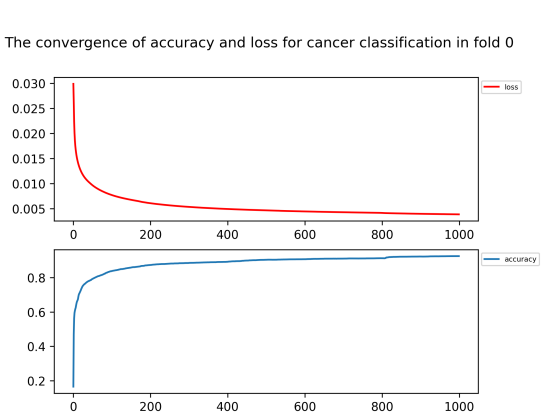


Figure 6.1: *The visualization of SBS signatures of 32 cancer types.*

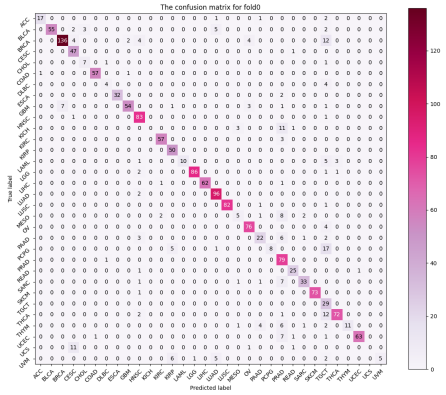
6.1 Cancer-types Classification

At the beginning,the sample size of each of the classes in the validation data set is visualized in Figure 6.4a to help with later analysis. In Figure 6.1, a heatmap is used to visualize the derived relationship between SBS signatures of each cancer types. As shown in this figure, most cancers can be easily identified by its own SBS signatures, e.g., COAD, ESCA, LIHC, PRAD, and SARC. On the other side, a few cancers also seems to be signified by the same type of SBS signatures, such as MESO and OV.

In each of the 5 fold evaluations, we obtained classification accuracy of 0.85,0.87,0.87,0.87,0.87 on each validation dataset, which is generally a stable prediction. Figure 6.2a show the update of accuracy and loss during the training process. The model finally converge at 85% accuracy and the error has been limited considerably. Due the the space limit, only one fold (first fold, also the worst performed fold) evaluations is displayed here to exam the performance of the model and recommend improvements. In Figure 6.2b The displayed outcome show the confusion matrix on the samples of all tumour types.Generally speaking, a obvious diagonal line is shown on the matrix, which identifies most samples have been accurately classified into their true classes. The confusion matrix on other data folds has also shown the similar pattern with that is shown on

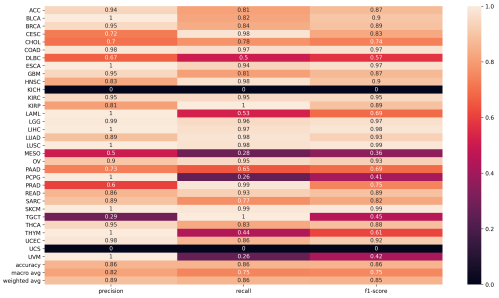


(a) Accuracy and loss curve.

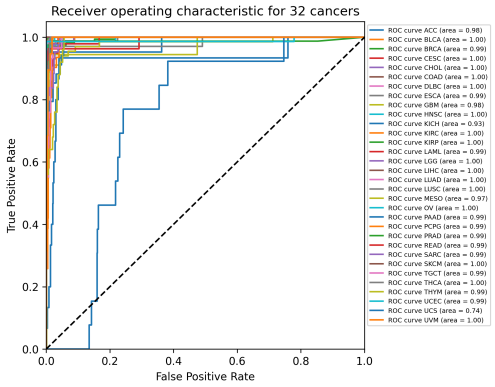


(b) The confusion matrix on the samples of the first fold.

Figure 6.2: *The cancer classification result in fold 0.*



(a) Precision, recall, and f1-score on the samples of the 0-th fold.



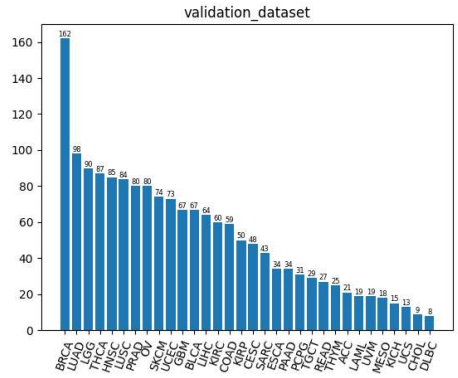
(b) The roc score on the samples of the 0-th fold.

Figure 6.3: *The cancer classification result in fold 0.*

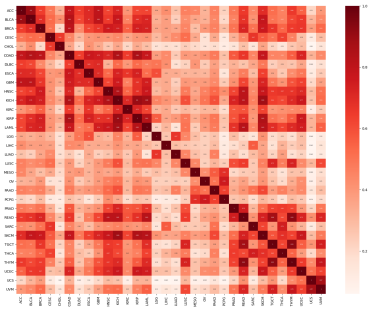
Figure 6.2b with some minor differences (see in Appendices). Although most samples has been accurately identified, there are still some visible mistakes, such the samples of BRCA which have got 12 out of 162 of the samples mis-classified to TGCT, the KICH,which has got 11 of 15 of its samples classified to PRAD, PCPG has also obtained bad prediction as it has missed 17 out of 31 of its samples to TGCT and UCS, the type of cancer class that have gained no accurate prediction and have most of its samples(11 out of 13 them) been classified to CESC.

In Figure 6.3a, For achieving the comprehensive evaluation, further visualization such as the score of precision, recall, and f1-measurement score on the samples of the 0-th fold are also provided as performance measurement metrics. The block filled with dark colour shows the failure cases, and the lighter coloured blocks are the cancer types that are accurately classified. As it could be observed from the graph, KICH and UCS, which are the two classes identified in the confusion matrix that having the severe misclassification problem, are showing the precision, recall and f1-score of 0.

