3-10-2021

A Deep Learning Approach for Multi-Omics Data Integration to Diagnose Early-Onset Colorectal Cancer

Noor Kammonah
University of Windsor

A Deep Learning Approach for Multi-Omics Data Integration to Diagnose Early-Onset Colorectal Cancer

By

Noor Kammonah

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2021

A Deep Learning Approach for Multi-Omics Data Integration to Diagnose Early-Onset Colorectal Cancer

by

Noor Kammonah

APPROVED BY:

M. Hlynka
Deptartment of Mathematics and Statistics

S. Samet
School of Computer Science

L. Rueda, Co-Advisor
School of Computer Science

A. Alkhateeb, Co-Advisor
School of Computer Science

February 11, 2021

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyones copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Colorectal cancer is one of the most common cancers and is a leading cause of death worldwide. It starts in the colon or the rectum, and they are often grouped together because they have many features in common. It has been noticed that colorectal cancer attacks young-onset patients who are less than 50 years of age in increasing rates lately. Rapid developments in omics technologies have led them to be highly regarded in the field of biomedical research for the early detection of cancer. Omics data revealed how different molecules and clinical features work together in the disease progression. However, Omics data sources are variants in nature and require careful preprocessing to be integrated. A convolutional neural network is a class of deep neural networks, commonly applied to analyze visual imagery. In this thesis, we propose a model that converts one-dimensional vectors of omics into RGB images to be integrated into the hidden layers of the convolutional neural network. The prediction model will allow all different omics to contribute to the decision making based on extracting the hidden interactions among these omics. These subsets of interacted omics can serve as potential biomarkers for young-onset colorectal cancer.

Keywords: Colorectal Cancer, Deep Learning, Omics, Convolutional Neural Network.

DEDICATION

يا أَيُّهَا الَّذينَ آمَنوا كُلوا مِن طَيِّباتِ ما رَزَقناكُم وَاشكُروا لِلَّهِ إِن كُنتُم إِيّاهُ تَعبُدونَ ﴿١٧٢﴾

O believers, eat what is good of the food We have given you, and be grateful to God, if indeed you are obedient to Him. (172)

وَاللَّهُ أَخرَجَكُم مِن بُطونِ أُمَّهاتِكُم لا تَعلَمونَ شَيئًا وَجَعَلَ لَكُمُ السَّمعَ وَالأَبصارَ وَالأَفئِدَةَ ۙ لَعَلَّكُم تَشكُرونَ ﴿٧٨﴾

God produced you from your mothers' wombs knowing nothing, but gave you ears and eyes and hearts so that you may be grateful. (78)

First & foremost, with my utmost honor, I would love to humbly dedicate this thesis to ALLAH, the most merciful the most compassionate. Only you, my lord, know what I've been through & only you helped me go through it.

Then and with great honor, I would love to dedicate this thesis to the unforgettable and rare beautiful souls I met along this journey that is called LIFE.
From the bottom of my heart, THANK YOU.

ACKNOWLEDGMENTS

I would also like to express my gratitude to my supervisor Dr. Luis Rueda for his support and professional guidance during my graduate studies. Special thanks to my co-supervisor Dr. Abedalrhman Alkhateeb for his continuous advice and for his great leadership.

I would like to thank my thesis committee members Dr. Samet and Dr. Hlynka, for taking the time to review my thesis and for attending my thesis proposal and defense sessions. Their valuable suggestions were of great importance.

Also, I would like to thank my colleague Musab Mushtaque Naik for his constant help and support whenever I needed them, my colloeague Alexandru Filip for answering my questions whenever I had any and our school's office coordinator, Ms. Margaret Garabon Cookson, for her help during my years of study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

The hot point in the medical research sector is cancer with a high economic and social burden. Some remarkable achievements have been made; the precise mechanisms of tumor initiation and growth, however, remain unclear. Cancer is a complex disease of the entire body that includes many defects in the DNA, RNA, protein, metabolite and medical imaging levels. Biological omics, including genomics, transcriptomics, proteomics, metabolomics and radiomics, attempt to systematically explain carcinogenesis at various biological levels, driving the shift of the paradigm of cancer research from a single parameter model to a systemic multi-parameter model (M. Lu & Zhan, 2018). How we view the cancer genome has been reshaped by the incredible scientific breakthroughs of the last decade; thus, our approach to the translation of this information must also be.

Cancer genomics refers to the analysis of tumor genomes using different profiling techniques, including (but not limited to) DNA methylation, copy number, transcriptome and whole-genome sequencing, technologies that can be described as omics collectively (Vucic et al., 2012). How we view the cancer genome has been reshaped by the incredible scientific breakthroughs of the last decade; thus, our approach to the translation of this information must also be. For patients, each type of omics data reflects a single view and it is difficult to use only a single omic to obtain precise prediction. Many experiments have sequenced several forms of omics from the same patient in order to obtain a holistic view of patients in genomics. Using the Cancer Genome Atlas (TCGA), which offers more than ten thousand samples of 33 cancer types, systematic studies have been carried out (Tomczak et al., 2015).

The useful data makes an integrated analysis for systematic cancer prognosis analysis based on multiple omics data (Chai et al., 2019). We consider the problem of classifying cancer patients in this thesis; framed as a image-based supervised classification. Our research aims to work on integrating images from multiomics data to classify cancer patients.

We present a novel supervised model that consists of three convolutional neural networks to handle multiple inputs (omics). The output of those three convolutional neural networks will be discriminant features (from each input/omic) where they will be flattened then fed into the last CNN to predict the young-onset colorectal cancer patients.

In this chapter, we first discuss cancer and one of its types, colorectal cancer. We then describe mutliomics, and their use in research, and biomarkers. In later sections of this chapter, we will describe the problem at hand, followed by thesis objective & organization.

1.1 Cancer

Cancer is a global epidemic. It is the first or second leading cause of death before the age of 70 in ninety-one countries as of 2015, and is predicted to be the "leading cause of death in the 21st century in all countries of the world" (Bray et al., 2018). Over the past half century, epidemiology has not only helped researchers to ferret out many of the non-inherited environmental causes of cancer, but also to estimate how many annual cancer deaths can be attributed to each person. Although the research can not be used to predict what will happen to any single individual, it nevertheless offers widely useful data for people seeking to reduce their exposure to known cancer-causing agents or carcinogens. It seems that cancer develops from the effects of two distinct forms of carcinogens agents that damage genes involved in the regulation of cell proliferation and migration are part of one of these groups. Cancer occurs when a single cell, usually over several years, accumulates a number of these mutations and eventually escapes from most proliferation restraints. The mutations

cause additional alterations to be produced by the cell and its descendants and to accumulate in increasingly large numbers, forming a tumor composed entirely of these abnormal cells. Another group contains agents that do not damage genes, but instead improve the growth of tumor cells or their precursors selectively. The biggest risk of malignancies is that they can metastasize, causing some cells to spread and thereby infect other areas of the body (Trichopoulos et al., 1996). Among the differences between normal cells and cancerous cells in the human body are:

- Cancerous cells are basically many cells that continue to grow and divide.
- Cancerous cells differ from normal cells in size and shapes.
- Cancerous cells' nuclei are larger and darker than normal cells' nuclei.
- Cancerous cells have abnormal number of chromosomes that are arranged in a disorganized fashion.
- Cancerous cells are shaped like a cluster of cells with no boundary.

1.1.1 Colorectal Cancer

Colorectal cancer (CRC) is one of the world's most common cancers, with between one and two million new cases diagnosed annually, making CRC the third most common cancer and the fourth most common cause of cancer-related death, trailing only lung, liver and stomach cancers with 700,000 deaths each year. "By gender, CRC is the second most common cancer in women (9.2%) and the third in men (10%)" (Stewart & Wild, 2014). From 1990 to 2012, the incidence of CRC increased by over 200,000 new cases per year with 55% of them detected in western countries. But this pattern is shifting due to some of those countries' rapid growth over the past few years (Brody, 2015). Figure 1 shows the four areas of a human colon (Ascending, Traverse, Descending and Sigmoid) and the human rectum with a cross-sectional image of a colon caner taken during colonoscopy.

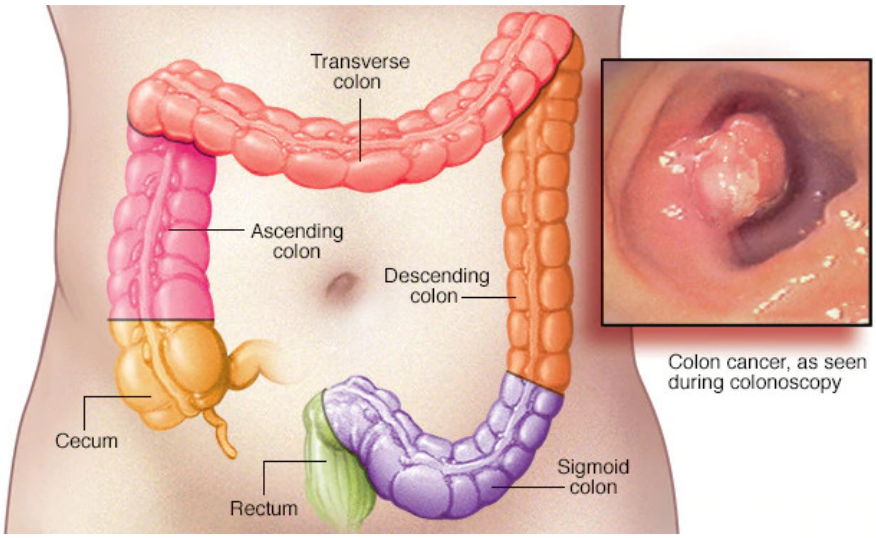

Figure 1: Human colon and rectum.

Colonoscopy is an exam used to diagnose changes or irregularities in the large intestine (colon) and rectum, during which a long and flexible tube (colonoscope) is inserted into the rectum. A small video camera at the tip of the tube helps the doctor to see the inside of the entire colon. Polyps or other forms of irregular tissue may be removed (if possible) during a colonoscopy via the lens. Also, during the same procedure, tissue samples (biopsies) may be taken as well (**?**).

