**Multi-omics data in the IDENTIFICATION of kidney cancer subgroups**

UNDERGRADUATE RESEARCH THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE Requirements of the DEGREE OF BACHELOR OF THE SCIENCE OF ENGINEERING

**Submitted by:**

Maduranga W.P.N. (2018/E/073)

Rodrigo S.M. (2018/E/102)

**DEPARTMENT OF COMPUTER ENGINEERING**

**FACULTY OF ENGINEERING**

**UNIVERSITY OF JAFFNA**

February, 2023

**“MULTI-OMICS DATA IN THE IDENTIFICATION OF**

**KIDNEY CANCER SUBGROUPS”,**

**Research proGRESS REPORT**

**Supervisor(s):**

Supervisor : Dr. Pratheeba J.

Co-Supervisor : Dr Thuseethan Selvarajah

**Examination Committee:**

Lecturer 1 ..........................................

Lecturer 2 ..........................................

CONTRIBUTION TO THE REPORT BY THE MEMBERS IN GROUP

|  |  |  |
| --- | --- | --- |
| Sections | 2018/E/073 | 2018/E/102 |
| **CHAPTER 1: INTRODUCTION** |  |  |
| 1.1 Motivation and Overview |  | 🗸 |
| 1.2 Aims and Objectives |  | 🗸 |
| 1.3 Research Scope |  | 🗸 |
| **CHAPTER 2: LITERATURE REWIEW** |  |  |
| 2.1 Introduction | 🗸 |  |
| 2.2 Prediction Models | 🗸 |  |
| 2.3 Performance Analysis | 🗸 |  |
| 2.4 Available Datasets | 🗸 |  |
| **CHAPTER 3 : METHODOLOGY AND RESEARCH PLAN** |  |  |
| 3.1 Overview of Methodology | 🗸 | 🗸 |
| 3.2 Detailed Methodology | 🗸 | 🗸 |
| 3.2.1 Dataset Selection (Single Omic) | 🗸 |  |
| 3.2.2 Data Preparation (Single Omic) |  | 🗸 |
| 3.2.3 Data Preprocessing (Single Omic) | 🗸 | 🗸 |
| 3.2.4 Test the models (Single Omic) | 🗸 | 🗸 |
| 3.2.5 Extract Features (Single Omic) | 🗸 |  |
| 3.2.6 Dataset Preparation (Multi Omic) | 🗸 |  |
| 3.2.7 Data Preprocessing (Multi Omic) | 🗸 |  |
| 3.2.8 Test the models (Multi Omic) | 🗸 |  |
| 3.2.9 Finalize the model |  | 🗸 |
| 3.3 Timeline |  | 🗸 |
| **CHAPTER 4: PROGRESS TO DATE** |  |  |
| 4.1 Literature Review | 🗸 |  |
| 4.2 Dataset Collection | 🗸 |  |
| 4.3 Dataset analysis | 🗸 |  |
| 4.4 Research Proposal | 🗸 |  |
| 4.5 Implementations - Single Omic | 🗸 | 🗸 |
| 4.5.1 Dataset Preparation |  | 🗸 |
| 4.5.2 Preprocessing | 🗸 | 🗸 |
| 4.5.3 Testing the models |  | 🗸 |
| 4.5.4 Extract Features |  | 🗸 |
| 4.6 Results - Single Omic |  | 🗸 |
| **REFERENCE** | 🗸 | 🗸 |

TABLE OF CONTENT

[TABLE OF CONTENT 3](#_Toc126360749)

[LIST OF FIGURES 4](#_Toc126360750)

[LIST OF TABLES 5](#_Toc126360751)

[ABBREVIATIONS AND ACRONYMS 6](#_Toc126360752)

[Chapter 1: Introduction 8](#_Toc126360753)

[1.1 Motivation and Overview 8](#_Toc126360754)

[1.2 Aims and Objectives 8](#_Toc126360755)

[1.3 Research Scope 9](#_Toc126360756)

[Chapter 2: Literature Review 10](#_Toc126360757)

[2.1 Introduction 10](#_Toc126360758)

[2.2 Prediction Models 11](#_Toc126360759)

[2.3 Performance Analysis 15](#_Toc126360762)

[2.4 Available Datasets 17](#_Toc126360763)

[Chapter 3: Methodology and Research Plan 19](#_Toc126360764)

[3.1 Overview of Methodology 19](#_Toc126360765)

[3.2 Detailed Methodology 19](#_Toc126360766)

[3.3 Time line 24](#_Toc126360776)

[Chapter 4: PROGRESS TO DATE 25](#_Toc126360777)

[4.1 Literature Review 25](#_Toc126360778)

[4.2 Dataset Collection 25](#_Toc126360779)

[4.3 Dataset analysis 25](#_Toc126360780)

[4.4 Research Proposal 25](#_Toc126360781)

[4.5 Implementations - Single Omic 25](#_Toc126360782)

[4.6 Results - Single Omic 30](#_Toc126360787)

[REFERENCES 33](#_Toc126360788)

LIST OF FIGURES

[Figure 1: ﻿Basic Architecture Of RNN 12](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360887)

[Figure 2: Basic Architecture of RNN with LSTM cell 13](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360888)

[Figure 3: Architecture of LSTM cell (STATE) 13](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360889)

[Figure 4: The basic structure of a multilayer perceptron 13](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360890)

[Figure 5: Illustration of a Bayesian Neural Network 14](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360891)

[Figure 6: Structure of confusion matrix 16](#_Toc126360892)

[Figure 7: Receiver Operating Characteristic (ROC) curves and AUC 16](#_Toc126360893)

[Figure 8: Overview of the Methodology (Single Omic) 19](#_Toc126360894)

[Figure 9: Overview of the Methodology (Multi Omic) 19](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360895)

[Figure 10: Data preprocessing steps 20](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360896)

[Figure 11: Behavior of K-Fold Cross Validation 22](#_Toc126360897)

LIST OF TABLES

[Table 1: Performance analysis of models in literature 15](#_Toc126360917)

[Table 2: AUC values from different proteins 17](file:///C:\Users\HP\Desktop\2018e073_2018e102_Final_Research_Progress.docx#_Toc126360918)

[Table 3: Timeline 24](#_Toc126360919)

[Table 4: Data Shapes after merging on each omic type 26](#_Toc126360920)

[Table 5: Cleaning process on each omic type 27](#_Toc126360921)

[Table 6: Records in each omic type before and after Sampling process on each omic type 28](#_Toc126360922)

[Table 7: Optimized parameters for each algorithm 30](#_Toc126360923)

[Table 8: Complete results for Protein 31](#_Toc126360924)

[Table 9: Complete results for miRNA 32](#_Toc126360925)

[Table 10: Complete results for methylation 32](#_Toc126360926)

[Table 11: Complete results for Genomic 33](#_Toc126360927)

ABBREVIATIONS AND ACRONYMS

AUC : Area under curve

BNN : Bayesian neural network

ccRCC : clear cell RCC

chRCC : chromophobe RCC

CT : Computerized tomography

DNA : Deoxyribonucleic acid

FN : False Negative

FP : False Positive

FPKM : fragments per kilobase per million

FPR : False positive rate

GDC : Genomic Data Commons

KICH : kidney chromophobe

KIRC : kidney renal clear cell carcinoma

KIRP : kidney renal papillary cell carcinoma

KNN : K-Nearest Neighbor

LSTM : Long Short-Term Memory

MCC : Matthews Correlation Coefficient

meth : methylation data

miRNA : microRNA expression data

ML : Machine Learning

MLP : multi-layer perceptron

mRNA : messenger RNA / gene expression data

NCA : Neighborhood Component Analysis

pRCC : papillary RCC

Q/A : Question and answer

RCC : Renal cell carcinoma

RNN : recurrent neural network

ROC : Receiver Operating Characteristic

RF : random forest

RNA : Ribonucleic acid

RPPA : Reverse phase protein array

SVM : support vector machine

TCGA : The Cancer Genome Atlas data repository

TMA : tissue microarray

TN : True Negative

TP : True positive

TPR : True Positive Rate

UCSC : University of California, Santa Cruz

# Introduction

## 1.1 Motivation and Overview

Kidneys are two bean-shaped organs, each about the size of a fist. They're located behind the abdominal organs, with one kidney on each side of the spine (*Source:*[*https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work*](https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work) ). Kidney cancer is a cancer that begins in the kidneys. Cancers mainly start when cells in the body begin to grow out of control. Kidney cancer can be called as one of the common cancer variants in the world.