Another inferior prediction made on classes such as TGCT, which has gained 0.29 in precision but 1 in the recall is identified. Moreover, the cancer type termed MESO (Mesothelioma) has received the partially wrong prediction result as it only displayed 0.5 in precision and 0.28 in the recall. As mentioned above, the PCPG class has also demonstrated a negative result as it only



(a) The sample distributions of the validation data set.



(b) The similarity of SBS signatures between cancer types.

Figure 6.4: *Two causes of cancer-types classification misses.*

reached a recall of 0.26, which has validated the formal analysis that many samples from other classes are misclassified to this specific class. considerably,Figure 6.3b shows the ROC score and the area under the curve for each of the cancer classes by taking the binary comparison with itself and the rest of the classes,It is clear to see that most of the classes has received the perfect classification as they have obtained AUC as 1.0,In this scenario, the only one lesser performed cases across all cancer prediction is the class of UCS (Uterine Carcinosarcoma) which has illustrated the inferior results in each of the above evaluation metrics and still gained the AUC as 0.74.

To explain the poor prediction of some cases.There are two main possible reasons to be deducted from both statistical respective and pathological respective:

(1) One of the possible reason has been elicited is that some cancer types has a small-scale number of samples and therefore caused the imbalanced learning of the model and caused the mis-prediction). (2) Another related reason from the aetiological respective is that The SBS signatures exposure of samples from two cancer types are likely to be very similar.

For validating the first presumption, the Figure 6.4a is provided to shows the data distribution in the validation data set. The examples that could be used to prove the theory would be like the prediction for the UCS and MESO classes which are only occupying 13 and 18 of samples across overall validation data samples(total of 1673 samples), whereas BRCA has up to 163 samples and has demonstrated preferable results over almost all of the evaluation metrics.

other classes that displayed the bad classification results such as TGCT and KICH are also appearing to be in lack of the sufficient amount of samples as the formal has 29 samples and the later one only has 15 samples in account.

For validating the second possible hypothesis, which is that the reason why there are misclassifications for some of the cancer cases is that some cancer types share the most common SBS signatures, a similarity comparison between different cancer classes based on SBS signatures' weights in each cancer is adopted to help with observation and analysis.

As it could be identified in Figure 6.4b , the previous finding discussed in confusion matrix evaluation could be addressed, the reason why TGCT class has recieved some of the samples misclassified from BRCA class is that they have achieved similarity of 77.49% by taking the cosine similarity between their related SBS signatures. also, by observing the heatmap between the cancer classes and all of the 49 SBS signatures, the discovery is made that both classes are signified by the SBS signature termed as SBS 3 which is strongly associated with germline and somatic BRCA1 and BRCA2 mutations and BRCA1 promoter methylation.

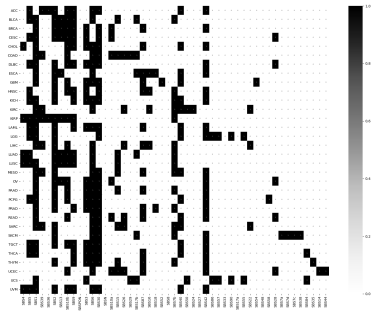


Figure 6.5: *The visualization of top 10 masked significant SBS signatures in 32 cancer types.*

for other cancers such as LUAD has also got 2 of its samples misclassified to the cancer type termed HNSC,the reason is because they are both signified by the signature named SBS 4 , where the suggested aetiology for the specific signature is tobacco smoking that causes multiple cancer types in addition to lung and head and neck.

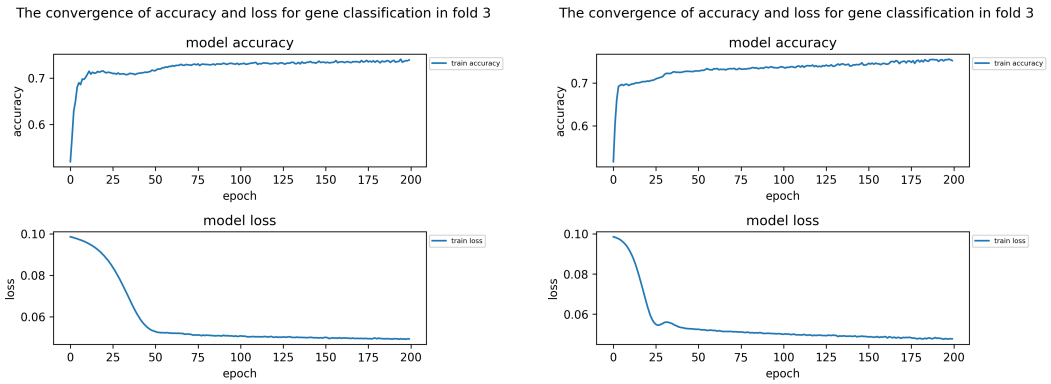
Instead of focusing on the errors that have been made by the classifier, we could also discover the pathological explainable decent classification results from our model. for instance, as we have obtained almost perfect precision, recall in both SARC(sarcoma) and SKCM(Skin Cutaneous Melanoma), with the heatmap between cancer type and the SBS signature provided, we could observe the reasonable prediction made by our model as those two cancers are highly related to SBS-7b which is likely to be caused by the exposure to ultraviolet light. Moreover, 2 of the lung cancer type LUAD,LUSC are shown to be higly related to the SBS4 signature which is explained as tobacco smoking as mentioned before.Other Ultra-violet exposure related SBS signature such as SBS-7a/SBS-7c/SBS-7d are also identified as important factor of causing the Skin Cutaneous Melanoma. In this case, it has provided our accurate basis and encouraged us to continuing working on the prediction of the driver gene in each cancer class based on those explanatory characteristics.

6.2 Gene Mutation Status Prediction

This section focuses on prediction of gene mutation status. The job to be done at initial stage is mainly using the extracted weights of SBS signatures from the cancer-types classification model to find the most relevant signatures for identification of the cancer types. At next stage, the top 10 of the most related SBS signatures discovered during the previous classification of cancer types are used as the mask vector. These SBS signatures are determined by the rank of weights from the cancer-types classification model. With the knowing cancer type, each sample will apply the corresponding mask vector and only leave the values of significant SBS signatures for the prediction of gene mutation status,the remaining 39 features will be settled as the 1% of its original values to form the spatial local feature.