1.2 Multiomics

Multiomics is a modern approach during which the data sets of various omic groups are combined. Genome, proteome, transcriptome, epigenome and microbiome are the various omic strategies employed during multiomics. The addition of "omics" to a molecular term implies a detailed evaluation of a group of molecules. The first discipline of omics to emerge, genomics, concentrated on the study of whole genomes as opposed to "genetics" that challenged human variants or single genes. A very useful mechanism for mapping and analyzing particular genetic variants that lead to both mendelian and complex diseases was provided by genomic studies. The omics field was primarily driven by technological advances that made biological molecules possible for cost-effective, high-throughput analysis (Hasin et al., 2017). The number

of citations of articles containing the term "multiomics", found on Pubmed (a search engine that accesses primarily MEDLINE (a bibliographic database of life sciences and biomedical information) references and abstracts on life sciences and biomedical topics), has gone up by almost 400% from 2010 until the year 2018.

Each type of omics data, on its own, usually provides a list of differences pertaining to the disease. The data can be used as markers of the disease phase and to provide insight into which biological mechanisms or processes vary between the disease and control groups. One major variance between the disease and control groups is that cancer cells (in disease groups) are less specialized than normal cells (in control groups). Meaning the latter mature with specific functions into very distinct cell types, but cancer cells do not. This is one reason why cancer cells begin to divide without stopping. Furthermore, cancer cells may ignore signals that typically tell cells to stop dividing or to start a process known as programmed cell death, or apoptosis, used by the body to get rid of unneeded cells.

Analysis of only one data form, however, is restricted to associations which often represent reactive rather than causative processes. Integration of multiple forms of omics data is also used to elucidate possible causative changes that contribute to illness, or treatments' objectives, which can then be evaluated in further molecular studies (Hasin et al., 2017).

1.3 Biomarkers

A biomarker is defined as a measurable indicator of a biological process. There are various types of cancer biomarkers. Some of which include:

- **Prognostic biomarkers**: those that predict the growth of cancer (Hahn & MacLean, 1955).
- **Diagnostic biomarkers**: those that predict the existence of a disease or condition of interest or a cancer subtype (Milioli et al., 2015).
- **Predictive biomarkers**: those that predict the chances of survival a patient who is being treated with a specific drug (Dao et al., 2011).
- **Progression biomarkers**: those that predict whether the cancer is spreading or not.
- **Recurrence biomarkers**: those that predict whether cancer will recur later in a patient's life (Umetani et al., 2006).

1.4 Problem Statement

Recent research has shown that CRC rates continue to increase in younger Canadians. According to the government of Canada's website, CRC is Canada's second most common cancer, with a man out of 14 and a woman out of 18 to be diagnosed with it in their lifetimes. It has been estimated that 93% of CRCs occur in adults who are 50 years or older. Multiomics data integration for Computer-aided diagnosis (CAD) systems has been applied in cancer research. The main challenge faced by that research is how to incorporate different resources of data, which may have different characteristics, in one prediction model. The advances of deep neural networks helped in integrating the multiomics data sets in different levels of the network layers. What needs to be done is the extraction of the most discriminant features from each omic and omitting the non relevant data, which includes noise and redundant features.

1.5 Research Objective

In this thesis, we construct a multi-input classification system that can classify the young-onset CRC patients from the older patients. The main objectives of this research are:

- Building a CAD system to diagnose young-onset CRC patients with high performance measurements for detecting them. This system can extract discriminant features from each omic then integrate these features in the prediction system using a dedicated convolutional neural network for each omic. The extracted discriminant features will then get flattened, concatenated and fed to a merging convolutional neural network for a collective prediction process.
- Finding a set of biomarkers for young-onset CRC patients. These biomarkers can be clinical or genomic features.

1.6 Thesis Organization

The organization of the thesis is as follows. In Chapter 2, we present background information about the pillar concepts of our project. We also summarize the literature review of some existing works in images classification using deep learning. In Chapter 3, we talk about the data used in this project and we describe our proposed methodology in detail. In Chapter 4, we provide the results we obtained from conducting several experiments on the data sets that we collected from a publicly provided web portal. We also present a comprehensive analysis and some insights of those acquired results. In Chapter 5, we provide our conclusions from this project as well as discuss some future research directions.

CHAPTER 2

Literature Review

This chapter discusses some background information about the main concepts that are considered fundamental in our project. It also discusses previous work that was done in regards to solving the problem stated in the previous chapter.

2.1 Artificial Intelligence

Artificial Intelligence (AI) refers to the intelligence shown by machines and that tries to mimic human's intelligence. Leading AI textbooks refer to AI as the study of intelligent agents; agents refer to any system that perceives its environment and acts in a way that maximizes how it successfully achieves its goals (Poole et al., 1998).

2.2 Machine Learning

Machine Learning (ML) is a branch of AI that aims to find patterns in vast quantities of data. ML algorithms use statistics on data, which can include numbers, words, pictures, clicks, etc. This data can be fed into an ML algorithm if it can be processed digitally. These algorithms are categorized into three main areas which are supervised, unsupervised, and semi-supervised ML algorithms. Supervised ML learning basically happens when an ML algorithm tries to learn a function based on example input-output pairs that maps an input to an output (Russell & Norvig, 2010). A function is inferred, by the ML algorithm, that consists of a group of training examples from the labeled training data (Mohri et al., 2012). In this type of learning, each example

is a pair made up of an input object (typically a vector) and a desired output value. A supervised ML algorithms analyzes the training data then an inferred function is generated, which can be used for mapping new examples. For unseen instances, an optimal scenario would allow the algorithm to correctly determine the class labels. This allows the learning algorithm to reasonably generalize from the training data to unseen circumstances. After implementing an ML algorithm, it needs to be evaluated for efficiency. There are many metrics for evaluating the performance of a ML algorithm, some of these measurements are based on the **confusion matrix**. Figure 2 shows an illustration of a confusion matrix.

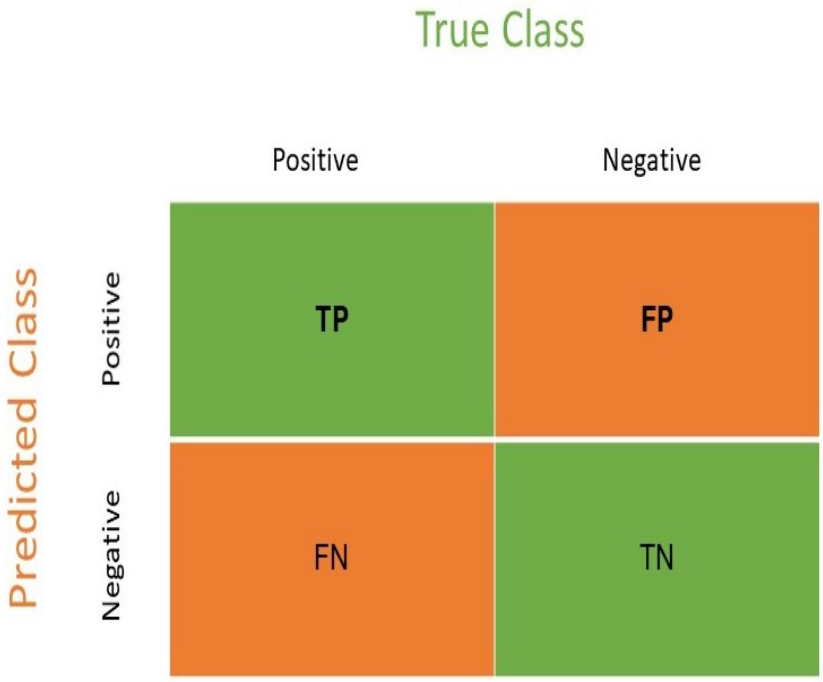

Figure 2: An illustration of a confusion matrix.

A confusion matrix, also know as an error matrix (Stehman, 1997), is a unique table structure that allows for the visualization of a machine learn algorithm, usually a supervised one, where the instances in a predicted class are expressed in each matrix row and each column represents the instances present in an actual class (or vice versa) (Powers, 2008). Some of the measurements used for evaluating ML algoritms'

performances -that are based on the confusion matrix- include accuracy, sensitivity (or recall), and specificity. Unsupervised machine learning is a type of machine learning that looks for previously undetected patterns in a data set that doesn't have preexisting labels and with a minimum of human supervision. This type of machine learning, also known as self-organization, allows for modeling of probability densities over inputs, in contrast to supervised learning that typically uses human-labeled data (Hinton & Sejnowski, 1999). Principal component and cluster analysis are two of the key methods used in unsupervised learning.

In unsupervised learning, cluster analysis is used to group or segment data sets with common attributes to extrapolate algorithmic relationships. Principal components are, basically, the directions of the data that explain a maximal amount of variance, i.e., the lines that capture most information of the data. The correlation between variance and data here is that the greater the variance held by a line, the greater the variance of the data points along it, and the greater the variance along a line, the greater the information it has. In cluster analysis, data that hasn't been labeled will be grouped, classified or categorized. Cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

In semi-supervised learning approaches, during training, a small amount of labeled data is combined with a large amount of unlabeled data. This type of ML falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data). When used in combination with a small amount of labeled data, unlabeled data can yield significant changes in the learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent or a physical experiment. The cost associated with the labeling process may become large. Also, acquiring fully labeled training sets is an infeasible process, while acquiring unlabeled data is relatively inexpensive. So, in such situations, semi-supervised learning can be of great practical value.

Machine learning approaches were proposed to predict the outcomes of cancer. Examples of such approaches inlcude the work of Hamzeh et al. (Hamzeh et al., 2019) who proposed a supervised learning model to predict prostate cancer Gleason score and the work of Abou Tabl et al. (Tabl et al., 2018) who proposed a semi-supervised learning to predict the survivability of breast cancer treatments.