Identified cases of kidney cancer seems to be increasing annually. One reason for this may be that imaging techniques such as computerized tomography (CT) scans are being used more often.

People who are elder than 60 are the most affected group from Kidney cancers and about 79,000 cases are identified annually. Usually, 14,000 deaths are recorded among them. Several Kidney cancer subtypes have been identified so far as follows:

* Kidney Clear Cell Carcinoma [1,2,3,4,5,6,7,8,9,10]
* Kidney Papillary Cell Carcinoma [1,2,3,4,6,7,8,910]
* Kidney Chromophobe [1,2,3,4,6,7,8,910]
* Rhabdoid Tumor [1]
* High-Risk Wilms Tumor [1]
* Clear Cell Sarcoma [1]

Fortunately, there are considerable possibilities of getting cured of Kidney cancer, if it can be detected in the early stages and able to find the affected variant properly. There are some symptoms occur in the early stages of the cancer. Some of them are shown below.

* Blood in urine, which may appear pink, red or cola coloured
* Pain in back or side that doesn't go away
* Loss of appetite
* Unexplained weight loss
* Tiredness
* Fever

## 1.2 Aims and Objectives

Our research aim is to help the medical society to identify the kidney cancer subtypes using the omic data of the patients.

The main objective of our research is to create an accurate model for predicting the subgroup of the kidney cancer.

## 1.3 Research Scope

Our hope to develop a better model to classify Kidney cancer subgroups. To achieve that task, we followed some constrains.

* + Consider the 3 most common variants due to the rareness of other variants.
  + Focused on 4 omics approaches known a s DNA methylation, gene expression, protein expression, and miRNA due to dataset limitation.

# Literature Review

## 2.1 Introduction

Kidney cancer is one of the deadliest diseases and unfortunately it is hard to detect early through normal clinical means [1]. Renal cell carcinoma (RCC) accounts for 90% of all kidney cancers [5]. Renal cell carcinomas are derived from the renal tubular epithelium [16].

Renal cell cancers are classified on the basis of morphology and growth patterns [16]. However, recent advances in the understanding of the genetic basis of renal carcinomas have led to a new classification that takes into account the molecular origins of these tumours [16]. The three most common forms of kidney cancer are kidney renal clear cell carcinoma or clear cell RCC (KIRC or ccRCC, accounting for 70–75% of all kidney cancers), kidney renal papillary cell carcinoma or papillary RCC (KIRP or pRCC, accounting for 10–16% of all kidney cancers) and kidney chromophobe or chromophobe RCC (KICH or chRCC, accounting for 5% of all kidney cancers) [5].

Current technologies allow us to measure various molecular data, which is also called as omic data. In recent years, the reduction of costs for the sequencing of biological molecules including DNA, RNA and proteins has allowed the widespread of huge amounts of data in the form of large structured databases and in form of repositories (specially created for the study of particular pathologies) [4].

Generally, single omic data is selected and used in the cancer related studies. Some single omic data used in literature review are,

* miRNA Genome Data [1]
* transcriptomic data [2]
* Genomics [3]
* Methylation [3]
* Proteins [6,8,9]

Some researchers used a combination of omic data which is called as multi-omic data in their cancer related studies. Summary of such studies are as follows:

* mRNA, miRNA and meth data: classification on kidney samples exploiting uncertainty aware models [4]
* DNA meth and mRNA data: A Gene Signature of Survival Prediction for Kidney Renal Cell Carcinoma [5]
* Genomics, transcriptomic, epigenomics, metabolomics: precision of kidney cancer therapies [10]

In our study, we focus on kidney cancer sub-typing using multi-omics data with the help of machine learning models.

## 2.2 Prediction Models

Machine learning methods are mostly used in cancer related studies, especially as a prediction model. As we know there are various type of such models in machine learning. Our approach is to identify the kidney cancer subtype. So, for this section we only added the models which were focused the classification of kidney cancer subtype. They used both conventional machine learning models and deep learning models in their studies.

### 2.2.1 Conventional Models

Conventional machine-learning techniques have limited in capability of processing the data in their original form [11]. Here are the conventional models that we identified in literature.

* K-Nearest Neighbor (KNN) [6]

In [6] they tested the possibility of using numeric data acquired from software-based quantification of certain marker proteins (key autography proteins - ATG) for discriminating renal cell carcinoma subtypes. They used KNN algorithm for discrimination among RCC subtypes.

One of the most fundamental yet important categorization techniques in machine learning is K-nearest neighbors.

In KNN, the entire training dataset is stored. When a prediction is required, the k-most similar records to a new record from the training dataset are then located. From these neighbors, a summarized prediction is made.

Similarity between records can be measured many different ways. A problem or data-specific method can be used. Generally, with tabular data, a good starting point is the Euclidean distance.

Once the neighbors are discovered, the summary prediction can be made by returning the most common outcome or taking the average. As such, KNN can be used for classification or regression problems *(Source:*[*https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/*](https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/)*).*

* Support vector machine (SVM) [4]

In [4], they proposed a tree MLP model for classification on kidney samples exploiting uncertainty aware models and they used SVM for compare their model.

Support Vector Machine is an example of a supervised machine learning technique that offers data analysis for regression and classification. SVM is mostly used for classification, while it can also be used for regression. Plotting is performed in n-dimensional space. Each feature's value corresponds to the value of the specified coordinate. The ideal hyperplane that differentiates between the two classes is then identified.

The coordinate representations of each observation are represented by these support vectors. It is a frontier method for segregating the two classes.

* Random forest (RF) [4]

In [4], they used RF also to compare their model (tree MLP model). Here is an introduction to RF.

A type of ensemble learning technique called random forest classifiers is used for classification, regression, and other tasks that may be carried out with the use of decision trees. These decision trees can be built during training, and the class output can either be regression or classification. These random forests can be used to overcome the bad tendency of overfitting the training set.

### 2.2.2 Deep learning Models

Deep learning is an advance machine learning approach that is used to make computers able to automatically extract, analyse and understand the useful information from the raw data [11]. Here are the deep learning models found in literature.

* Long Short-Term Memory (LSTM) [1]

In [1], LSTM was used for grouping the kidney cancer subtypes. Here is an introduction to LSTM.