6.2.1 Significant SBS signatures

At this sections , we introduce the significant SBS signatures in each cancer type that can be used to identify the driver gene's gene mutation status within that cancer.Figure 6.5 visualizes the top 10 significant SBS signatures in 32 cancer types. In each cancer type (each row), the black blocks means the SBS signature is important in that cancer type, while the white blocks means the SBS signature could be ignored.



(a) Accuracy and loss curve of the simple CNN. *(b) Accuracy and loss curve of the complex CNN.*

Figure 6.6: *The performance of evaluating two CNNs on masked SBS signatures to predict the gene mutation status.*

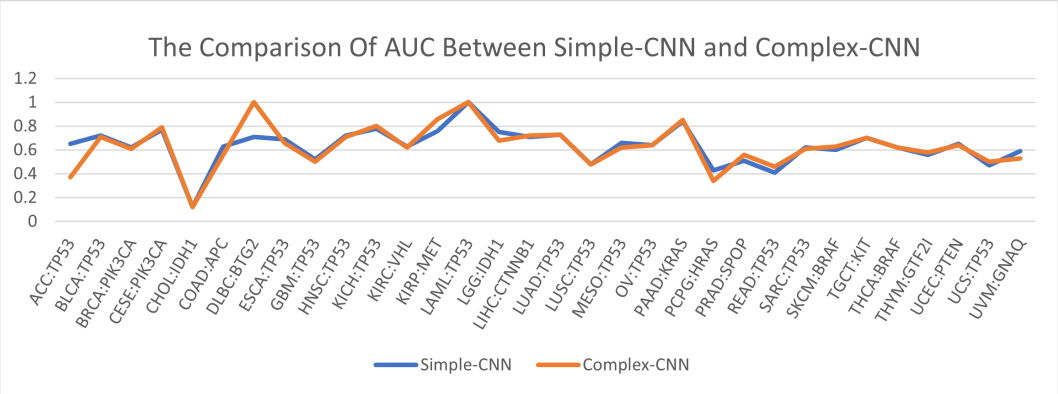


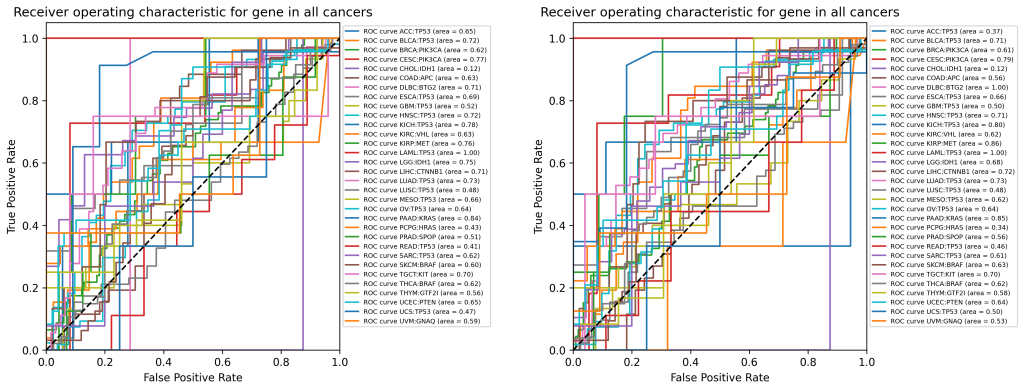
Figure 6.7: *The accuracy comparison of driver genes in 32 cancer types between the simple CNN model and the complex CNN model.*

From this diagram, it can be observed that some SBS signatures are hold by few cancer types, while the others are important in almost any cancer type. For example, SBS-4 is associated with tobacco smoking and, thus it affects LUAD,LUSC (cancers of the lung). It confirms the widely known fact: smoking could cause lung disease. Furthermore, all of SBS-7a/SBS-7c/SBS-7d are likely to be caused by the exposure to ultraviolet light. Therefore, they could be found associated with Skin Cutaneous Melanoma.

On the other hand, some SBS signatures are general important in many cancers. SBS-2 is deep related to activity of the AID/APOBEC family of cytidine deaminases. APOBEC3A/APOBEC3B could be the most likely the cause of the gene mutations in human cancers. Therefore, there are 24 out of 32 cancers has SBS-2 as the top significant SBS signatures. Similar SBS signatures includes SBS-30 (28 cancer types) and SBS-6 (21 cancer types), which are very general signature in tumours and cancers, their advised aetiology such as defective DNA mismatch repair and deficiency in base excision repair are also noticed.

6.2.2 CNN models

With the work around finding the extracted significant SBS signatures in each cancers been made.Two CNN models are trained on the masked SBS signatures and similar to cancer clas-



(a) *The roc score of simple CNN on the samples of the 3-rd fold.*

(b) *The roc score of complex CNN on the samples of the 3-rd fold.*

Figure 6.8: *The roc score of two CNNs on the samples of the 3-th fold.*

sification evaluation, one fold of the assessment is supplied (some other folds results could be seen in Appendices).The result performance is shown in Figure 6.6. with the comparison been made,we found the complex CNN could outperform the simple one by improving the convergence threshold and achieve higher accuracy slightly during the training. The simple CNN can achieve 71.78% accuracy in the weighted accuracy across all labels, while the complex CNN can achieve 71.85% accuracy. besides, The complex CNN model also converged faster, After the third epoch, the overall weighted prediction accuracy of using the simple CNN is 60.1% and that of using the complex CNN is already 70.8%. From these results, the knowledge has been established is that the complicated CNN structure could have higher capability of fitting itself on the same dataset, both accuracy and loss performed well by using complex CNN. In that case,it could be derived that deeper neural network could achieve better results in the future.

Furthermore, the ROC curve is also presented in Figure 6.8. It could be noticed that some gene mutation status are acceptably predicted by using the simple CNN model, such as KIRC:IDH1 (63%), MESO:SPOP (66%), ACC:IDH1 (65%), and so on. Powered by the complex CNN model, these mutation status predictions become better, while some bad cases in the simple model are slightly improved, e.g., CESC:PIK3CA(from 77% to 79%), DLBC:BTG2 (from 71% to 100%), KIRP:TP53 (from 76% to 86%), and so on. However, even under the complex CNN model, a few cases still could not achieve accurate prediction. For example, UCS:TP53 only hit 50% of all its cases (the same result of using a random guess).