2.2.1 Feature Selection

Feature selection is a mechanism in ML in which a subset of relevant features are selected to be used in constructing a model. Feature selection is used to simplify models so that researchers and analysis can better understand them (James et al., 2013). This process also shortens training time and reinforce generalization by reducing overfitting. The core concept behind using feature selection is that the data includes some features that are either redundant or irrelevant, and can thus be omitted without much information loss (Spiliopoulou et al., 2015). Figure 3 shows an overview of feature selection techniques.

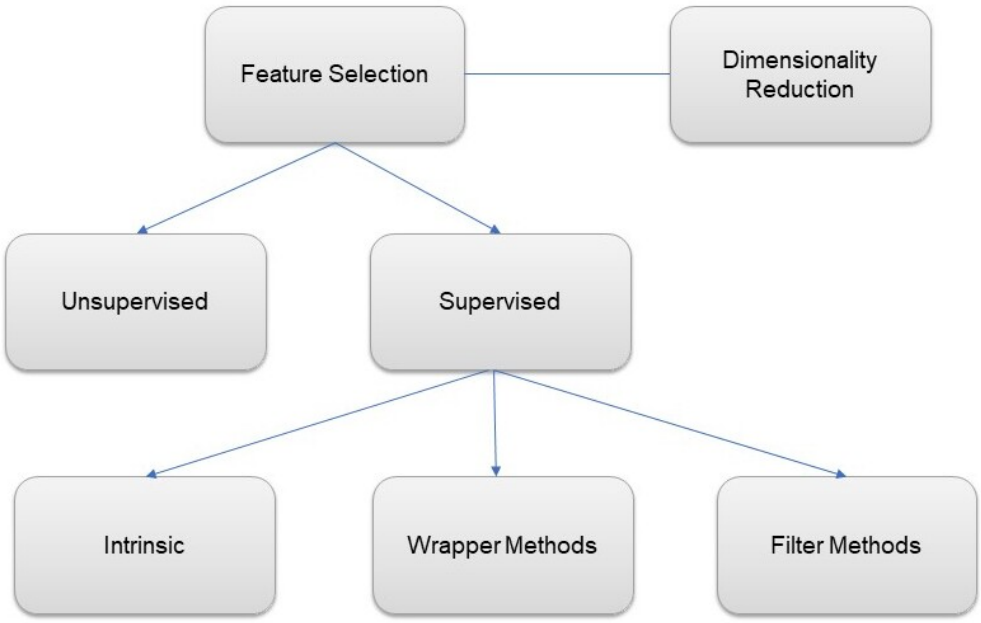

Figure 3: Overview of feature selection techniques.

2.2.2 Overfitting and Underfitting

Overfitting (in statistics) is the production of an analysis which corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably. It refers to a model that models the training data too well. This phenomenon arises when a model learns the detail and noise of the training data so well that that it negatively impacts the performance of the model on new data. This means that the noise or spontaneous variations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not extend to new data; thus, have a negative effect on the ability of the model to generalize.

Underfitting refers to a model that can neither model the training data nor generalize to new data. An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data. Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of overfitting.

2.2.3 Handling Class Imbalance

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to a severe imbalance where there is one example in the minority class for hundreds, thousands, or millions of examples in the majority class or classes. Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models with a poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class.

To handle this problem, there are two main techniques:

- **Upsampling** the majority class using (for example) the SMOTE method. SMOTE (Synthetic Minority Over-Sampling Technique) performs the basic task of basic resampling (creating new data points for the minority class) not by duplicating observations, but by creating new observations along the lines of a randomly chosen point and its nearest neighbors.

- **Downsampling** the majority class, which can be achieved by omitting samples from the majority class and assembling balanced bags of samples from both the majority and minority classes then averaging the results. Example of this work has been done by Elkarami (Elkarami et al., 2016).

2.3 Deep Learning

Deep Learning (DL) is a function of AI that mimics the functions of the human brain in how it processes data and how it builds patterns to use them for the decision-making process. DL is a subset of ML that has networks capable of unsupervised learning from unstructured or unlabeled data. Also known as deep neural learning or deep neural network. DL architectures such as deep neural networks, and Convolutional Neural Networks (CNNs) have been applied to many fields such as bioinformatics, drug design, medical image analysis and board game programs, where they have achieved outcomes comparable to and in some cases exceeding the performance of human experts (Krizhevsky et al., 2012).

2.3.1 Artificial Neural Networks

Artificial Neural Networks (ANNs), mostly referred to as Neural Networks (NNs), are computing systems that were loosely inspired by the animals brains' neural networks (Chen et al., 2019). An ANN is based on a collection of connected nodes called artificial neurons, which vaguely resemble a biological brain's neurons. Each connection between the nodes can transmit a signal to other neurons. An artificial neuron receives a signal, processes it and then signal neurons connected to it. This signal is a real number, and the output of each neuron is computed by a non-linear function of the sum of its inputs. The connections between the neurons are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons can have a threshold such that a signal is sent only if that threshold is crossed by the aggregate signal. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times. Figure 4 shows an illustration of an ANN. Each circle in the figure represents an artificial neuron and each arrow represents a connection from the input of one artificial neuron to the output of another.

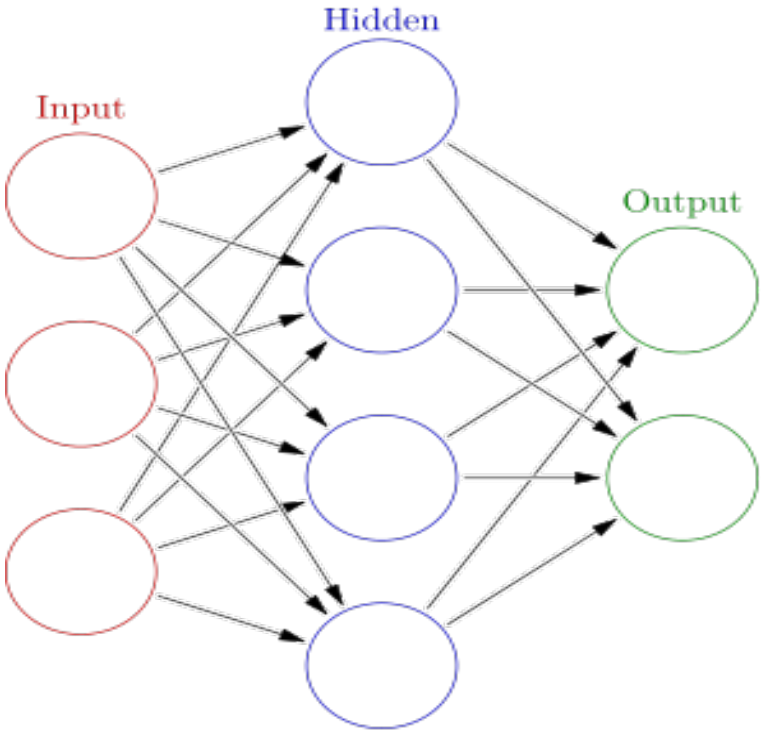

Figure 4: A shallow ANN with one hidden layer.

2.3.2 Self-Organizing Maps

A Self-Organizing Map (SOM) is type of ANN that is trained (using unsupervised learning) to generate a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples called a **map**. SOMs differ from other ANNs in that they apply competitive learning as opposed to error-correction learning which means that only a single node is activated at each iteration in which the features of an instance of the input vector are presented to the neural network. A node is chosen according to the similarity, between the current input values and all the nodes in the grid. The node with the smallest Euclidean difference between the input vector and all nodes is chosen. By going through all the nodes present on the grid, the entire grid will eventually match the complete input data set, with similar nodes grouped together towards one area, and dissimilar ones separated.

2.3.3 Convolutional Neural Networks

The use of Convolutional Neural Networks (CNNs) for image processing was proposed in 1990 (Cun et al., 1990). The computing capacity revolution has promoted the advancement of deep learning; deep learning with a CNN has shown outstanding success in image classification in particular (Krizhevsky et al., 2012). A CNN is a class of deep neural networks that is applied most often to analyze visual imagery (Valueva et al., 2020). CNNs have applications in image and video recognition, image classification, medical image analysis, natural language processing, and financial time series. CNNs are regularized versions of multilayer perceptrons, which means they are fully connected networks, i.e., each neuron in one layer is connected to all neurons in the next layer. The fact that those networks are fully-connected makes their data prone to be overfitted. Typical ways of regularization include adding some form of magnitude measurement of weights to the loss function. CNNs regularize the data by taking advantage of the hierarchical pattern in data and assembling more complex patterns using smaller and simpler patterns. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme. CNNs vary from other pattern recognition algorithms in that they combine feature selection with classification (Hertel et al., 2015). Figure 5 shows an illustration of the basic architecture of a CNN.

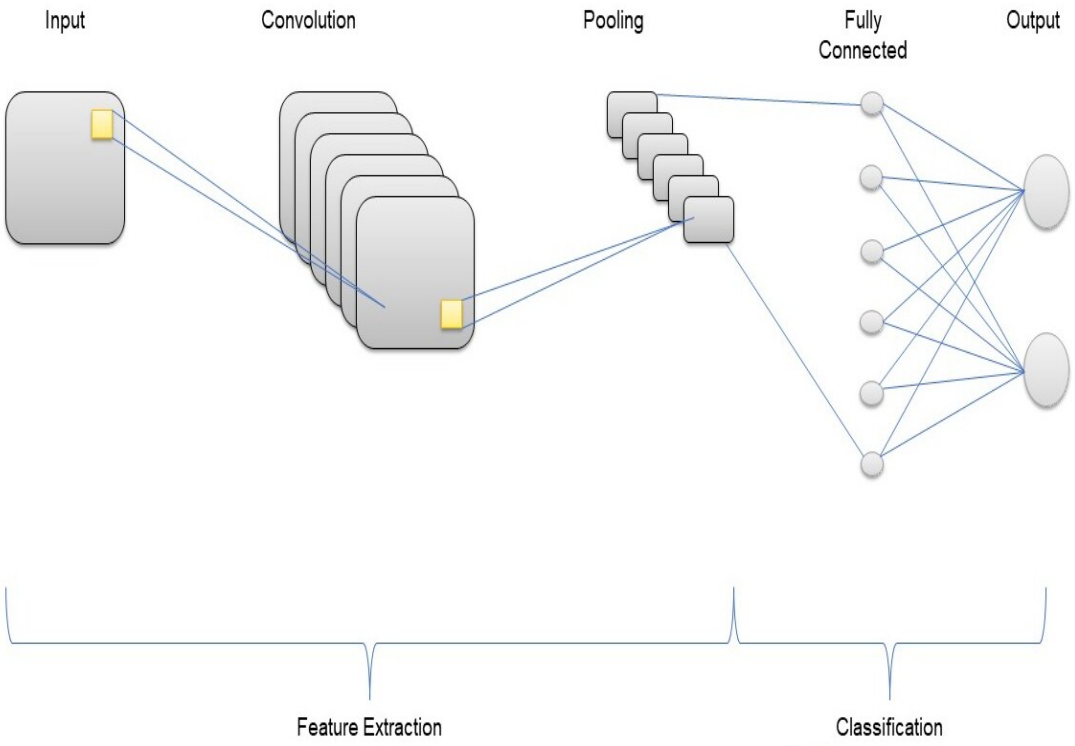

Figure 5: A schematic diagram of a basic convolutional neural network (CNN) architecture.