LSTM allows to a neural network to remember the stuff that it needs to keep hold of context and also forget the stuff that is no longer applicable. It’s a type of recurrent neural network (RNN) (Figure 1). RNN requires long-term memorization. LSTM provides more additional special units that can hold information longer using an internal state (Figure 2). State has 3 gates (Figure 3) as,

* + Forget gate: stuff which can forget
  + Input gate: new information for add or update
  + Output date: which part of instance output in a particular instance

Applications of LSTM

* + Machine translation
  + Q/A chatbot

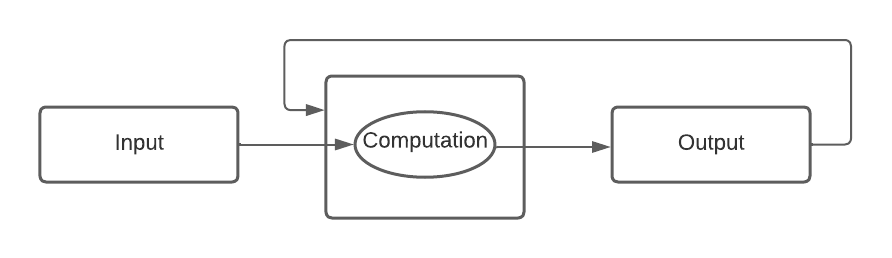
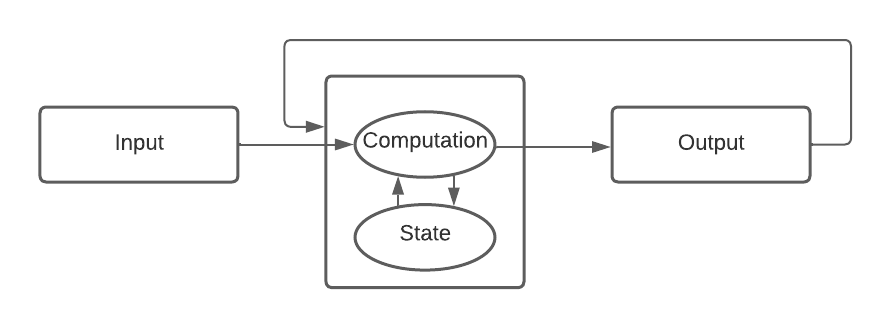


Figure 1: [﻿Basic Architecture Of RNN](https://pydeeplearning.weebly.com/blog/basic-architecture-of-rnn-and-lstm)



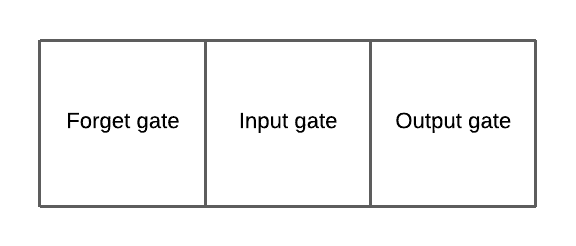


Figure 2: Basic Architecture of RNN with LSTM cell

Figure 3: Architecture of LSTM cell (STATE)

* Neighborhood Component Analysis (NCA) [1]

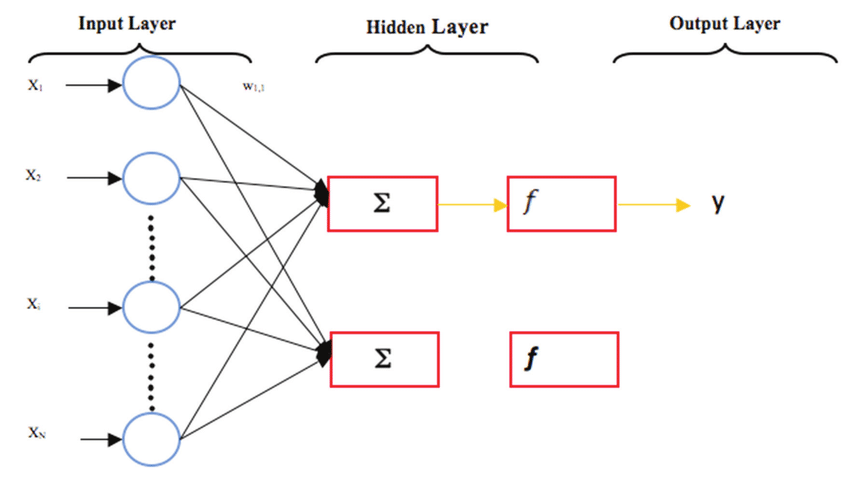
In [1], they classified the kidney cancer into its corresponding subtype using miRNA. NCA is used to extract discriminative features from miRNAs.

NCA is a supervised learning method and it is a nearest neighbour-based feature weighting algorithm.

* Multi-layer perceptron (MLP) [4]

In [4], they proposed a model which?? an extension of the multi-layer perceptron (MLP) combining several MLPs in a tree architecture (tree MLP). And they compare their model with a standard MLP.

MLP is a perceptron with multiple layers. It has 3 layers, one input layer, one output layer and some hidden layers.



; where b is the bias associated with the neurons [14]. (*Figure 4*)

Figure 4: The basic structure of a multilayer perceptron

* Bayesian neural network (BNN) [4]

In [4], they used BNN to compare their model (tree MLP model) Here is an introduction to BNN.

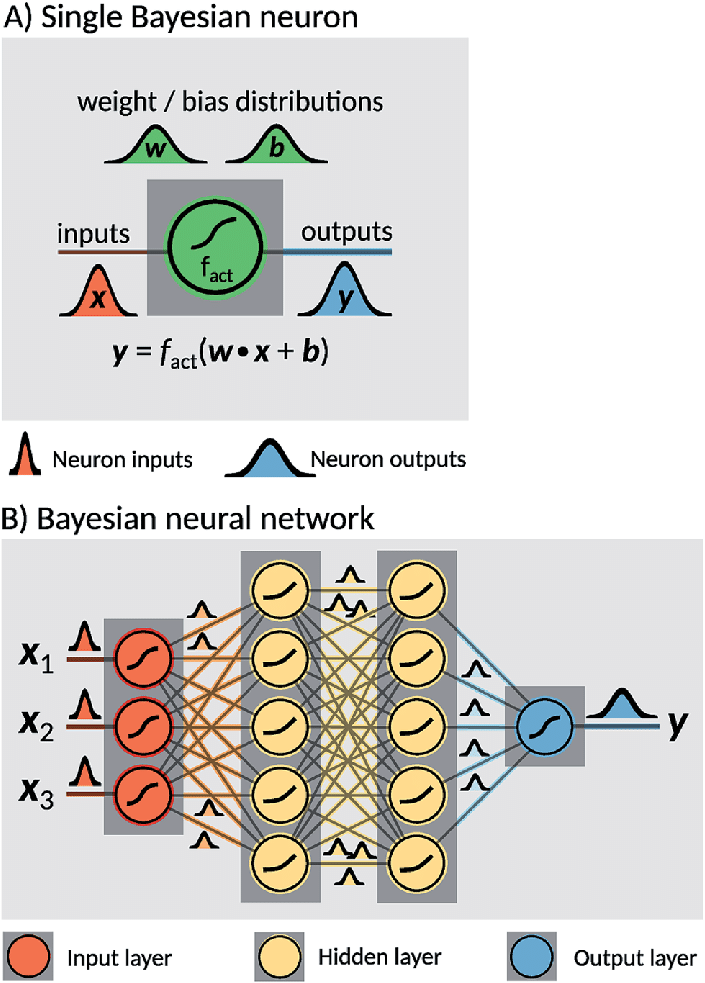
A Bayesian neuron defines a mathematical operation based on an activation function fact, a distribution of weights w and a distribution of biases b intrinsic to the neuron. Every input x is processed by sampling one instance of weights and biases from the distributions and applying the activation function. A BNN consists of a set of interconnected Bayesian neurons. The neurons in the network are organized in layers, and can differ in their activation functions as well as their weight and bias distributions [15]. Here is as illustration of BNN (*Figure 5: [15]*)

Figure 5: Illustration of a Bayesian Neural Network

## 2.3 Performance Analysis

Here we are focusing the performance analysis methods and their performances of the models that we discussed earlier. (Table 1) we can clearly see that deep learning models perform better than conventional models.