In Fig 6.7, it is clear to see that both model could achieve good accuracy, except for a few cases, such as CHOL:IDH1, LUSC:TP53, and USCS:TP53.Also, it is plain to recognize that although there are refinement been made by the complex model compare to the uncomplicated one, it has not shown significant improvement.

There could several reasons for the misclassification of few cases:

(1) The proposed Convolutional neural network is still not complex enough to deal with the complicated relation between the mutational signatures and the top frequently mutated driver gene in each cancer.Therefore with the improvement been identified from above experiment,more advanced Deep learning models and mechanisms (e.g., RMSprop optimizer) could be deployed.

(2) The selected high weight features may not be representative for expressing the complicated mutation process happened in the process of gene mutation.Just because some SBS signatures are common in many cancers, it does not mean the signatures represent important information for the cause of these cancers and the corresponding gene mutation process.

(3) The diverse top K significant SBS signatures should be exploited to adjust the selected Top-K mask vectors to suit the better extraction of the features in different combinations.

(4) The single SBS signatures may not be informative enough in helping with explaining the comprehensive mutation happened in the gene.The fusion of it with other mutational signature including Doublet base substitution (DBS) signature, Small insertions and deletions (ID) Signatures might be able to expressed the whole mutation process happened in the genes completely.

(5) The data distribution is skewed and some cancer types only have a few samples. which might caused the disproportionate learning to our classifier and therefore caused biased study.The proposed reason could be validated as it is clear to see that our classifier has performed better result on those cancers that are sharing more samples. For instance, the complex CNN has achieved AUC of 0.79 in CESC:PIK3CA which shares 48 samples across validation data set, achieved score of 0.71 in HNSC:TP53 which occupied as 5th greatest data set among all classes, achieved AUC score of 0.86,0.73 for KIRP:MET,LUAD:TP53 which are both the large portion of classes in the overall data file, the later is even the cancer class that shared the second largest population (98 out of 1672 in validation data set and 563 out of 9637 samples in all data).

(6) The imbalanced mutation of the driver gene in each sample could also inferred to be one of the main reason why the inferior classification has been made. Although the proposed focal loss function could address the imbalanced label problem,it cannot be generally utilized for total of 32 models at once.The reasoning argument could be elicited from the observation of the mutation frequency in each driver gene in the cancer class in Fig 3.2. For illustration, the mutation frequency of TP53 in HNSC is 0.53 which is considered as balance mutation distribution and it has gained relative higher AUC score of 0.71,whereas the mutation frequency of HRAS in PCPG is 0.1 and therefore the classifier has potentially learned few resource from it and giving the AUC score of 0.34.

(7) The misclassified cancer classes such as UCS and MESO in the previous evaluation performed by the softmax regression model has implied the unpleasant result and the extracted SBS signatures could therefore not be precise enough to be exploited in the later gene mutation status prediction of those cancers.

Nevertheless, with the experiments have been done, the theory of identifying the driver gene mutation status in each cancer through their related mutational signatures is proved, the deep relationship between the driver gene and their corresponding cancer is reassured.The potential possibility of identifying other related gene's mutation status could be established.

6.3 Chapter Summary

In This chapter, we summarize the evaluations of two AI procedures in the project. It firstly presented and analyzed the results of the cancer-types classification model in both computing science perspective and pathological respective. Then, the comparison between performance of the simple convolutional neural network and complex convolutional neural network is presented to facilitate and exam the potential problems and the improvements for future development.

The next chapter will be the final chapter of this dissertation. It concludes the content of the whole project and provides potential future ideas based on the works have been done in this project.

7 Conclusion

7.1 Project Summary

In this project, we initially introduced the background of the cancer genome researched, proposed the motivation of the research which is aimed to understand what is tumour and studied the deep relationship between the genes and cancers to help with formulating the hypothetical solution of reducing the expensive tumour sequencing and provided the theory that the involvement of the mutational signatures could produce a easier and efficient way in the development of machine learning and deep learning approaches in the study of human diseases such as cancer.

In our research, we have validated that with the involvement of the softmax regression method, the cancer-types classification model demonstrated very promisable accuracy to identify the cancer type of a sample based on its SBS signature exposures. It proved the meaningful information which carried by those characteristic signals are immense and prolonging. Moreover, In our model, Given a cancer type, the weight value of the trained model could suggest which SBS signatures could highly impact on this cancer and which of the SBS signatures could be lesser important in identification of cancer type,while it is not generally workable with other models. Furthermore, Based on these valuable information, we deployed the strategy to extract top 10 most weighted SBS signatures that is effective in identification of each cancer and enhanced the concept that those mutational signature is of great usage in explaining the tumour condition in aetiological way and could be potentially usable in the process of the identification of the gene mutation status.

Nevertheless, with the experiments of using different structured convolutional neural network to classifying on the mutation status of the driver gene within each cancer types we collected. We realized the process of involving extracted SBS signatures in gene classification is not as good as it is expected. still, We discovered the acceptable predictive power and even some excellent predictions made to some of the driver gene's mutation status in certain cancer types.Subsequently, Some possible reasons for the inferior classification results of the driver gene were discussed in Section 6.2.2. Finally, with the study and analysis performed in the research, we have validated the deep relationship between cancers and their diver genes through the mutation signature and further facilitated the understanding of the tumours and established the foundation for the later discovery of this specific disease.