A CNN consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of a series of convolutional layers that convolve with a multiplication or other dot product. The activation function is usually **ReLU**, and is subsequently followed by additional convolutions such as pooling layers, fully connected layers and normalization layers, referred to as hidden layers because their inputs and outputs are masked by the activation function and final convolution.

CNNs' layers can be summarized as:

- **Input Layer**: the input layer in CNNs contain image data which are represented by a three dimensional matrix. These need to be reshaped into a single column. For example, if we have an image of dimension 28 x 28 (which equals 784), we need to convert it into 784 x 1 before feeding into the input. If we have **m** training examples, then input's dimension will be 784 X m.

- **Convolution Layer**: in this layer (sometimes called feature extractor layer) features of the image get extracted. First, a part of image is connected to this layer to perform convolution operation and to calculate the dot product between receptive field (a local region of the image that has the same size as that of the filter) and the filter. Result of the operation is a single integer of the output volume. Then we slide the filter over the next receptive field of the same input image by a **Stride** and repeat the same operation. We will repeat the same process until we go through the whole image. The output will be the input for the next layer. Convo layer also contains ReLU activation to make all negative value to zero.

- **Pooling Layer**: this layer is used to reduce the spatial volume of input image after convolution. It is used between two convolution layer. **Max pooling** is used to reduce the spatial volume of input image.

- **Fully connected layer**: this layer involves weights, biases, and neurons. It connects neurons in one layer to neurons in another layer.

- **Softmax layer**: this layer performs classification.

- **Output layer**: this layer contains the class label.

2.3.3.1 Activation Functions

In a neural network, inputs are fed into the neurons in the input layer. Each neuron has a weight, and multiplying the input number with the weight gives the output of the neuron, which is transferred to the next layer. The activation function is a mathematical gate in between the input feeding the current neuron and its output going to the next layer. The most common activation functions can be divided into three groups: ridge functions, radial functions and fold functions.

- **Ridge activation functions**: these functions are made up of univariate functions acting on a linear combination of the input variables. Often used examples of these activation functions include: Linear activation, ReLU activation, Heaviside activation, and Logistic activation.

- **Radial activation functions**: these functions are used in RBF networks, which are ANNs with an output that linear combination of radial basis functions of the inputs and neuron parameters. These activation functions can take many forms, but they are usually found as one of the following functions: Gaussian, Multiquadratics, Inverse multiquadratics and Polyharmonic splines.

- **Folding activation functions**: these functions are exhaustively used in the pooling layers in CNNs and in output layers of multi-class classification networks. They perform aggregation over the inputs. In multiclass classification the softmax activation is often used.

2.4 Summary of Previous Work

In this section, we summarize the papers that were studied for conducting this project's literature review.

- In the paper titled: **iSOM-GSN: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps** the authors introduced a systematic, generalized method, called **iSOM-GSN**, used to transform multi-omic data with higher dimensions onto a two-dimensional grid. They applied a CNN to predict disease states of various types and based on the idea of Kohonens SOM, they generated a two-dimensional grid for each sample for a given set of genes that represents a gene similarity network. Their model produced nearly perfect classification accuracy and provided an enhanced scheme for representation learning, visualization, dimensionality reduction and interpretation of multi-omic data (Fatima & Rueda, 2020).

- In the paper titled: **Deep Learning Approach for Breast Cancer InClust 5 Prediction based on Multiomics Data Integration** the authors presented a deep learning model based on multiomics data integration to predict the five-year interval survival of breast cancer InClust 5. Their method was an expansion of the iSOM-GSN model, where they created a feature map for each omic data set instead of only one. The model incorporated the prediction of the three CNNs using an integration layer. Their model was able to learn from all the omic data sets and was able to classify one-dimensional sample vectors using CNNs (Alkhateeb et al., 2020).

- In the paper titled: **CancerSiamese: one-shot learning for primary and metastatic tumor classification** the authors proposed **CancerSiamese**, which is a new one-shot learning model, to predict the cancer type of a query primary or metastatic tumor sample. CancerSiamese received pairs of gene expression profiles and learned a representation of similar or dissimilar cancer types through two parallel CNNs joined by a similarity function. Their work demonstrated, for the first time, the feasibility of applying one-shot learning for expression-based cancer type prediction when gene expression data of cancer types are limited (Mostav et al., 2020).

- In the paper titled: **Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks** the authors presented a feature representation approach termed **REFINED** (REpresentation of Features as Images with NEighborhood Dependencies) to arrange high-dimensional vectors in a compact image form. They generated a concise feature map in the form of a two-dimensional image. They illustrated the superior predictive capabilities of REFINED in drug sensitivity prediction scenarios (Bazgir et al., 2020).

- In the paper titled: **DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture** the authors proposed **DeepInsight** which converts non-image samples into a

well-organized image-form. DeepInsight enabled feature extraction through the application of CNN for non-image samples to seize imperative information and had shown promising results. To their knowledge, this was the first work to apply CNN simultaneously on different kinds of non-image data sets (Sharma et al., 2019).

- In the paper titled: **A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis** the authors addressed the inherent limitations of current pathological reporting systems, which made patient outcomes vary considerably within similarly staged patient cohorts, through the use of machine learning. They introduced a data driven framework which makes use of a large number of diverse types of features. Their framework had an outstanding performance in predicting mortality in stage II patients (AUROC=0:94), which exceeded that of current clinical guidelines (AUROC=0:65), and their work was demonstrated on a cohort of 173 colorectal cancer patients (Dimitriou et al., 2018).

- In the paper titled: **Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy** the authors developed a DL algorithm using data from 1,290 patients, and validated it on newly collected 27,113 colonoscopy images from 1,138 patients with at least one detected polyp (per-image-sensitivity, 94.38%; per-image-specificity, 95.92%; area under the receiver operating characteristic curve, 0.984), on a public database of 612 polyp-containing images (per-image-sensitivity, 88.24%), on 138 colonoscopy videos with histologically confirmed polyps (per-image-sensitivity of 91.64%; per-polyp-sensitivity, 100%), and on 54 unaltered full-range colonoscopy videos without polyps (per-image-specificity, 95.40%) (Wang et al., 2018).

- In the paper titled: **Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data** the authors wanted to validate an ML CRC detection model on a US community-based insured adult population. Their model achieved 99% specificity and had the highest accuracy in identifying right-sided CRCs (Hornbrook et al., 2017).

- In the paper titled: **Deep learning based tissue analysis predicts outcome in colorectal cancer** the authors combined convolutional and recurrent architectures to train a deep network to predict CRC outcome based on images of tumour tissue samples. The novelty of their approach is that they directly predicted patient outcome, without any intermediate tissue classification. Their results suggested that state-of-the-art DL techniques can extract more prognostic information from the tissue morphology of CRC than an experienced human observer (Bychkov et al., 2018).

- In the paper titled: **Colorectal Cancer Detection Based on Deep Learning** the authors introduced a DL-based method in CRC detection and segmentation from digitized H&E (hematoxylin and eosin) - stained histology slides. They demonstrated that their NN approach produced a median accuracy of 99.9% for normal slides and 94.8% for cancer slides compared to pathologist-based diagnosis on HE-stained slides digitized from clinical samples (Xu et al., 2020).

CHAPTER 3

Materials and Methodology

In this chapter, we discuss our methodology and the data used in out project. We tried different models to improve the performance measurements and to select the best subset of features that present the problem. For each method, we tried to optimize the performance by utilizing different parameter values. We also relied on the literature to select the best practice at each step.

3.1 Data

The data set used in our project has been downloaded from the website **cBioPrtal**. The source of the data is GDAC Firehose (with reference number 20160128) (*cBioPortal for cancer genomics*, n.d.). The data is made up of 3 omics, which are clinical features, CNA (Copy Number Alteration) and gene expression, for each patient. There are 162 patients, 81 of them are younger than 50 years of age, the rest are older than 50. Figure 6 shows he block diagram of the proposed methodology.