Table 1: Performance analysis of models in literature

|  |  |  |  |
| --- | --- | --- | --- |
| Study | ML model | Performance Analysis | |
| **Method** | **Performance** |
| [1] | LSTM | average accuracy | 95% |
| Matthews Correlation Coefficient value (MCC) | 0.92 |
| [4] | MLP | Precision | 98% |
| Recall | 99% |
| F1-score | 99% |
| Accuracy | 99% |
| BNN | Precision | 98% |
| Recall | 98% |
| F1-score | 98% |
| Accuracy | 98% |
| RF | Precision | 95% |
| Recall | 95% |
| F1-score | 95% |
| Accuracy | 95% |
| SVM | Precision | 95% |
| Recall | 95% |
| F1-score | 95% |
| Accuracy | 95% |
| [6] | KNN | AUC | *(In Table 2)* |
| Accuracy | 82% |
| Kappa | 0.32 |

( is the accuracy and is the hypothetical probability of chance agreement)

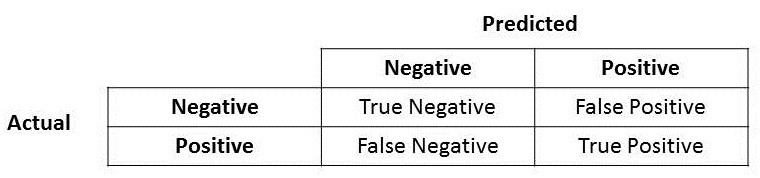


Figure 6: Structure of confusion matrix

Area Under the ROC Curve (*Figure 7: [17])*

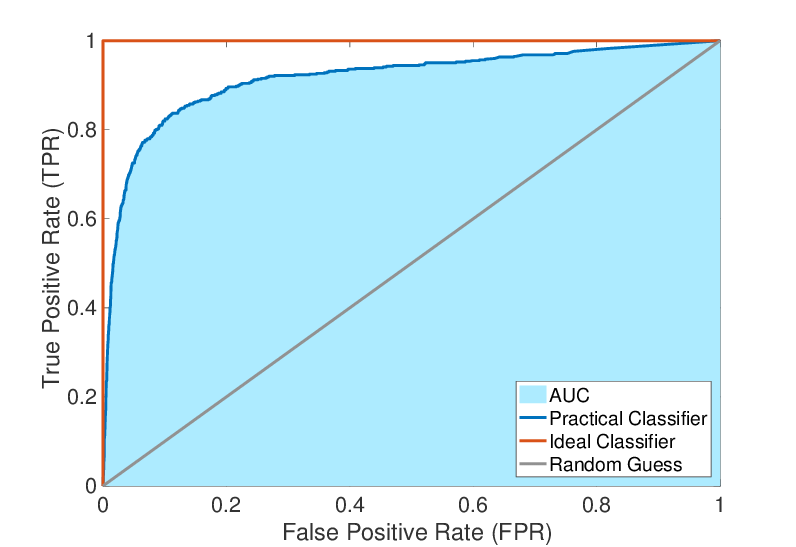


Figure 7: Receiver Operating Characteristic (ROC) curves and AUC

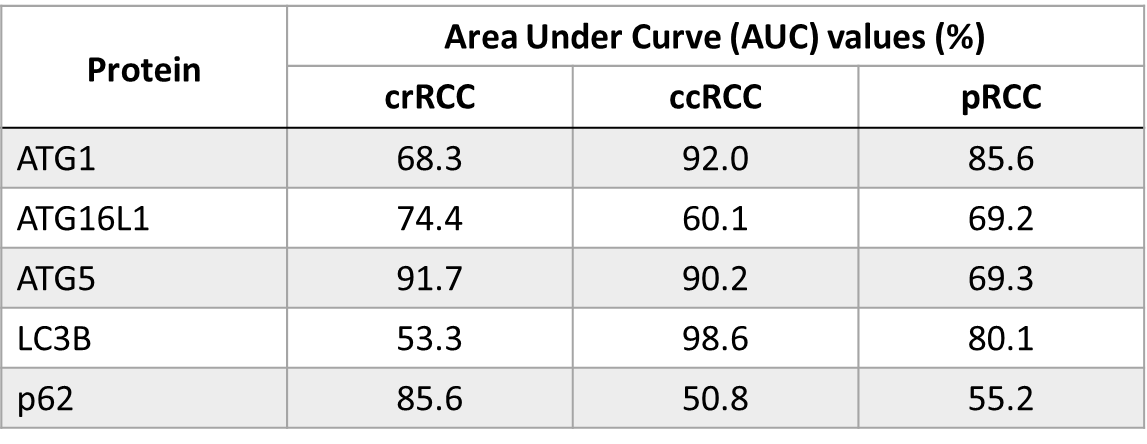


Table 2: AUC values from different proteins

## 2.4 Available Datasets

miRNA quantitative read counts data [1]

* Provided by The Cancer Genome Atlas data repository (TCGA)
* 1881 features

FPKM (fragments per kilobase per million) files [2]

* Derived from the ccRCC, pRCC and chRCC cohorts of the TCGA database
* Representing transcriptomic data of 891 patients
* Contained 20,501 genes

Kidney tumor samples from the Genomic Data Commons (GDC) database [4]

* For KIRP, KIRCH and KICH subtypes, only samples available are selected for mRNA, miRNA and meth data
* Obtained a dataset of 909 samples

The Cancer Genome Atlas (TCGA, <http://cancergenome.nih.gov/>) expression data [5]

* 529 KIRC samples and 79 normal samples

Somatic mutation information [5]

* 336 samples
* 26,693 somatic mutations involving 9290 genes.

Clinical information [5]

* 537 patients

Tissue microarray (TMA) of RCC [6]

* Containing 237 RCCs from untreated patients
* Containing 18 normal kidney tissues from healthy donors

An external validation dataset for ccRCC [7]

* Obtained from the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC; ref. 21).
* 782 ccRCC slides
* 222 patients of both normal and malignant tissue

An independent dataset [7]

* 131 patients (41 pRCC, 59 ccRCC, and 31 chRCC)
* Collected from the Brigham and Women’s Hospital Department of Pathology

# Methodology and Research Plan

## 3.1 Overview of Methodology

Our overall methodology we divided into 2 stages.