7.2 Future Work

Although the results obtained have achieved the expected goals and displayed the extensive predictive power of the SBS signatures and paved way of future research that can be deployed with this molecular characteristic feature, while more improvements and studies could be done based on this project. Therefore, some future works are suggested and recommended:

- With known mutation status of cancer driven genes that can be predicted well by deploying the mutational signatures such as single base substitution signature, a molecular recommendation system could be built to derive the mutation status of unknown genes. With help of this system and the improvement that could be made, doctors could reduce the cost of tumor sequencing to check the whole genome mutation status.
- Although the softmax regression could provides us accurate prediction of the cancer type of the sample given their SBS signatures.there are still portion of misclassifications of the model due to the limited data size.Thus, the suggested work would be gaining more data and deploying different model to observe if another stage of improvement could be made.
- Despite the promising classification result given by the model, there are some special cases which are not considered by our experiments, For illustration, if the sample has more than one cancer type, will the model and the usage of SBS signature still give us acceptable result? will we be able to distinguish the significant SBS signatures that carried the information of specific cancer ? Therefore, many works and research is encouraged to be made to test the feasibility of the approach.
- As shown in the previous chapter, although there are some well distinguished cases predicted by the built model. we have got percentage of the gene mutation status predicted inaccurately. However, with the comparison of the simple and complex models, we realized the increasing of the neural network layers could potentially improve the convergence of the learning as well as ascending the prediction accuracy.Thus, the suggestion is made that more complicated machine learning models could be explored to acquire a better prediction of gene mutation status, such as residual learning structure, attention mechanism, and so on.
- Another thing could be noticed during our research is that we are suffering from the unbalanced data distribution, although several technics such as stratified sampling and the deployment of focal loss function are made,there are still existing certain type of driver genes that are sharing the extreme unbalanced mutation frequency in the cancer class. Accordingly, more data set should be collected and processed reasonably to solve the imbalance issue. As it is always said.The success of machine learning is on top of enough samples.
- Even though the experiments performed has clearly addressed the informative nature of the SBS signature in gene mutation as well as the cancer type classification. Other form of base substitution signature such as Doublet base substitution (DBS) signature, Small insertions and deletions (ID) Signatures as well as the rearrangement signatures should also be considered as important factors in the identification of different type of somatic mutation process, with the prominent results given by the HRDetect model mentioned before, the proposal is made that the combination of those powerful mutational signatures might be the final answer of explaining the mutation status of those driver gene embedded during the process of forming the tumours.

Furthermore, other works should be deployed is to further reconstruct the model and the complicated structure and build a easy-to-use system for other researchers to test their data on. Additionally,while the performance of individual models in the project is gaining the reasonable result, other technics including transfer learning could be applied to fine-tune the machine learning models and determined to achieve better overall performance. Finally, as the model should not be only constrained at only one area of studying. Our systems should also be extended to support different platforms: internet of things, mobile phones, and smart living systems.

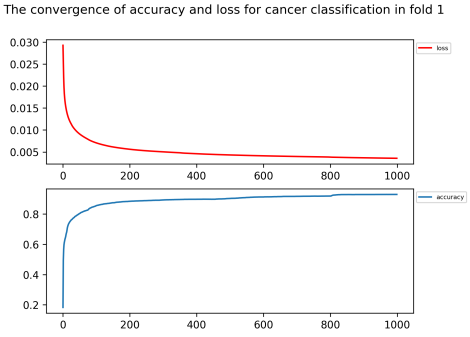
7.3 Final Reflection

At the early stage of this project, many different approaches and strategies are attempted and most of them has not produced the affirmative result. through, After analysis of these results and the researches has been made, many problems have been identified and solved, such as the identification of the local feature patterns, the plans that have been conducted at imbalanced dataset, and so on. At those stages, many knowledge around the biological process as well as the machine learning has been learned and exploited. Therefore, it is proved the meaningfulness of the research. After finished this project, many developing and machine learning skill are mastered to address the challenges and issues proposed in Section 1.2,those question are addressed by:

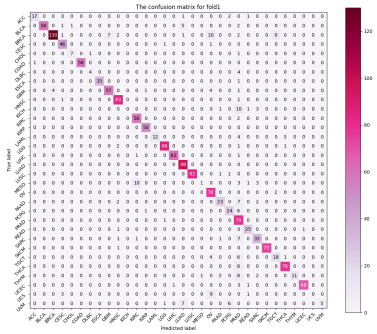
1. **Mastered the skill of using Sigminer,maftools which are the R tools mainly used to studying on the Cancer Genomics, to extract the mutational signature such as SBS signatures from samples.**
2. **learned the skill of using the proportion of data to represent the overall distribution of the data by performing stratified sampling**
3. **Understood the functionality of softmax regression model as well as exploited it to solve the problem of multi-class classfication brought by formal studies and used the extracted weight to identify the significant SBS signatures in each cancer type. This model is aimed to remove the noisy SBS signatures from samples.**
4. **Understood CNN models to learn the spatially local feature patterns and applied it to the processed data to minimize the loss and also established several ways to improve the accuracy, e.g., more complex neural networks and focal_loss.**
5. **Deployed several metrics to evaluate our results, such as K-fold cross validation, precision/recall/f1 scores/ROC and AUC.**
6. **Understood the pathological phenomenon by analyze the prediction result of both cancer and gene mutation status classification.**

A Appendices

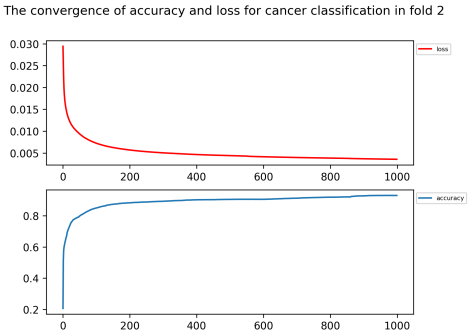
A.1 The classification result in other folds



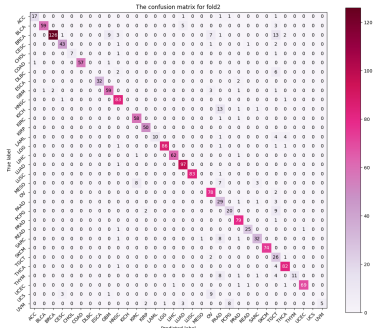
(a) Accuracy and loss curve.



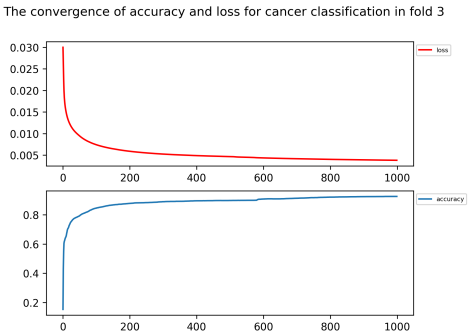
(b) The confusion matrix of cancer classification of the 1-th fold.