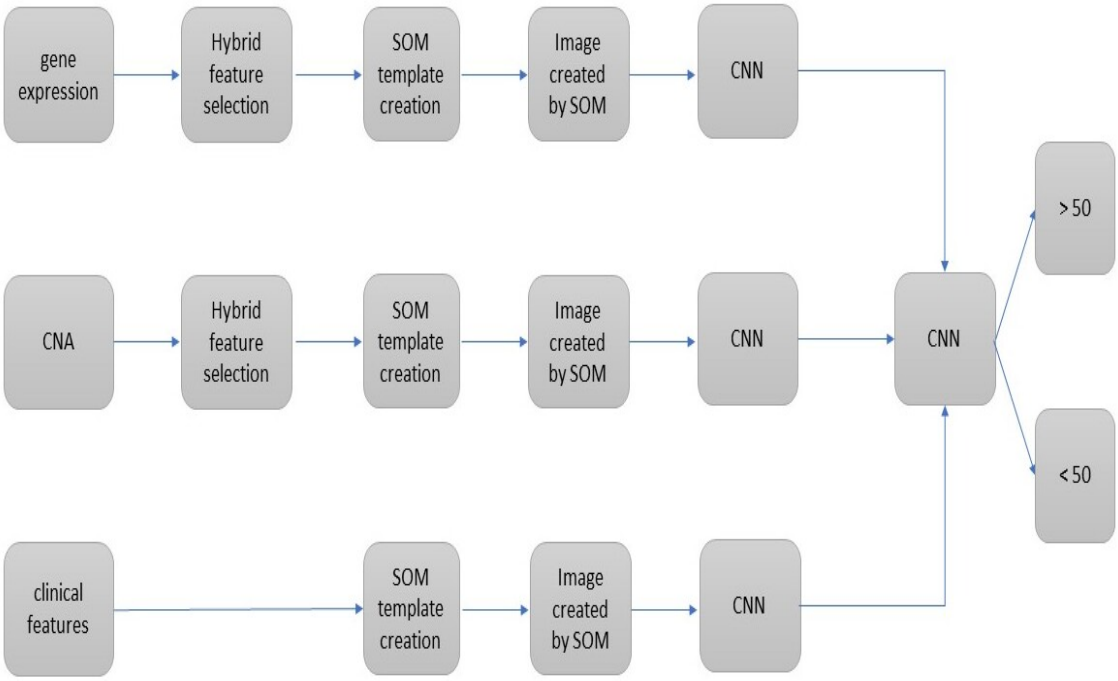

Figure 6: The block diagram of the proposed methodology.

3.2 Feature Selection

The first step in our project is utilizing a hybrid feature selection technique. In this step we combine ranking and wrapping methods. More about this hybrid approach in the following subsections.

3.2.1 Ranking

In this sub-step we rank the features that are correlated with the class using **chi-square** to remove non-relevant features. The chi-square technique measures the degree of independence of each feature in relation to the class. The following formula explains how the measurement is done:

$$\chi^2(Y,X) = \frac{N \times (AD - CB)}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

Where A is the number of times feature X occurs and Y denotes the classes (less than 50 or greater than 50). B is the number of times the feature X occurs without class Y. C is the number of times class Y occurs without X. D is the number of times neither X nor Y occurs. N is the total number of samples (In our project, the total number of samples is 162).

The number of features in the omics gene expression and CNA is very high, which will complicate the classification model and and which will lead to the problem of **Curse of Dimensionality**, which means that the number of features is way larger than the number of samples. The other reason (behind us performing ranking) is that many features do not show the discriminant behavior required to identify the classes of the CRC patients. To overcome those complications, we reduce the number of features using chi-square.

3.2.2 Wrapping

On the resulting relevant features (that resulted from the ranking process), a wrapper method is utilized to extract a potential subset of features that can represent the problem (**that is the class**). The wrapper method incorporates a random forest classifier with the **Minimum Redundancy Maximum Relevance (mRMR)** technique.

The mRMR feature selection is performed at this stage to infer the interactions among the features in each omic so that we are able to select the features with higher relevance. mRMR also reduces the redundancy among the features by not considering the correlated features. mRMR minimizes redundancy using the following formula:

$$\min W_I = \frac{1}{|S|^2} \sum_{i,j \epsilon S} I(i,j) \tag{2}$$

Where S is the set of features and $I(i,j)$ is the mutual information between features i and j. A high value of W_I indicates that the feature is redundant.

The mRMR finds the relevant features by considering the class labels. It maximizes relevance using the following formula:

$$\max V_I = \frac{1}{|S|} \sum_{i \epsilon S} I(h,i) \tag{3}$$

where h is the class label.

Lastly, the algorithm maximizes the mutual information difference to select the features using the following formula:

$$\max(V_I - W_I) \tag{4}$$

If the difference of mutual information has a high value, then it indicates that the features can effectively separate our classes. We also use random forest with mRMR feature selection that uses forward search to further reduce the number of selected features. The following two tables show the results of performing our hybrid feature selection process on the two omics: **CNA** and **gene expression**.

Note: We did not apply our hybrid feature selection approach on the clinical features omic because this particular omic has few features.

3.3 Creating Images Templates using SOM

We applied SOM on each omic data set to represent the data in a two-dimensional space. We created a template image for the features in each of the three omics which is then used to present the feature map (image) for each sample in that omic. We obtained three omics templates, one for each of the used omic in this study. These templates are depicted in Figures 9, 10 and 11.

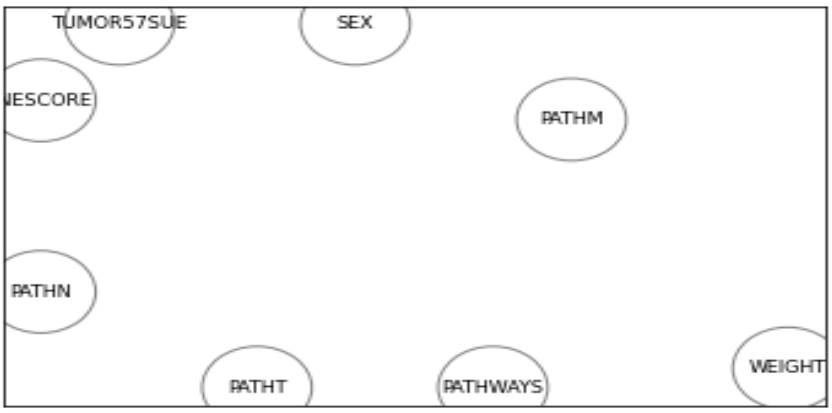

Figure 7: Images template of the clinical features omic.

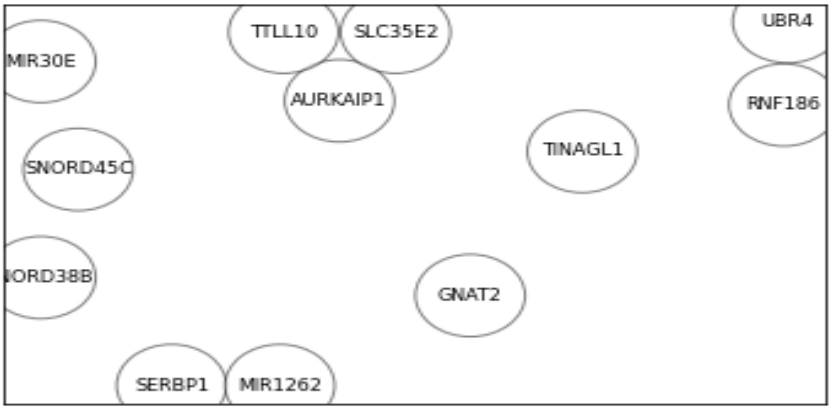

Figure 8: Images template of the CNA omic.

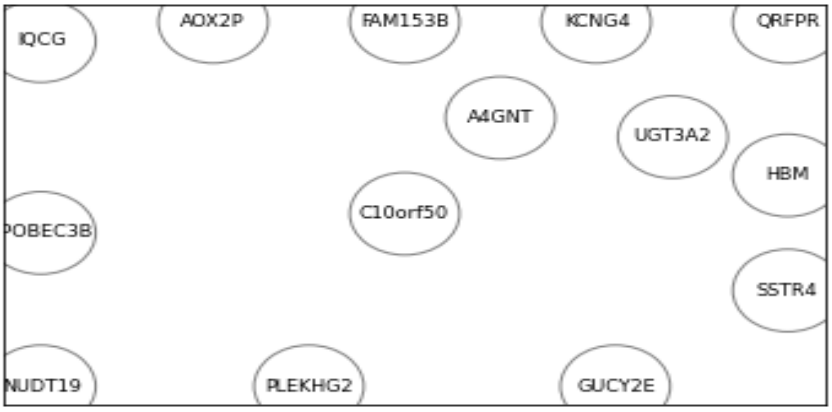

Figure 9: Images template of the gene expression omic.

For the clinical features omic, eight features (genes) resulted in its image template. They were (TUMOR57SUE, WESCORE, SEX, PATHM, PATHN, PATHT, PATHWAYS, WEIGHT). For the CNA omic, twelve features (genes) resulted in its image template. They were (MR30E, SNORD45O, SNORD38B, SERBP1, TTLL10, SLC35E2, AURKAIP1, MR1262, GNAT2, TINAGL1, UBR4, RNF186). For the gene expression omic, fourteen features (genes) resulted in its image template. They were (IQCG, POBEC3B, NUDT19, AOX2P, FAM153B, A4GNT, C10orf50, PLEKHG2, KCNG4, UGT3A2, QRFPR, HBM, SSTR4, GUCY2E).

3.4 Omics' Images Creation

Using the templates (which we mentioned in the previous section), SOM scans the values of each feature. Then the values are converted into a color and the nodes in each template are filled with that color. As an example for that, Figures 12, 13 and 14 show the resulting images for one of the young-onset CRC patients for the three omics.

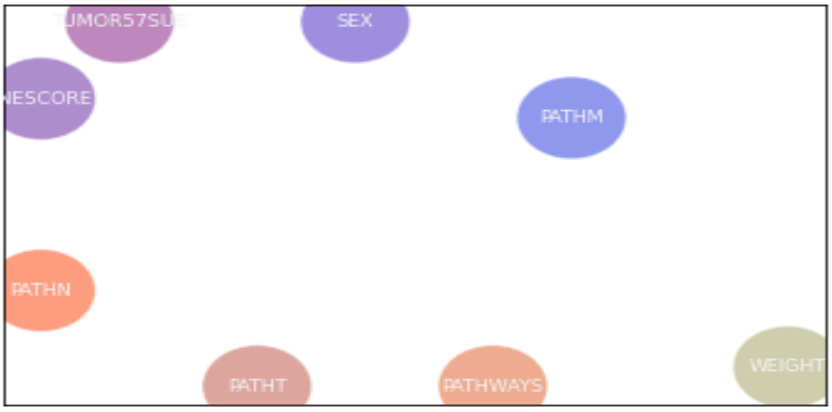

Figure 10: Image of a patient sample clinical features omic.