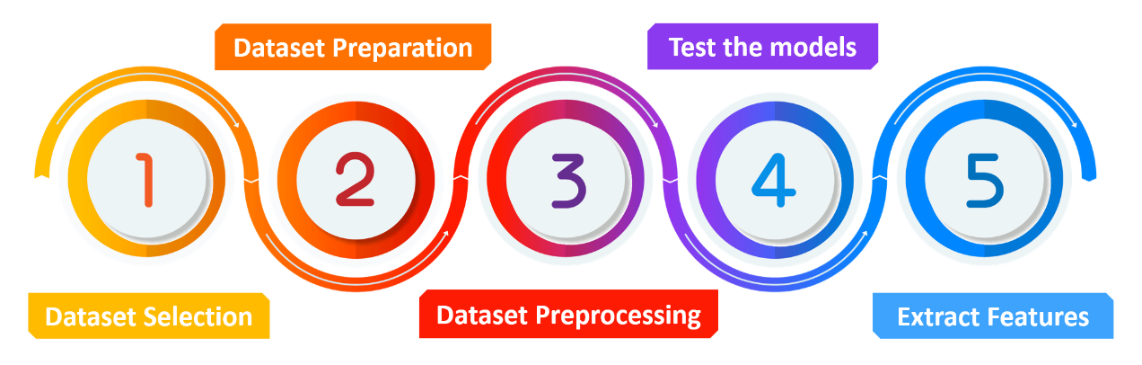
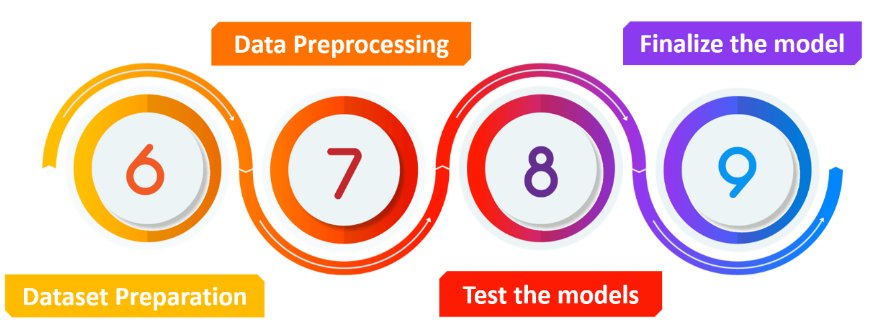
1. Single omic methodology
2. Multi omic methodology

Figure 8: Overview of the Methodology (Single Omic)

Figure 9: Overview of the Methodology (Multi Omic)



## 3.2 Detailed Methodology

The methodology can divide into 9 steps likewise the Figure 8 and 9. This section explains each step with examples.

### 3.2.1 Dataset Selection (Single Omic)

We have used the University of California, Santa Cruz (UCSC) to download the required datasets. UCSC consists of datasets in The Cancer Genome Atlas (TCGA) which is known as a reliable data repository with 33 cancer types. Among them, we focused on the dataset belonging to Kidney cancer and it included data for 3 major subtypes of relevant cancer which are known as,

* Kidney Clear Cell Carcinoma (KIRC) [1,2,3,4,5,6,7,8,9,10]
* Kidney Papillary Cell Carcinoma (KIRP) [1,2,3,4,6,7,8,9,10]
* Kidney Chromophobe (KICH) [1,2,3,4,6,7,8,9,10]

In our research, we use 4 omics variants among many others.

* DNA methylation (Methylation450k)
* gene expression (RNAseq)
* protein expression (RPPA)
* miRNA (IlluminaHiseq)

### 3.2.2 Data Preparation (Single Omic)

In our research we consider 4 omic types and 3 subtypes. So altogether we downloaded 12 datasets. Each dataset may not be use directly in the implementation process. There can be have adding some parts or removing or renaming kind of works in csv or excel format. In this step we focused on those steps.

### 3.2.3 Data Preprocessing (Single Omic)

Data Preprocessing is the process of simply transforming raw data into an understandable format. Real-world data is sometimes incomplete, inconsistent, redundant, and noisy. Data preprocessing involves various steps that help to convert raw data into a processed and sensible format. The diagram shows the various steps involved in data preprocessing [12].

Dataset Integration

Dataset Cleaning

Feature Selection

Sampling

Figure 10: Data preprocessing steps

* Data Integration

Data integration focuses on delivering a uniform view of the data from many sources and bringing them together. Conflicts occurring from the combination of data with various representations are resolved. This procedure is crucial in several scientific and industrial applications. Integrating data becomes even more important as it grows exponentially in amount.

* Data Cleaning

Finding inaccurate records and corrupt data in a record set or database table is the process of "data cleaning."

We hope to use Data cleaning to find,

* + - * Incomplete
      * Inaccurate
      * Inconsistent
      * Irrelevant data in our dataset.

Then we hope to modify or remove that data considering the requirement.

* Feature Selection

Feature selection is a dimensionality reduction technique, which is to choose a small subset of the relevant features from the original features by removing irrelevant, redundant, or noisy features [13]. Mostly, feature selection can result in improved learning performance, such as increased learning accuracy, reduced computing expense, and improved model interpretability. Researchers in the fields of computer vision, text mining, and other fields have recently presented a range of feature selection algorithms and demonstrated the efficacy of their works through theory and experiment. The objective of this study is to review the current state of the art for these techniques. Additionally, a complete experiment is run to see if feature selection can enhance learning performance while taking into account some of the methods discussed in the literature.

**Pearson correlation** is the feature selection technique which we hope to use in our research.

The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation [14].

* Sampling

Sampling is a technique used in machine learning to select a subset of data from a larger dataset for training or testing models. The purpose of sampling is to reduce the size of the data to make the process of training and evaluating models more efficient, without sacrificing too much accuracy. Different sampling techniques are used in machine learning, such as random sampling, stratified sampling, and systematic sampling. The choice of sampling technique depends on the nature of the data and the problem being solved.

### 3.2.4 Test the models (Single Omic)

There are several techniques used to test machine learning models:

1. Split data
2. Cross-Validation
3. Performance Metrics
4. Model Comparison

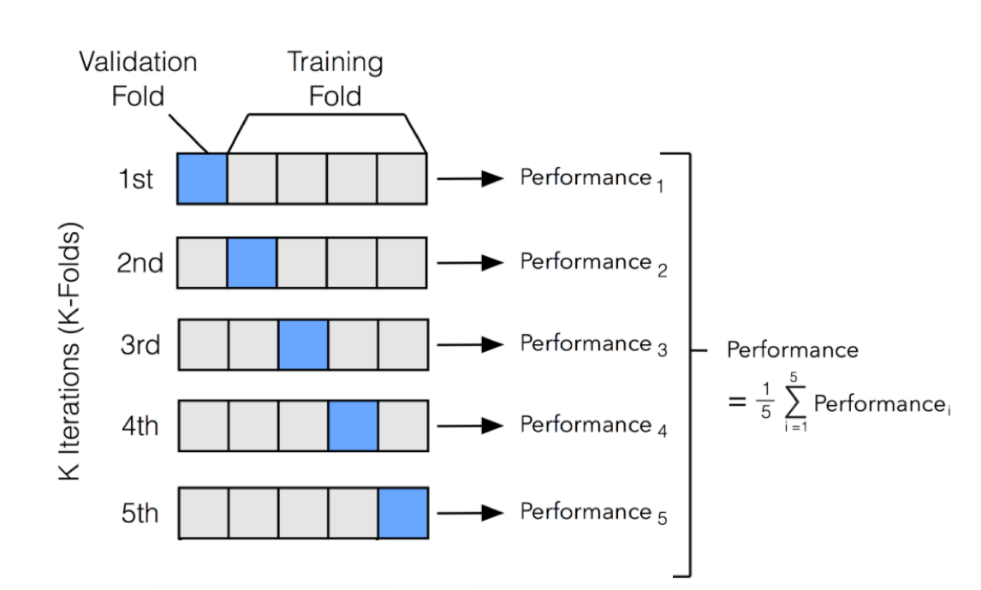
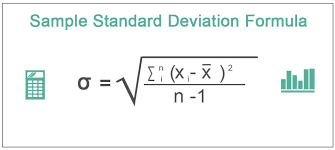
In our work we hope to use K-Fold Cross-Validation to evaluate the best performing model. K-Fold Cross-Validation is a model validation technique used in machine learning to assess the performance of a model. In this technique, the data is divided into k equal-sized parts or folds. The model is trained on k-1 of these folds and tested on the remaining one. This process is repeated k times, with each fold used once as the test set. The average performance metric across all k iterations is then used to assess the model's performance. This method helps in reducing the chances of overfitting and ensures that the model is not trained or tested on the same data, providing a better estimate of its performance(accuracy) on unseen data.