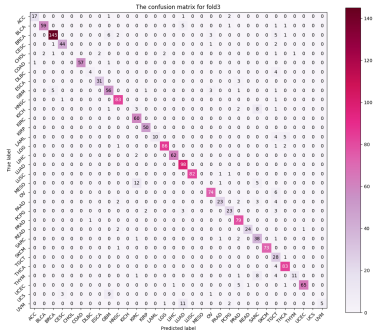
(c) Accuracy and loss curve.



(d) The confusion matrix of cancer classification of the 2-th fold.

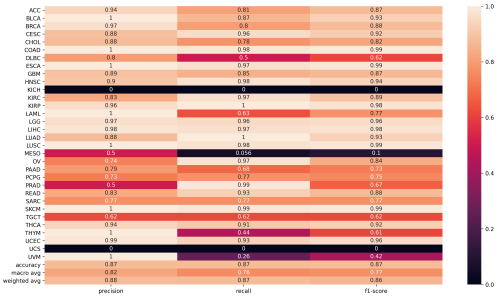


(e) Accuracy and loss curve.

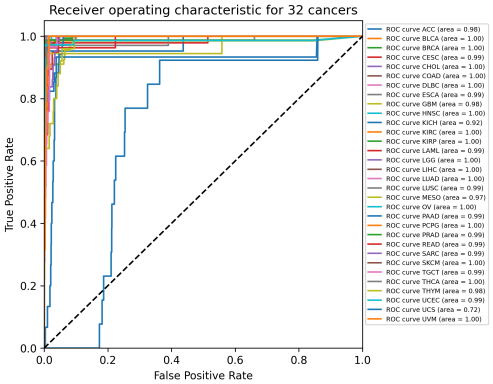


(f) The confusion matrix of cancer classification of the 3-th fold.

Figure A.1: *The classification result of cancer types in other folds.*

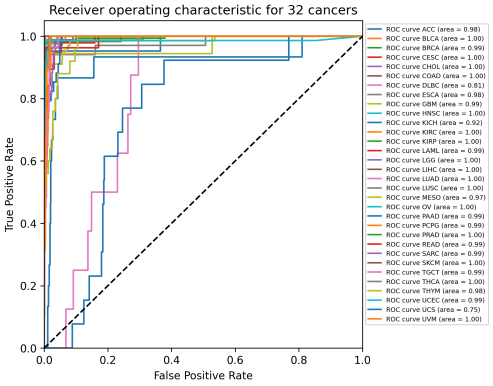


(a) Precision, recall, and f1-score on the samples of 1-th fold.



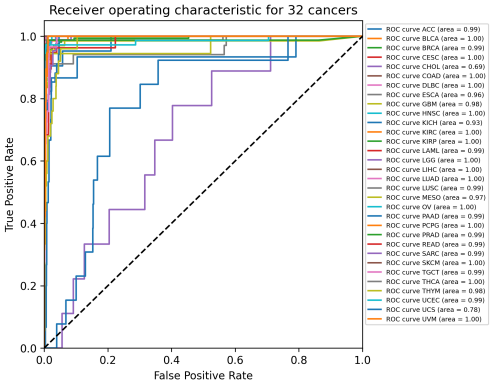
(b) The roc score on the samples of 1-th fold.

(c) Precision, recall, and f1-score on the samples of 2-th fold.



(d) The roc score on the samples of 2-th fold.

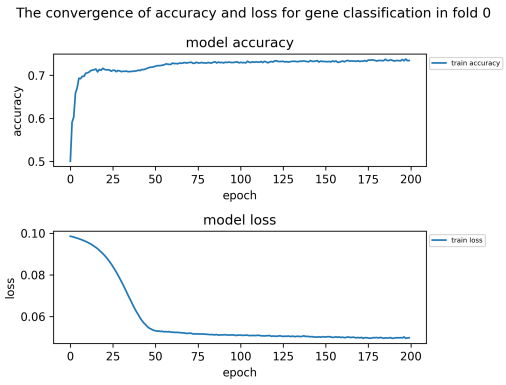
(e) Precision, recall, and f1-score on the samples of 3-th fold.



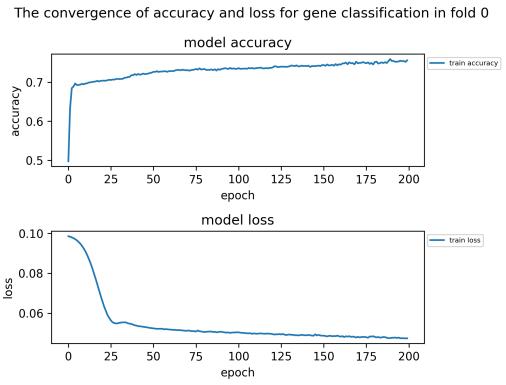
(f) The roc score on the samples of 3-th fold.

Figure A.2: *The classification result of cancer types in other folds.*

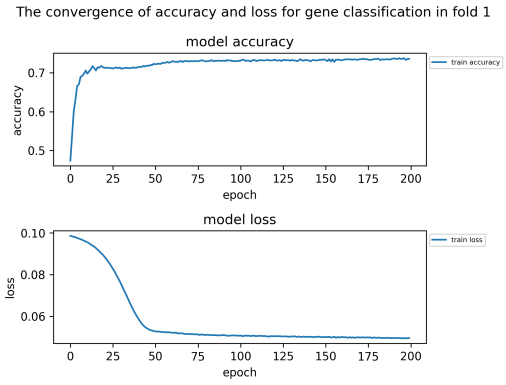
A.2 Gene status prediction on other folds



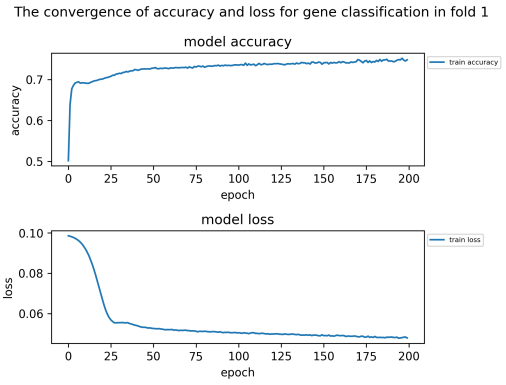
(a) Accuracy and loss curve of the simple CNN.