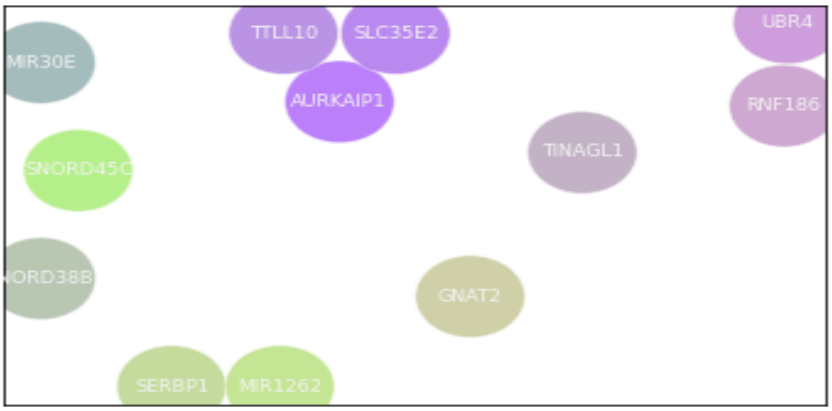

Figure 11: Image of a patient sample CNA omic.

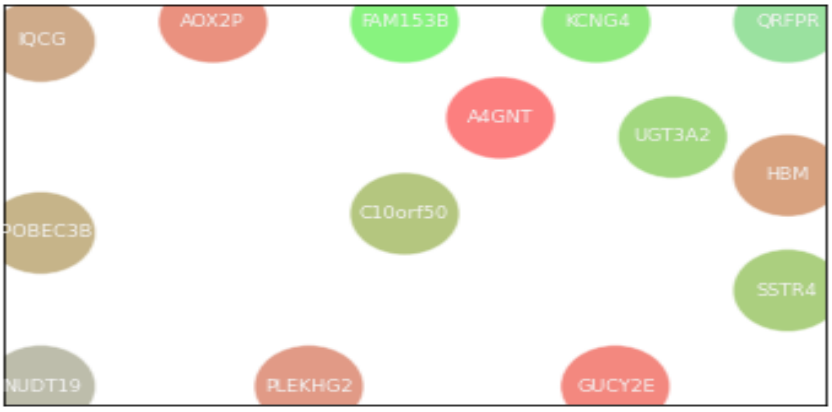

Figure 12: Image of a patient sample gene expression omic.

For the clinical features omic, a patient's sample's features (genes) were assigned purple, blue, orange, light brown and light green colors. For the CNA omic, a patient's sample's features (genes) were assigned different shades of the colors purple and green. For the gene expression omic, a patient's sample's features (genes) were assigned different shades of the colors green, red and brown.

3.5 Designing a Multi-Input CNN

We have created a model that consists of three CNNs, each of those CNNs takes an individual omic as an input then extracts the discriminant features from each omic. These **"extracted"** features are then merged to be fed into the main CNN after flattening these features. Figure 13 depicts the our model's flow diagram.

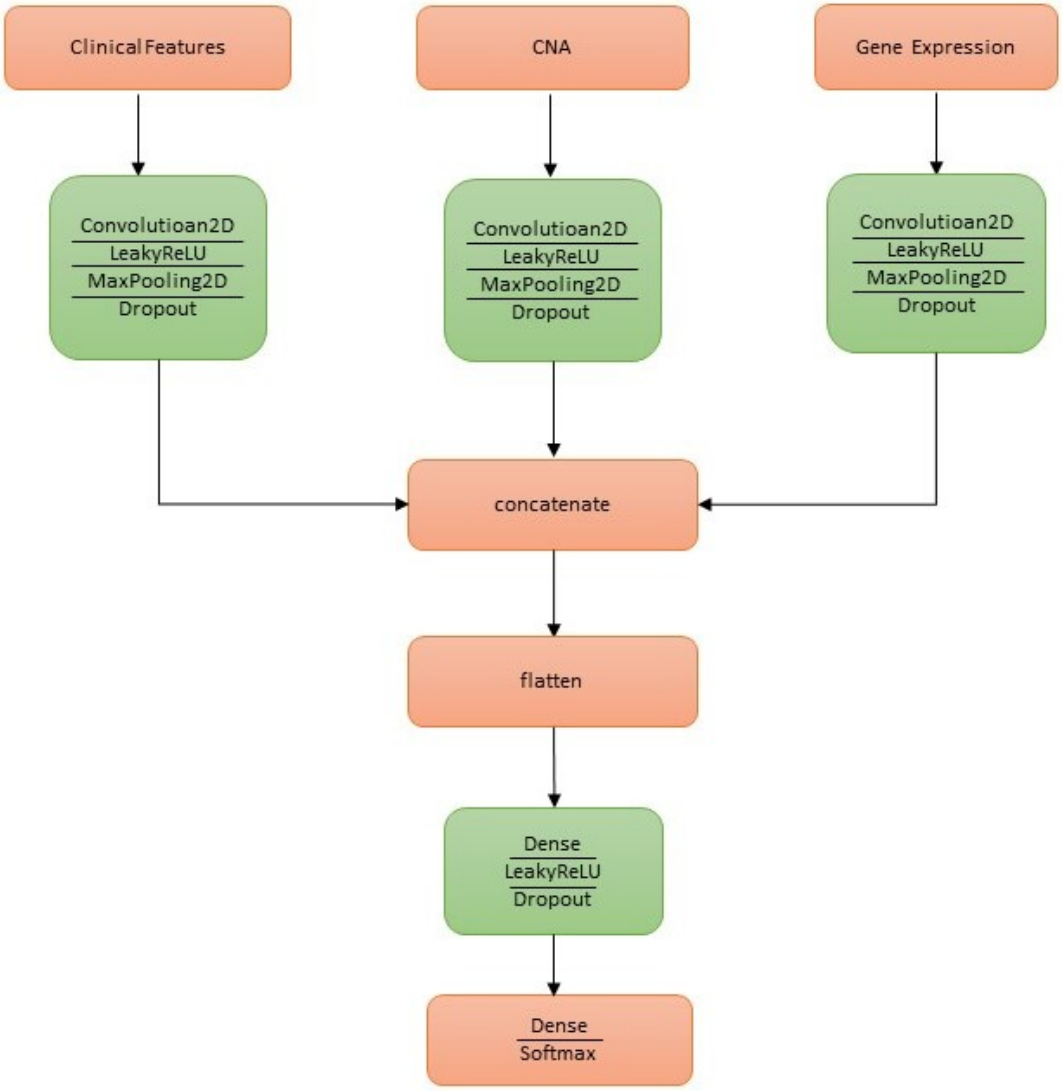

Figure 13: Design of multi-input CNNs.

For each omic we have a CNN that consists of:

- A convolutional layer that is called **Convolution2D**. This layer has 32 filters and a kernel of size 3 X 3.
- An activation function called **LeakyReLU**, which is a leaky version of a Rectified Linear Unit (**ReLU**), which allows a small gradient when the unit is not active. This layer is the input layer that expects images.

- A pooling layer called **MaxPooling2D** that takes the maximum and is configured with a pool size of 2 X 2 (it halves the input in both spatial dimensions).
- A regularization layer (using dropout) called **Dropout** which is configured to randomly exclude 25% of the neurons in the layer to reduce overfitting.

The output of the convolutional layers for the three inputs goes through a concatenation layer, and the result is converted from a two-dimensional matrix to a vector by the flatten layer. This step allows standard fully connected layers to process the output. Next we have a fully connected layer with 512 neurons, a Leaky rectifier activation function followed by another Dropout layer configured to exclude 50% of neurons, this integrated entire unit randomly processes the output from the flatten layer. Finally, the output layer has neurons and a softmax activation function to output probability-like predictions for each class.

For the optimization algorithm, we chose **Adam** (Adaptive Moment Estimation) and for the loss function, we chose **categorical_crossentropy**. The optimization algorithm minimizes the loss function. For the implementation, we used the default setting of Google Colab. The structure of the model was built using Keras 2.3.1 on TensorFlow1.15.2. We run the code on Google Colab utilizing TPU with 8 cores.

CHAPTER 4

Results

In this chapter we present the results of our research in terms of determining set of features (genes) that are associated with CRC as well as some performance measurements for ML algorithms in general and how well our model performed in accordance to those measurements.

4.1 9-Fold Cross Validation

Cross validation is any of many model validation techniques for estimating how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the model is predicting and the model's creator wants to assess how accurately that model can predict. In a prediction problem, a model is usually given a data set of known data on which training is run called **training data set**, and a data set of unknown data against which the model is tested called the **testing data set**. The goal of cross-validation is to test the model's ability to predict new data in order to point out problems like overfitting or selection bias (Cawley & Talbot, 2010) and to give an insight on how the model will generalize to an independent data set.

In k-fold cross-validation, the original sample is randomly partitioned into k equal sized sub-samples. Of which, a single sub-sample is retained as the validation data for testing the model, and the remaining sub-samples are used as training data. The cross-validation process is then repeated k times, with each of the k sub-samples used exactly once as the validation data. The k results can then be averaged to produce a single estimation.

One advantage of k-fold cross validation is that all observations are used for both training and validation, and each observation is used for validation exactly once. For our model, We split the data set into nine equal sections with each section containing sixteen samples. Then we perform a loop for nine times. Each time, we take one of the sections for the testing, and the remaining are used for the training. Then, we return the testing section and take the next section and use it for testing and the remaining sections for training and so on. This technique maximizes the training phase to learn from 88.88% of the samples each time, and, overall, tests all the samples. In other words, 100% of the samples are tested.

Tables 1 and 2 show the results of performing our hybrid feature selection process on the two omics: **CNA** and **gene expression**.

Table 1: Selected features from CNA omic.