Figure 11: Behavior of K-Fold Cross Validation

In cross validation we hope to use following Performance Measures to our model testing. provides below results after executing its process;

* Accuracy
  + Accuracy obtains from below formula;
* Standard Deviation
  + Standard Deviation results from below formula;

And here we are going to solve a classification problem. So many algorithms exist related to these categories. And we have planned to use the 3 most popular algorithms in that category for our project. Because in the literature review, we observed these algorithms were well performed in same category that we are focusing. Those models are as below,

* K-Nearest Neighbor (KNN) algorithm [6]
* Random forest algorithm [4]h
* Support Vector Machine (SVM) [4]

In the literature we found some deep learning models also in same category. As we see in the literature, they performed better than conventional models. But those models require very large amount of data and more computational power. So, we focused only conventional models.

### 3.2.5 Extract Features (Single Omic)

Using selected prediction models, we hope to test for different number of feature sets. Among them we hope to get the best features on each omic type.

### 3.2.6 Dataset Preparation (Multi Omic)

After extracting the features for each omic type, we can use them to create a new dataset which contain all omic type (all selected features). In this step, we hope to create a proper dataset for multi-omic process.

### 3.2.7 Data Preprocessing (Multi Omic)

For created dataset for multi-omics also, we hope to do the data preprocessing. Here we hope to use follow the same procedure that we followed in single omic.

### 3.2.8 Test the models (Multi Omic)

Here also we hope to test the models using k-ford cross validation. And hope to use same prediction method used in single omic, in this process.

### 3.2.9 Finalize the model

In this step, we need to analyze the performances to choose the most suitable model for our study. The final results depend on the model we chose in this step. So we need to finalize the model considering the reliability.

## 3.3 Time line

Table 3: Timeline

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Weeks**  **Tasks** | Semester 06 | | | | | | | | | | | | | | | | | | | | Semester 07 | | | | | | | | | | | | | | | | Semester 08 | | | | | | | | |
| 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
| Literature review |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Bibliography writing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data collection |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Proposal writing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Model implementation  (Single omics) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Model implementation  (Multi omics) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Finalize the model (Multi omics) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Report writing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Research paper writing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# PROGRESS TO DATE

## 4.1 Literature Review

We referred from several research articles (all together more than 15 articles), books and educational websites which are related to our topic and have written the annotated bibliographies for some articles (10 the annotated bibliographies). The literature review will be conducted throughout the research.

## 4.2 Dataset Collection

We are considering 4 omics approaches to classify the kidney cancer subtype among 3 subtypes in our project. So, we downloaded 12 datasets, i.e., for each subtype (#3) four omic types were considered (3 \* 4 = 12). All datasets were downloaded from TCGA.

## 4.3 Dataset analysis

Now we are on dataset analyzing part. Since we are going to do a multi omic classification we have to consider several datasets. Now we are referring the datasets and analyzing them to do the preprocessing.

## 4.4 Research Proposal

With the knowledge of literature review we planned our future works and created our methodology. And we documented our previous works, ongoing works and our plans to achieve our target and here you are viewing it as our research proposal.

## 4.5 Implementations - Single Omic

We used 4 omic types in our research. For all omic type we follow almost same procedure.

### 4.5.1 Dataset Preparation

In the datasets, columns contained patients and rows contained the features. Since in TCGA, they divided the datasets for each subtype, in each dataset there were not a target row. So, we added a target row as ‘Subtype’ for each dataset with some integer values. Values can define as.

* ‘0’: KIRC
* ‘1’: KIRP
* ‘2’: KICH

We started to create our models in python 3 environment. There we imported all 3 (for each subtype) datasets.

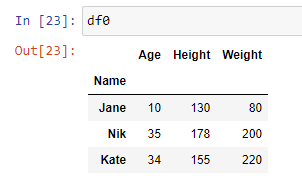
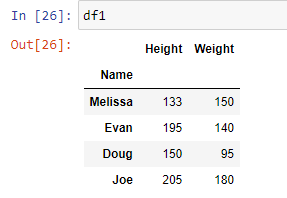
### 4.5.2 Preprocessing

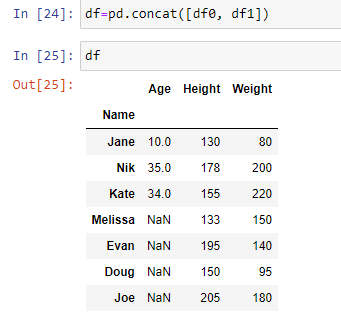
* **Dataset Integration**

We had to combine 3 datasets related to each subtype for create a model in relevant omics type. Since the features are not exactly same in 3 datasets, we could not use merge method. So we used *pandas.concat* method for combine the datasets.

In concat, it simply merges the datasets if the features (column names) are same and if not, it just adds the column. By adding columns which are not in other datasets, makes some ‘NULL’ values (missing values) to the dataset.

**Ex**: How *concat* works (note that, these data not related to our data)



Here concat order of the columns are not mattered. And if a column is not in both data frame, combined data frame has some missing values.

Table 4: Data Shapes after merging on each omic type

|  |  |  |
| --- | --- | --- |
| Omics type | Total Records | Features |
| Protein | 757 | 278 |
| miRNA | 721 | 2184 |
| Gene expression | 1020 | 20530 |
| DNA methylation | 867 | 485577 |

* **Dataset Cleaning**

In here we considered 3 main procedures.

* + Dropping duplicates
  + Handle null values
  + Handle zero values

Table 5: Cleaning process on each omic type

|  |  |  |
| --- | --- | --- |
| Cleaning procedure | Omics type | Description |
| Dropping duplicates | Protein | * No duplicates |
| miRNA | * No duplicates |
| Gene expression | * No duplicates |
| DNA methylation | * No duplicates |
| Handle null values | Protein | * Removed all null values containing features. |
| miRNA | * Removed features which has more than 20 null values. * Replaced the remaining null values by mean. |
| Gene expression | * No null values |
| DNA methylation | * Removed all null values containing features. |
| Handle zero values | Protein | * Very few zero values found. * Decided to remain them. |
| miRNA | * No zero values |
| Gene expression | * Removed the features which has more than 20 zero values |
| DNA methylation | * No zero values |

* **Feature Selection**

We got the correlation coefficients in dataset using Pearson correlation.

Then we got the highly correlated features with each other considering a threshold value which varies for different omics types.

* + Protein -> 0.8
  + miRNA -> 0.9
  + Gene expression -> 0.8
  + DNA methylation -> 0.9

Since the selected correlation values are very high, it means these are somewhat linearly dependent with other features. hence those effect on the dependent variable is almost same. So, we dropped one of them.

Then we got the absolute correlation of features with the target variable. Here we planned to test for best **100,50,20** features first. Then we changed the number of selecting features, considering the accuracy.

* **Sampling**

Initially all omic types are unbalance over each subtype. We handle these unbalancing using both Oversampling and Undersampling to get a fair and equal subtype value count.