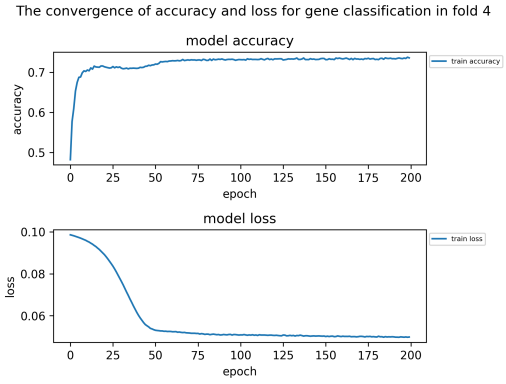
(b) Accuracy and loss curve of the complex CNN.



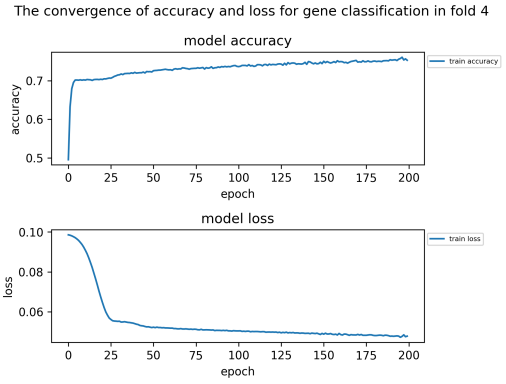
(c) Accuracy and loss curve of the simple CNN.



(d) Accuracy and loss curve of the complex CNN.

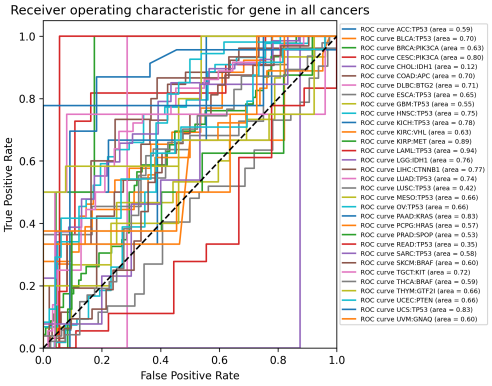


(e) Accuracy and loss curve of the simple CNN.

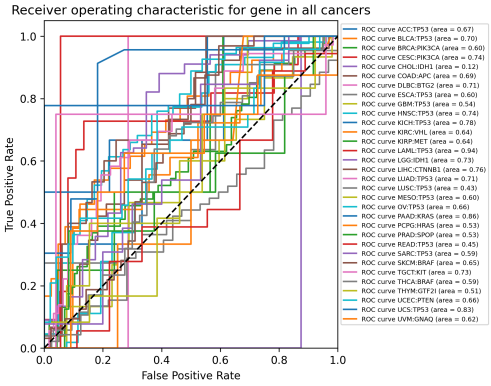


(f) Accuracy and loss curve of the complex CNN.

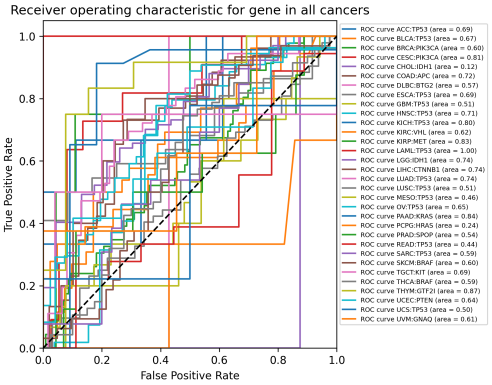
Figure A.3: *The performance of evaluating two CNNs on masked SBS signatures to predict the gene mutation status on other data folds.*



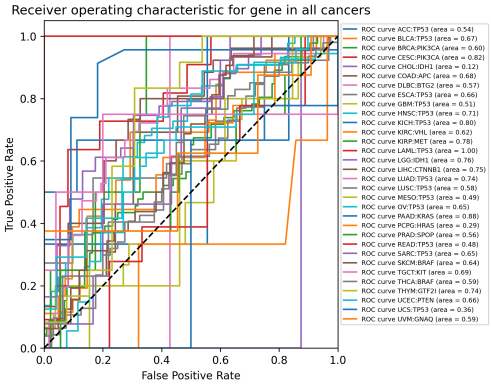
(a) The roc score of simple CNN in 0-th fold.



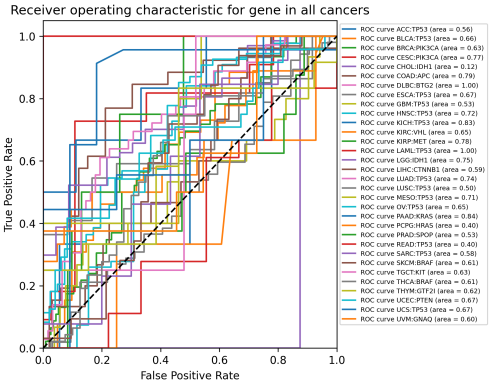
(b) The roc score of complex CNN in 0-th fold.



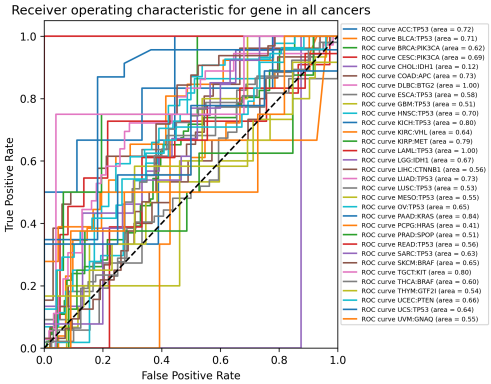
(c) The roc score of simple CNN in 1-th fold.