Omic	Selected Features	Count
CNA	MR30E, SNORD45O, SNORD38B, SERBP1, TTLL10, SLC35E2, AURKAIP1, MR1262, GNAT2, TINAGL1, UBR4, RNF186	12

Table 2: Selected features from gene expression omic.

Omic	Selected Features	Count
gene expression	IQCG, POBEC3B, NUDT19, AOX2P, FAM153B, A4GNT, C10orf50, PLEKHG2, KCNG4, UGT3A2, QRFPR, HBM, SSTR4, GUCY2E	14

4.2 Performance Measurements

To evaluate our model, we are using the following performance measurements:

- **Sensitivity**: (also called Recall) is a measure of the proportion of actual positive cases that got predicted correctly (also called true positive **(TP)** cases). The existence of true positives implies that there will be another proportion of actual positive cases, which would get predicted incorrectly as negative (also described as false negative **(FN)** cases). This can also be represented in the form of a false negative rate. The sum of sensitivity and false negative rate would be 1. Sensitivity is calculated using the following equation:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

- **Specificity**: is the proportion of actual negatives which were predicted correctly (also called true negative **(TN)** cases). The existence of true negatives implies that there will be another proportion of actual negative cases, which got predicted falsely as positive; thus, termed as false positives **(FP)** cases. This proportion could also be called a false positive rate. The sum of specificity and false positive rate would always be 1. Specificity is calculated using the following equation:

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

- **Precision**: this measure is calculated as the number of **TP**s divided by the summation of total number of **TP**s and **FP**s. Precision is calculated using the following equation:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- **F-score**: this measure is calculated from the precision and sensitivity of the model. F-score is calculated using the following equation:

$$F - score = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \tag{4}$$

- **Accuracy**: is the most intuitive performance measure and it is simply a ratio of correctly predicted observation **(TP & TN)** to the total observations. A high accuracy may suggest that our model is the among the best, but only when we have symmetric data sets where values of false positive and false negatives are almost same. Otherwise, we need to use other performance measurements to evaluate the performance of our model. Accuracy is calculated using the following equation:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

Next table is the confusion matrix that resulted from our experiments.

Table 3: Resulting confusion matrix.

	True Positive	True Negative
Predicted Positive	73	8
Predicted Negative	7	74

.

- **Sensitivity** was calculated as 91.25%
- **Specificity** was calculated as 90.24%
- **Precision** was calculated as 90.12%
- **F-score** was calculated as 90.68%
- **Accuracy** was calculated as 90.74%

Comparing our model's performance measures with another model that was built for predicting the (**FOLFOX**) treatment response in metastatic or recurrent CRC patients via machine learning algorithms (W. Lu et al., 2020), we establish that our model achieved higher sensitivity and specificity that the latter. The treatment response prediction model produced an average of 85% in the sensitivity measure and an average of 64% in the specificity measure.

Resulting features from our hybrid feature selection approach on the two omics (**CNA** and **gene expression**) were checked if they are associated with the CRC disease. Tables 4 and 5 specify which of the resulting genes are associated with CRC and which are not. Both tables indicate that five genes in the CNA omic and five genes in the gene expression omic are associated with CRC.

Table 4: Association of genes (resulting from hybrid feature selection on CNA omic) with CRC.

Gene	Associated with CRC (Yes or No)
MR30E	No
SNORD45O	No
SNORD38B	No
SERBP1	No
TTLL10	No
SLC35E2	Yes
AURKAIP1	Yes
MR1262	No
GNAT2	Yes
TINAGL1	No
UBR4	Yes
RNF186	Yes

Table 5: Association of genes (resulting from hybrid feature selection on gene expression omic) with CRC.

Gene	Associated with CRC (Yes or No)
IQCG	No
POBEC3B	No
NUDT19	No
AOX2P	No
FAM153B	Yes
A4GNT	No
C10orf50	No
PLEKHG2	Yes
KCNG4	No
UGT3A2	Yes
QRFPR	Yes
HBM	No
SSTR4	Yes
GUCY2E	No

We used the literature to verify the associations between the findings (resulting genes) and CRC. The verification of whether a gene is associated with CRC or not, was done by finding a peer-reviewed paper that mentioned how a certain gene contributes in the advancement of CRC, or if it (the specific gene) is found to be significantly expressed in the CRC tissue. The expression of the gene can be in terms of gene expression, DNA methylation or DNA gene copy. These verifications are summarized below:

- For the gene **SLC35E2**, it was found to be associated with lymphovascular invasion in CRC. Also, DNA copy number was decreased for this gene in CRC (Nakao et al., 2011).
- For the gene **AURKAIP1**, it was found to be negatively correlated to irinotecan sensitivity for irinotecan sensitivity/resistance in CRC cell lines (X.-X. Li et al., 2014).
- For the gene **GNAT2**, it was among a list of 524 genes for panel sequencing found to be associated with CRC (Ge et al., 2019).
- For the gene **UBR4**, (Protein name: Retinoblastoma-associated factor 600), it was found differentially-expressed in CRC metastatic cells (Luque-Garcia et al., 2010).
- For the gene **RNF186**, it was found that the expression of **RNF186** was decreased in CRC tissues compared to corresponding normal tissues. It was also demonstrated that upregulation of **RNF186** suppressed CRC cell growth and migration in vitro and in vivo (Ji et al., 2020).
- For the gene **FAM153B**, it was among a list of 524 genes for panel sequencing found to be associated with CRC (Ge et al., 2019).
- For the gene **PLEKHG2**, it was found to be a part of a lncRNA-miRNA-mRNA ceRNA network and ceRNAs suggest the transformation from IBD to CRC. Also it was found to be responsible of positive regulation of the apoptotic process(Sun et al., 2019).

- For the gene **UGT3A2**, it was among a list of 524 genes for panel sequencing found to be associated with CRC (Ge et al., 2019).
- For the gene **QRFPR**, it was among a list of 524 genes for panel sequencing found to be associated with CRC (Ge et al., 2019).
- For the gene **SSTR4**, the hypermethylation of this gene is frequently observed in CRC (J. Li et al., 2017).

We analyzed the protein expression of some of the the resulting genes. Chen-Chu et al. (Chu et al., 2018) reported that gene **SLC35E2** was among another 27 genes that were found to be differentially expressed in CRC. Using GO and KEGG pathway analyses, the researchers performed a comprehensive analysis to discover the biological functions of differentially expressed circular RNAs. They found that **SLC35A2** created 9 protein interactions with the rest of the 27 reported genes. For the gene **AURKAIP1** (AUrora Kinase A-Interacting Protein), it was found among other mitochondrial ribosomal proteins that perform synthesis in many human diseases (including CRC) (Gopal & Rajkumar, 2016). The gene **PLEKHG2** was found among other Rho family guanine nucleotide exchange factors (RhoGEFs) genes and those are found to be over-expressed and responsible for the the protein synthesis in the CRC tissue in comparison to adjacent normal tissues. This set of genes facilitates the progression of CRC (Lee, 2016). The **SSTR4** protein expression has an association with the clinico-pathological factors, cell proliferation, Bcl-2 and p53 expression in CRC cells (Qiu et al., 2006).

CHAPTER 5

Conclusion and Future Work

In this chapter we present what we concluded from this project and suggest directions for future work.

5.1 Conclusion

Our supervised learning model was able to consider all omics with no bias or directive from a specific omic. We were able to achive this thesis's objective of taking data from a publicly available web portal, pre-process that data, create SOM features images which were fed into individual CNNs and then we were able to (eventually) classify the cancer data sets into specific labels (classes). The main advantages that we were able to achieve in this project are handling multiple input data sets and the ability to select discriminant features that best represent the data set. We examined scientific papers about how those discriminant features (genes) participate in the biological process of advancing or suppressing cancer.

5.2 Future Work

This work can be further extended by:

- Verifying the findings of the gene expression and CNA omics using wet lab experiments (Tabl et al., 2018) and pathway analysis. Pathway analysis can be done from the literature and using **KEGG** database (Ogata et al., 1998). **KEGG** database "is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for metabolism, genetic information processing, cellular processes, organismal systems, human diseases and drug development" (*KEGG PATHWAY Database*, n.d.).

- Trying a different dimensionality reduction technique for mapping, other than SOM, like Principal Component Analysis (PCA), Fisher Linear Discriminant or Multiple Discriminant Analysis (MDA).

- Applying the proposed methodology on multi-input data sets (more than three) to verify the prediction strength of the model. Then, comparing the results with state-of-the-art methods for multi-input data sets.

REFERENCES

Alkhateeb, A., Zhou, L., Tabl, A. A., & Rueda, L. (2020). Deep Learning Approach for Breast Cancer InClust 5 Prediction Based on Multiomics Data Integration. New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3388440.3415992`

Atikukke, G., Alkhateeb, A., Porter, L., Fifield, B., Cavallo-Medved, D., Facca, J., ... Misra, S. (2020). P-370 comprehensive targeted genomic profiling and comparative genomic analysis to identify molecular mechanisms driving cancer progression in young-onset sporadic colorectal cancer. *Annals of Oncology*, *31*.

Bazgir, O., Zhang, R., Dhruba, S. R., Rahman, R., Ghosh, S., & Pal, R. (2020). Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nature Communications*, *11*. doi: 10.1038/s41467-020-18197-y

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394 - 424.

Brody, H. (2015). Colorectal cancer. *Nature*, *521*.

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P., Verrill, C., ... Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, *8*. doi: 10.1038/s41598-018-21758-3

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*(70), 2079 - 2107. Retrieved from `http://jmlr.org/papers/v11/cawley10a.html`

cbioportal for cancer genomics. (n.d.). `https://www.cbioportal.org/study/summary?id=coadread_tcga`. (Accessed: 2021-01-20)

Chai, H., Zhou, X., Cui, Z., Rao, J., Hu, Z., Lu, Y., ... Yang, Y. (2019). Integrating multi-omics data with deep learning for predicting cancer prognosis. *bioRxiv*.