Table 6: Records in each omic type before and after Sampling process on each omic type

|  |  |  |  |
| --- | --- | --- | --- |
| Omics type | Subtype | Data records (Before) | Data records (After) |
| Protein | KIRC | 478 | **100** |
| KIRP | 216 | **100** |
| KICH | 63 | **100** |
| miRNA | KIRC | 311 | **120** |
| KIRP | 321 | **120** |
| KICH | 89 | **120** |
| Gene expression | KIRC | 606 | **130** |
| KIRP | 323 | **130** |
| KICH | 91 | **130** |
| DNA methylation | KIRC | 480 | **100** |
| KIRP | 321 | **100** |
| KICH | 66 | **100** |

### 4.5.3 Testing the models

We used cross validation for test the models. There as the performance measurement, we used accuracy and standard deviation. For this test we used 3 prediction algorithms and process we followed in each algorithm is listed below.

Support Vector Machine (SVM)

* Kernel Selection (rbf, linear, poly) – Selected suitable kernel for each omic type
* Check Data transformation – Check for each omic type, whether it needs a transformation or not, and if needed what is the transformation method (standardization or normalization) better.
* Validate for different number of features – Run the model on various number of feature subsets (best 100,50,20.. features).

Random Forest (RF)

* Parameter Selection – mainly focused criterion, max\_depth and n\_estimators here.
* Validate for different number of features

K – Nearest neighbours (KNN)

* K Selection
* Check Data transformation
* Validate for different number of features

### 4.5.4 Extract Features

By testing the models, we can find the best model with best number of features. So here, we have extracted the best performed features considering each omics types. Selected feature counts are shown below.

* + Protein -> 170
  + miRNA -> 308
  + Gene expression -> 290
  + DNA methylation -> 60

## 4.6 Results - Single Omic

Table 7: Optimized parameters for each algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| Omics type | Algorithm | Parameters / Transformation | Used Values |
| Protein | SVM | Kernel | RBF |
| Data Transformation | standardization |
| RF | criterion | gini |
| max\_depth | 8 |
| n\_estimators | 1000 |
| KNN | k | 4 |
| Data Transformation | normalization |
| miRNA | SVM | Kernel | Linear |
| Data Transformation | standardization |
| RF | criterion | entropy |
| max\_depth | 8 |
| n\_estimators | 100 |
| KNN | k | 4 |
| Data Transformation | normalization |
| Gene expression | SVM | Kernel | Poly |
| Data Transformation | not required |
| RF | criterion | entropy |
| max\_depth | 8 |
| n\_estimators | 500 |
| KNN | k | 7 |
| Data Transformation | standardization |
| DNA methylation | SVM | Kernel | linear |
| Data Transformation | standardization |
| RF | criterion | entropy |
| max\_depth | 8 |
| n\_estimators | 1000 |
| KNN | k | 7 |
| Data Transformation | not required |

Table 8: Complete results for Protein

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of features** | **Cross Validation Score** | | |
| **SVM (Standardized)** | **RF** | **KNN (Normalized with min max scaler)** |
| All (170) | 0.93 (+/- 0.06) | 0.97 (+/- 0.03) | 0.76 (+/- 0.12) |
| 160 | 0.93 (+/- 0.05) | 0.96 (+/- 0.01) | 0.76 (+/- 0.08) |
| 150 | 0.93 (+/- 0.03) | 0.96 (+/- 0.02) | 0.77 (+/- 0.10) |
| 140 | 0.94 (+/- 0.04) | 0.96 (+/- 0.02) | 0.76 (+/- 0.12) |
| 120 | 0.91 (+/- 0.06) | 0.95 (+/- 0.03) | 0.77 (+/- 0.09) |
| 100 | 0.90 (+/- 0.04) | 0.94 (+/- 0.05) | 0.78 (+/- 0.11) |
| 60 | 0.89 (+/- 0.04) | 0.93 (+/- 0.03) | 0.79 (+/- 0.05) |
| 50 | 0.88 (+/- 0.07) | 0.92 (+/- 0.05) | 0.81 (+/- 0.11) |
| 40 | 0.85 (+/- 0.03) | 0.88 (+/- 0.04) | 0.81 (+/- 0.09) |
| 30 | 0.82 (+/- 0.04) | 0.87 (+/- 0.05) | 0.79 (+/- 0.07) |
| 20 | 0.78 (+/- 0.12) | 0.85 (+/- 0.12) | 0.73 (+/- 0.11) |

Table 9: Complete results for miRNA

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of features** | **Cross Validation Score** | | |
| **SVM (Without transformation)** | **RF** | **KNN (Normalized with min max scaler)** |
| All (308) | 0.96 (+/- 0.03) | 0.97 (+/- 0.03) | 0.95 (+/- 0.07) |
| 300 | 0.96 (+/- 0.05) | 0.96 (+/- 0.05) | 0.95 (+/- 0.07) |
| 290 | 0.96 (+/- 0.04) | 0.95 (+/- 0.06) | 0.95 (+/- 0.06) |
| 270 | 0.96 (+/- 0.03) | 0.96 (+/- 0.07) | 0.95 (+/- 0.06) |
| 160 | 0.96 (+/- 0.04) | 0.95 (+/- 0.06) | 0.95 (+/- 0.05) |
| 150 | 0.97 (+/- 0.04) | 0.94 (+/- 0.07) | 0.95 (+/- 0.07) |
| 140 | 0.96 (+/- 0.04) | 0.96 (+/- 0.07) | 0.95 (+/- 0.07) |
| 130 | 0.96 (+/- 0.03) | 0.95 (+/- 0.05) | 0.96 (+/- 0.06) |
| 120 | 0.96 (+/- 0.03) | 0.95 (+/- 0.05) | 0.96 (+/- 0.07) |
| 110 | 0.94 (+/- 0.04) | 0.95 (+/- 0.07) | 0.96 (+/- 0.06) |
| 100 | 0.94 (+/- 0.05) | 0.95 (+/- 0.06) | 0.95 (+/- 0.05) |
| 90 | 0.94 (+/- 0.05) | 0.94 (+/- 0.07) | 0.95 (+/- 0.05) |
| 50 | 0.94 (+/- 0.05) | 0.94 (+/- 0.06) | 0.95 (+/- 0.05) |
| 20 | 0.94 (+/- 0.02) | 0.92 (+/- 0.03) | 0.95 (+/- 0.06) |

Table 10: Complete results for methylation

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of features** | **Cross Validation Score** | | |
| **SVM (Without transformation)** | **RF** | **KNN (Normalized with min max scaler)** |
| All (375385) | 0.96 (+/- 0.06) | 0.95 (+/- 0.04) | 0.71 (+/- 0.03) |
| 170 | 0.93 (+/- 0.08) | 0.92 (+/- 0.05) | 0.94 (+/- 0.05) |
| 130 | 0.93 (+/- 0.04) | 0.94 (+/- 0.06) | 0.93 (+/- 0.04) |
| 120 | 0.93 (+/- 0.06) | 0.94 (+/- 0.05) | 0.93 (+/- 0.04) |
| 110 | 0.93 (+/- 0.08) | 0.94 (+/- 0.06) | 0.93 (+/- 0.06) |
| 100 | 0.93 (+/- 0.06) | 0.93 (+/- 0.05) | 0.93 (+/- 0.06) |
| 90 | 0.93 (+/- 0.03) | 0.93 (+/- 0.04) | 0.93 (+/- 0.05) |
| 70 | 0.93 (+/- 0.04) | 0.93 (+/- 0.06) | 0.93 (+/- 0.06) |
| 60 | 0.92 (+/- 0.07) | 0.92 (+/- 0.05) | 0.94 (+/- 0.04) |
| 50 | 0.93 (+/- 0.06) | 0.93 (+/- 0.05) | 0.94 (+/- 0.06) |
| 40 | 0.93 (+/- 0.07) | 0.92 (+/- 0.05) | 0.94 (+/- 0.07) |
| 30 | 0.93 (+/- 0.06) | 0.92 (+/- 0.08) | 0.94 (+/- 0.07) |
| 20 | 0.91 (+/- 0.07) | 0.92 (+/- 0.07) | 0.92 (+/- 0.08) |

