**MULTI-OMIC DATA IN THE CERVICAL CANCER STUDIES**

UNDERGRADUATE RESEARCH PROPOSAL SUBMITTED

IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF

BACHELOR OF THE SCIENCE IN ENGINEERING

**Submitted by:**

Perera G.A.N.C (2019/E/094)

Sachinthani W.M.R. (2019/E/117)

**DEPARTMENT OF COMPUTER ENGINEERING**

**FACULTY OF ENGINEERING**

**UNIVERSITY OF JAFFNA**

31 MAY 2023

**MULTI-OMIC DATA IN THE CERVICAL CANCER STUDIES**

Supervisor(s):

Supervisor : Dr. P. Jeyananthan

External Supervisor :

Examination Committee:

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

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

**CONTRIBUTION TO THE PROPOSAL BY THE MEMBERS IN GROUP**

|  |  |  |
| --- | --- | --- |
| Sections | 2019/E/094 | 2019/E/117 |
| **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 Forecasting Models & Prediction Models | ✓ |  |
| 2.2.1 Forecasting Models | ✓ |  |
| 2.2.2 Prediction Models | ✓ |  |
| 2.3 Performance Analysis | ✓ |  |
| 2.4 Available Databases | ✓ |  |
| 2.5 Research Gap |  | ✓ |
| **CHAPTER 3 : METHODOLOGY AND RESEARCH PLAN** |  |  |
| 3.1 Methodology in Brief | ✓ | ✓ |
| 3.2 Detailed Methodology | ✓ | ✓ |
| 3.2.1 Data selection | ✓ |  |
| 3.2.2 Data preprocessing |  | ✓ |
| 3.2.3 Feature selection | ✓ | ✓ |
| 3.2.4 Apply machine learning methods | ✓ | ✓ |
| 3.2.5 Compare performance | ✓ |  |
| 3.2 Timeline | ✓ |  |
| **CHAPTER 4: PROGRESS TO DATE** |  |  |
| 4.1 Literature Review |  | ✓ |
| 4.2 Database Collection | ✓ | ✓ |
| 4.2.1 Phenotype Data selection | ✓ |  |
| 4.2.2 Dataset Selection | ✓ |  |
| 4.3 Database Preparation | ✓ | ✓ |
| **REFERENCE** | ✓ | ✓ |

## TABLE OF CONTENT

[TABLE OF CONTENT 4](#_Toc138070746)

[LIST OF FIGURES 5](#_Toc138070747)

[LIST OF TABLES 6](#_Toc138070748)

[Chapter 1: INTRODUCTION 8](#_Toc138070749)

[1.1 Motivation and Overview 8](#_Toc138070750)

[1.2 Aims and Objectives 8](#_Toc138070751)

[1.3 Research Scope 9](#_Toc138070752)

[Chapter 2: LITERATURE REVIEW 10](#_Toc138070753)

[2.1 Introduction 10](#_Toc138070754)

[2.2 Forecasting Models & Prediction Models 12](#_Toc138070755)

[2.2.1 Forecasting Models 12](#_Toc138070756)

[2.2.2 Prediction Models 12](#_Toc138070757)

[2.3 Performance Analysis 18](#_Toc138070758)

[2.4 Available Databases 23](#_Toc138070759)

[2.5 Research Gap 25](#_Toc138070760)

[Chapter 3: Methodology and Research Plan 26](#_Toc138070761)

[3.1 Methodology in Brief 26](#_Toc138070762)

[3.2 Detailed Methodology 27](#_Toc138070763)

[3.2. Data selection of cervical cancer 27](#_Toc138070764)

[3.2.2 Data preprocessing 28](#_Toc138070765)

[3.2.3 Feature Selection 30](#_Toc138070766)

[3.2.4 Apply machine learning methods 32](#_Toc138070767)

[3.2.5 Compare performance 32](#_Toc138070768)

[3.3 Timeline 34](#_Toc138070769)

[Chapter 4: PROGRESS TO DATE 35](#_Toc138070770)