(d) The roc score of complex CNN in 1-th fold.



(e) The roc score of simple CNN in 4-th fold.



(f) The roc score of complex CNN in 4-th fold.

Figure A.4: *The roc score of simple CNN and complex CNN on the samples of other data folds.*

Bibliography

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N. and Yu, W. (2020), 'The repertoire of mutational signatures in human cancer', *Nature* **578**(7793), 94–101.

Alexandrov, L., Nik-Zainal, S., Wedge, D., Campbell, P. and Stratton, M. (2013), 'Deciphering signatures of mutational processes operative in human cancer', *Cell reports* **3**, 246–259.

Anaconda, I. (2021), 'Conda', `https://docs.conda.io/en/latest/`.

Ayyad, S., Saleh, A. and Labib, M. (2019), 'Gene expression cancer classification using modified k-nearest neighbors technique', *Biosystems* **176**.

Bailey, M., Tokheim, C., Porta, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M., Kim, J., Reardon, B., Ng, P. K. s., Jeong, K., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L. and Mariamidze, A. (2018), 'Comprehensive characterization of cancer driver genes and mutations', *Cell* **173**, 371–385.e18.

Bray, F., Jemal, A., Grey, N., Ferlay, J. and Forman, D. (2012), 'Global cancer transitions according to the human development index (2008-2030): a population-based study', *Lancet Oncol* **13**, 790–801.

Davies, H., Glodzik, D., Morganella, S., Yates, L. R. and Nik-Zainal, S. (2017), 'Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures', *Nature Medicine* **23**(4), 517.

Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D., Lander, E., Van Allen, E. and Sunyaev, S. (2020), 'Identification of cancer driver genes based on nucleotide context', *Nature Genetics* **52**, 1–11.

Ghoneim, A., Muhammad, G. and Hossain, M. S. (2019), 'Cervical cancer classification using convolutional neural networks and extreme learning machines', *Future Generation Computer Systems* **102**.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep Learning*, MIT Press. `http://www.deeplearningbook.org`.

Hoadley, K., Yau, C., Hinoue, T., Wolf, D., Lazar, A., Drill, E., Shen, R., Taylor, A., Cherniack, A., Thorsson, V., Akbani, R., Bowlby, R., Wong, C., Wiznerowicz, M., Sánchez-Vega, F., Robertson, G., Schneider, B., Lawrence, M., Noushmehr, H. and Mariamidze, A. (2018), 'Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer', *Cell* **173**, 291–304.e6.

Hoeck, A., Tjoonk, N., Boxtel, R. and Cuppen, E. (2019), 'Portrait of a cancer: Mutational signature analyses for cancer diagnostics', *BMC Cancer* **19**.

Institute, N. C. (2015), 'What is cancer?', `https://www.cancer.gov/about-cancer/understanding/what-is-cancer#:~:text=Cancer%20is%20the%20name%20given,up%20of%20trillions%20of%20cells.`

Institute, N. C. (2021), 'Harmonized cancer datasets', https://portal.gdc.cancer.gov/.

Jung, W. E.-I. (2019), 'What are the differences between chromosomes, chromatids and chromatin?', https://www.quora.com/What-are-the-differences-between-chromosomes-chromatids-and-chromatin.

Kamps, R., Brandão, R., van den Bosch, B. J., Paulussen, A., Xanthoulea, S., Blok, M. J. and Romano, A. (2017), 'Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification', *International Journal of Molecular Sciences* **18**.

Loo, P., Nilsen, G., Nord, S., Vollan, H., Børresen-Dale, A.-L., Kristensen, V. and Lingjærde, O. (2012), 'Analyzing cancer samples with snp arrays', *Methods in molecular biology (Clifton, N.J.)* **802**, 57–72.

Lu, Y. and Han, J. (2003), 'Cancer classification using gene expression data', *Information Systems* **28**, 243–268.

Luo, P., Ding, Y., Lei, X. and Wu, F.-X. (2019), 'deepdriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks', *Frontiers in Genetics* **10**, 13.

Martinez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo Pac, C., Mularoni, L., Pich, O., Bonet Giner, J., Kranas, H., Gonzalez-Perez, A. and López-Bigas, N. (2020), 'A compendium of mutational cancer driver genes', *Nature Reviews Cancer* **20**, 1–18.

Matsutani, T. and Hamada, M. (2020), 'Parallelized latent dirichlet allocation provides a novel interpretability of mutation signatures in cancer genomes', *Genes* **11**(10).
URL: *https://www.mdpi.com/2073-4425/11/10/1127*

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L., Martin, S., Wedge, D., Loo, P., Ju, Y., Smid, M., Brinkman, A., Morganella, S., Aure, M., Lingjærde, O., Langerød, A., Ringnér, M. and Stratton, M. (2016), 'Landscape of somatic mutations in 560 breast cancer whole-genome sequences', *Nature* **534**.

Reis-Filho, J. and Pusztai, L. (2011), 'Breast cancer 2 gene expression profiling in breast cancer: classification, prognostication, and prediction', *Lancet* **378**, 1812–23.

Shixiang, W., Li, H., Song, M., He, Z., Wu, T., Wang, X., Tao, Z., Wu, K. and Liu, X.-S. (2020), 'Copy number signature analyses in prostate cancer reveal distinct etiologies and clinical outcomes'.

Sotiriou, C., Neo, S.-Y., Mcshane, L., Korn, E., Long, P., Jazaeri, A., Martiat, P., Fox, S., Harris, A. and Liu, E. (2003), 'Breast cancer classification and prognosis based on gene expression profiles from a population-based study', *Proceedings of the National Academy of Sciences of the United States of America* **100**, 10393–8.

Stewart BW, W. C. (2014), 'World cancer report 2014', https://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2014.

Tan, A. C. and Gilbert, D. (2003), 'Ensemble machine learning on gene expression data for cancer classification', *Applied bioinformatics* **2**, S75–83.

Vineis, P. and Wild, C. (2013), 'Global cancer patterns: Causes and prevention', *Lancet* **383**.

wiki (2021), 'Mutational signatures', https://en.wikipedia.org/wiki/Mutational_signatures.