Table 11: Complete results for Genomic

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of features** | **Cross Validation Score** | | |
| **SVM (Without transformation)** | **RF** | **KNN (Normalized with min max scaler)** |
| All (11,800) | 0.93 (+/- 0.03) | 0.93 (+/- 0.04) | 0.91 (+/- 0.07) |
| 300 | 0.93 (+/- 0.05) | 0.91 (+/- 0.07) | 0.91 (+/- 0.05) |
| 290 | **0.94 (+/- 0.06)** | 0.91 (+/- 0.07) | 0.90 (+/- 0.05) |
| 280 | 0.93 (+/- 0.05) | 0.91 (+/- 0.06) | 0.89 (+/- 0.06) |
| 270 | 0.93 (+/- 0.05) | 0.93 (+/- 0.05) | 0.89 (+/- 0.05) |
| 260 | 0.93 (+/- 0.05) | 0.93 (+/- 0.05) | 0.91 (+/- 0.07) |
| 250 | 0.93 (+/- 0.04) | 0.91 (+/- 0.07) | 0.91 (+/- 0.06) |
| 230 | 0.93 (+/- 0.04) | 0.90 (+/- 0.04) | 0.91 (+/- 0.06) |
| 200 | 0.92 (+/- 0.04) | 0.90 (+/- 0.04) | 0.91 (+/- 0.07) |
| 180 | 0.91 (+/- 0.05) | 0.90 (+/- 0.04) | 0.91 (+/- 0.03) |
| 170 | 0.90 (+/- 0.04) | 0.90 (+/- 0.04) | 0.93 (+/- 0.04) |
| 160 | 0.90 (+/- 0.04) | 0.91 (+/- 0.05) | 0.92 (+/- 0.06) |
| 150 | 0.91 (+/- 0.04) | 0.91 (+/- 0.07) | 0.91 (+/- 0.03) |
| 100 | 0.88 (+/- 0.06) | 0.90 (+/- 0.08) | 0.91 (+/- 0.07) |
| 50 | 0.88 (+/- 0.03) | 0.91 (+/- 0.07) | 0.89 (+/- 0.06) |
| 20 | 0.86 (+/- 0.05) | 0.88 (+/- 0.06) | 0.87 (+/- 0.08) |

REFERENCES

1. Muhamed Ali, Ali, Hanqi Zhuang, Ali Ibrahim, Oneeb Rehman, Michelle Huang, and Andrew Wu. 2018. "A Machine Learning Approach for the Classification of Kidney Cancer Subtypes Using miRNA Genome Data" Applied Sciences 8, no. 12: 2422. <https://doi.org/10.3390/app8122422>
2. Marquardt, André et al. “Subgroup-Independent Mapping of Renal Cell Carcinoma-Machine Learning Reveals Prognostic Mitochondrial Gene Signature Beyond Histopathologic Boundaries.” Frontiers in oncology vol. 11 621278. 15 Mar. 2021, doi:10.3389/fonc.2021.621278
3. Eloise Withnell, Xiaoyu Zhang, Kai Sun, Yike Guo, XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data, Briefings in Bioinformatics, Volume 22, Issue 6, November 2021, bbab315,<https://doi.org/10.1093/bib/bbab315>
4. Lovino, M., Bontempo, G., Cirrincione, G., Ficarra, E. (2020). “Multi-omics Classification on Kidney Samples Exploiting Uncertainty-Aware Models”. In: Huang, DS., Jo, KH. (eds) Intelligent Computing Theories and Application. ICIC 2020. Lecture Notes in Computer Science (), vol 12464. Springer, Cham. <https://doi.org/10.1007/978-3-030-60802-6_4>
5. Hu F, Zeng W, Liu X. A Gene Signature of Survival Prediction for Kidney Renal Cell Carcinoma by Multi-Omic Data Analysis. International Journal of Molecular Sciences. 2019; 20(22):5720. <https://doi.org/10.3390/ijms20225720>
6. He, Z., Liu, H., Moch, H. et al. Machine learning with autophagy-related proteins for discriminating renal cell carcinoma subtypes. Sci Rep 10, 720 (2020). <https://doi.org/10.1038/s41598-020-57670-y>
7. Eliana Marostica, Rebecca Barber, Thomas Denize, Isaac S. Kohane, Sabina Signoretti, Jeffrey A. Golden, Kun-Hsing Yu; Development of a Histopathology Informatics Pipeline for Classification and Prediction of Clinical Outcomes in Subtypes of Renal Cell Carcinoma. Clin Cancer Res 15 May 2021; 27 (10): 2868–2878. <https://doi.org/10.1158/1078-0432.CCR-20-4119>
8. Wu J, Jin S, Gu W, Wan F, Zhang H, Shi G, Qu Y and Ye D (2019) Construction and Validation of a 9-Gene Signature for Predicting Prognosis in Stage III Clear Cell Renal Cell Carcinoma. Front. Oncol. 9:152. doi: 10.3389/fonc.2019.00152
9. Terrematte, P.; Andrade, D.S.; Justino, J.; Stransky, B.; de Araújo, D.S.A.; Dória Neto, A.D. A Novel Machine Learning 13-Gene Signature: Improving Risk Analysis and Survival Prediction for Clear Cell Renal Cell Carcinoma Patients. Cancers 2022, 14, 2111. <https://doi.org/10.3390/cancers14092111>
10. Huang, Jennifer J, and James J Hsieh. “The Pan-Omics Landscape of Renal Cell Carcinoma and Its Implication on Future Clinical Practice.” Kidney cancer (Clifton, Va.) vol. 4,3 121-129. 16 Sep. 2020, doi:10.3233/KCA-200085
11. Chauhan, Nitin & Singh, Krishna. (2018). A Review on Conventional Machine Learning vs Deep Learning. 347-352. 10.1109/GUCON.2018.8675097.
12. Agarwal, Vivek. (2015). Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis. International Journal of Computer Applications. 131. 30-36. 10.5120/ijca2015907309.
13. Jianyu Miao, Lingfeng Niu, A Survey on Feature Selection, Procedia Computer Science, Volume 91, 2016, Pages 919-926, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2016.07.111>
14. Kayri, Murat. (2015). An Intelligent Approach to Educational Data: Performance Comparison of the Multilayer Perceptron and the Radial Basis Function Artificial Neural Networks. Educational Sciences: Theory and Practice. 15. 1247-1255. 10.12738/estp.2015.5.0238.
15. Häse, Florian & Galván, Ignacio & Guzik, Alan & Lindh, Roland & Vacher, Morgane. (2019). How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry. Chemical Science. 10. 10.1039/C8SC04516J.
16. Syed A Hoda, MD, Esther Cheng, DO, Robbins Basic Pathology, American Journal of Clinical Pathology, Volume 148, Issue 6, December 2017, Page 557, <https://doi.org/10.1093/ajcp/aqx095>
17. Chen, Hongge & Boning, Duane. (2019). Machine Learning Approaches for IC Manufacturing Yield Enhancement. 10.1007/978-3-030-04666-8\_6.