Chen, Y.-Y., Lin, Y.-H., Kung, C. M.-H., Chia-Ching, & Yen, I.-H. (2019). Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes. *Sensors (Basel, Switzerland)*, *19*. doi: 10.3390/s19092047

Chu, C., Chunshuai, W., Jiajia, C., Jinlong, Z., Pengfei, X., Jiawei, J., & Zhiming, C. (2018). Transcriptional information revealed differentially expressed circular RNAs in facet joint osteoarthritis. *Biochemical and Biophysical Research Communications*, *497*(2), 790 - 796. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0006291X18303942` doi: https://doi.org/10.1016/j.bbrc.2018.02.157

Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., & Henderson, D. (1990). *Handwritten digit recognition with a back-propagation network*. Morgan Kaufmann Publishers Inc.

Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., & Sahinalp, S. C. (2011). Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics (Oxford, England)*, *27*, i205 - i213. doi: 10.1093/bioinformatics/btr245

Dimitriou, N., Arandjelovic, O., Harrison, D., & Caie, P. (2018). A principled machine

learning framework improves accuracy of stage II colorectal cancer prognosis. *npj Digital Medicine*, *1*. doi: 10.1038/s41746-018-0057-x

Elkarami, B., Alkhateeb, A., & Rueda, L. (2016, 05). Cost-sensitive classification on class-balanced ensembles for imbalanced non-coding RNA data. In (p. 1 - 4). doi: 10.1109/EMBSISC.2016.7508607

Fatima, N., & Rueda, L. (2020). iSOM-GSN: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps. *Bioinformatics*, *36*(15), 4248 - 4254. Retrieved from `https://doi.org/10.1093/bioinformatics/btaa500` doi: 10.1093/bioinformatics/btaa500

Ge, W., Cai, W., Bai, R., Hu, W., Wu, D., Zheng, S., & Hu, H. (2019). A novel 4-gene prognostic signature for hypermutated colorectal cancer. *Cancer management and research*, *11*, 1985 - 1996. doi: https://doi.org/10.2147/CMAR.S190963

Gopal, G., & Rajkumar, T. (2016). Mammalian mitochondrial ribosomal small subunit (MRPS) genes: A putative role in human disease. *Gene*, *589*. doi: 10.1016/j.gene.2016.05.008

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (p. 855 - 864).

Hahn, M. E., & MacLean, M. S. (1955). *Counseling psychology* (2nd ed.). McGraw-Hill.

Hamzeh, O., Alkhateeb, A., Zheng, J., Kandalam, S., Lueng, C., & Rueda, L. (2019). A Hierarchical Machine Learning Model to Discover Gleason Grade Group-specific Biomarkers in Prostate Cancer.
doi: 10.20944/preprints201911.0298.v1

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*.

Hertel, L., Barth, E., Käster, T., & Martinetz, T. (2015). Deep convolutional neural networks as generic feature extractors. In (p. 12 - 16).

Hinton, G., & Sejnowski, T. J. (Eds.). (1999). *Unsupervised Learning: Foundations of Neural Computation* (1st ed.). Bradford Books.

Hornbrook, M., Goshen, R., Choman, E., OKeeffe-Rosetti, M., Kinar, Y., Liles, E., & Rust, K. (2017). Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. *Digestive Diseases and Sciences*, *62*, 1 - 9. doi: 10.1007/s10620-017-4722-8

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Introduction to Statistical Learning*. Springer.

Ji, Y., Tu, X., Hu, X., Wang, Z., Gao, S., Zhang, Q., ... Chen, W. (2020). The role and mechanism of action of RNF186 in colorectal cancer through negative regulation of NF-κB. *Cellular Signalling*, *75*. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0898656820302412` doi: https://doi.org/10.1016/j.cellsig.2020.109764

Kegg pathway database. (n.d.). `https://www.genome.jp/kegg/pathway.html`. (Accessed: 2021-01-20)

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems* (Vol. 25, p. 1097 - 1105). Curran Associates, Inc.

Lee, S. (2016). *Expression and function of the Rho guanine nucleotide exchange factor ECT2 in colorectal cancer* (Unpublished doctoral dissertation). Convergence Science and Technology.

Li, J., Chen, C., Bi, X., Zhou, C., Tao, H., Ni, C., ... Duan, S. (2017). DNA methylation of CMTM3, SSTR2 and MDFI genes in colorectal cancer. *Gene*, *630*. doi: 10.1016/j.gene.2017.07.082

Li, X.-X., Zheng, H.-T., Peng, J.-J., Huang, L.-Y., Shi, D.-B., Liang, L., & Cai, S. (2014). RNA-seq reveals determinants for irinotecan sensitivity/resistance in colorectal cancer cell lines. *International journal of clinical and experimental pathology*, *7*, 2729 - 2736.

Lu, M., & Zhan, X. (2018). The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *European Association for Predictive Preventive Personalised Medicine (EPMA)*(9), 77 - 102.

Lu, W., Fu, D., Kong, X., Huang, Z., Hwang, M., Zhu, Y., ... Ding, K. (2020). FOLFOX treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms. *Cancer Medicine*, *9*. doi: 10.1002/cam4.2786

Luque-Garcia, J., Martinez-Torrecuadrada, J., Epifano, C., Cañamero, M., Babel, I., & Casal, J. (2010). Differential protein expression on the cell surface of colorectal cancer cells associated to tumor metastasis. *Proteomics*, *10*, 940 - 952. doi: 10.1002/pmic.200900441

Milioli, H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., & Moscato, P. (2015). The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. *PloS one*, *10*. doi: 10.1371/journal.pone.0129711

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning* (4th ed.). The MIT Press.

Mostav, M., Chiu, Y.-C., Chen, Y., & Huang, Y. (2020). CancerSiamese: one-shot learning for primary and metastatic tumor classification. *bioRxiv*. doi: 10.1101/2020.09.07.286583

Nakao, M., Kawauchi, S., Uchiyama, T., Adachi, J., Ito, H., Chochi, Y., ... Sasaki, K. (2011). Dna copy number aberrations associated with the clinicopathological features of colorectal cancers: Identification of genomic biomarkers by array-based

comparative genomic hybridization. *Oncology Reports*, 1603 - 1611. doi: https://doi.org/10.3892/or.2011.1246

Ogata, H., Goto, S., Fujibuchi, W., & Kanehisa, M. (1998). Computation with the KEGG pathway database. *Biosystems*, *47*, 119 - 128. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0303264798000173` doi: https://doi.org/10.1016/S0303-2647(98)00017-3

Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational Intelligence: A Logical Approach*. Oxford University Press.

Powers, D. (2008). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness Correlation..

Qiu, C.-Z., Wang, C., Huang, Z.-X., Zhu, S.-Z., Wu, Y.-Y., & Qiu, J.-L. (2006). Relationship between somatostatin receptor subtype expression and clinicopathology, Ki-67, Bcl-2 and p53 in colorectal cancer. *World journal of gastroenterology*, *12*. doi: 10.3748/wjg.v12.i13.2011

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson.

Sharma, A., Vans, E., Shigemizu, D., Boroevich, K., & Tsunoda, T. (2019). Deepinsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific Reports*, *9*. doi: 10.1038/s41598-019-47765-6

Spiliopoulou, A., Nagy, R., Bermingham, M. L., Huffman, J. E., Hayward, C., Vitart, V., ... Haley, C. S. (2015). Genomic prediction of complex human traits: relatedness, trait architecture and predictive meta-models. *Human molecular genetics*, *24*, 4167 - 4182. doi: https://doi.org/10.1093/hmg/ddv145

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77 -

89. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0034425797000837` doi: https://doi.org/10.1016/S0034-4257(97)00083-7

Stewart, B. W., & Wild, C. P. (Eds.). (2014). *World cancer report 2014*. International Agency for Research on Cancer (IARC).

Sun, F., Liang, W., Tang, K., Hong, M. H., & Qian, J. (2019). Profiling the lncRNA-miRNA-mRNA ceRNA network to reveal potential crosstalk between inflammatory bowel disease and colorectal cancer. *peerJ*, *7*.

Tabl, A. A., Alkhateeb, A., Pham, H. Q., Rueda, L., ElMaraghy, W., & Ngom, A. (2018). A novel approach for identifying relevant genes for breast cancer survivability on specific therapies. *Evolutionary Bioinformatics*, *14*.

Tomczak, K., Czerwiska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland)*, *19*(1A), A68 - A77.

Trichopoulos, D., Li, F. P., & Hunter, D. J. (1996). What causes cancer? *Scientific American*, *275*(3), 80 - 87.

Umetani, N., Giuliano, A. E., Hiramatsu, S. H., Amersi, F., Nakagawa, T., Martino, S., & Hoon, D. S. B. (2006). Prediction of breast tumor progression by integrity of free circulating dna in serum. *Journal of Clinical Oncology*, *24*, 4270 - 4276. doi: 10.1200/JCO.2006.05.9493

Valueva, M. V., Nagornov, N. N., Lyakhov, P. A., Valuev, G. V., & Chervyakov, N. I. (2020). Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, *177*, 232 - 243. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0378475420301580`

Vucic, E. A., Thu, K. L., Robison, K., Rybaczyk, L. A., Chari, R., Alvarez, C. E., & Lam, W. L. (2012). Translating cancer 'omics' to improved outcomes. *Genome research*, *22*(2), 188 - 195.

Wang, P., Xiao, X., Brown, J., Berzin, T., Tu, M., Xiong, F., ... Lai, L. (2018). Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature Biotechnology*, 741 - 748. doi: 10.1038/s41551-018-0301-3

Xu, L., Walker, B., Liang, P.-I., Tong, Y., Xu, C., Su, Y., & Karsan, A. (2020). Colorectal cancer detection based on deep learning. *Journal of Pathology Informatics*, *11*. doi: 10.4103/jpi.jpi_68_19

VITA AUCTORIS

Name: Noor Kammonah

Education: B.Sc Computer Science, Yarmouk University, Irbid, Jordan

M.Sc. Computer Science, Yarmouk University, Irbid, Jordan

M.Sc. Computer Science, University of Windsor, Windsor, Ontario, Canada, 2021