[4.1 Literature Review 35](#_Toc138070771)

[4.2 Database Collection 35](#_Toc138070772)

[4.2.1 Phenotype Data selection 35](#_Toc138070773)

[4.2.2 Data set selection 35](#_Toc138070774)

[4.3 Database Preparation 36](#_Toc138070775)

## LIST OF FIGURES

[Figure 1 : Random forest algorithm workflow 14](#_Toc136351331)

[Figure 2 : The workflow of data preprocessing and prediction model development. A-The workflow of data preprocessing. B- The workflow of miRNAs based machine learning cervical survival prediction model development 15](#_Toc136351332)

[Figure 3 : Decision tree workflow 16](#_Toc136351333)

[Figure 4 : Overview of the Methodology 25](#_Toc136351334)

[Figure 5 : Data preprocessing workflow 28](#_Toc136351335)

## LIST OF TABLES

[Table 1 : Performance analysis of models in study 3 17](#_Toc136353469)

[Table 2 : Confusion matrix for SVM (1 indicates as tumor while 0 indicates as normal) 17](#_Toc136353470)

[Table 3 : Confusion matrix for RF (1 indicates as tumor while 0 indicates as normal) 18](#_Toc136353471)

[Table 4 : Performance analysis of models in study 11 19](#_Toc136353472)

[Table 5 : Performance analysis of models in study 10 20](#_Toc136353473)

**ABBREVIATIONS AND ACRONYMS**

HPV: Human papillomavirus

DNA: Deoxyribonucleic acid

miRNA: microRNA expression data

ICC: Invasive Cervical Cancer

CNA: Copy number alterations

SVM: support vector machine

VC: Vapnik-Chervonenkis

PCA: Principal Components Analysis

RF: Random forest

KNN: K-Nearest Neighbor

RS: Remote Sensing

AUC: Area under curve

DT: Decision tree

ASM: Attribute selection measures

CART: Classification and Regression Tree

TN: True Negative

TP: True positive

FP: False positive

FN: False negative

TCGA: The Cancer Genome Atlas data repository

SNP: Single-nucleotide polymorphism

GEO: Gene Expression Omnibus database

RPPA: Reverse phase protein array

INDELs: Insertions/Deletions

FS: Feature selection

ML: Machine Learning

ROC: Receiver Operating Characteristic

# Chapter 1: INTRODUCTION

## Motivation and Overview

## Cancer can be identified as a disease when abnormal cells grow rapidly that don't have usual cell growing boundaries.[6]Cervical cancer is the most commonly diagnosed cancer which occurs in a woman’s cervix and because of cervical cancer, 260,000 women die annually, with over 520,000 new cases reported worldwide. [3][9][6]Most cases of cervical cancer begin increasing in women between the ages of 20 and 29, reach their highest between the ages of 55 and 64, and then begin decreasing after the age of 65.[11]The high cervical cancer death rate is mostly a result of poor knowledge of early detection methods and bound access to healthcare.[2]

## The primary risk factor for cervical cancer, which affects more than 500,000 women annually worldwide, is human papillomavirus (HPV) infection. Immunosuppression, smoking, a history of pregnancies, being married before turning 18, having children before turning 18, having several sexual partners in addition to long-term birth control pill use are additional risk factors.[3][5][11]Additionally, one in five Indian women was discovered to have a cervical illness, and the country has the highest number of cervical cancer patients.[11]

## The two most common kinds of cervical cancer currently are squamous cell carcinoma, which accounts for 90% of cases, and adenocarcinoma, which accounts for 10% of cases.[3]Surgery is the main form of treatment for cervical cancer in its early stages. Radiotherapy alone or combined with chemotherapy is also required for patients with a high risk of recurrence.[9]Commercially accessible vaccinations against the highly carcinogenic human papillomavirus (HPV) are also available, although the proportion of women who get the vaccine is still low, especially in developing countries.[6]

## Studies with machine learning have used molecular data to predict cervical cancer outcomes, however many of these studies have concentrated on single-omic data or had trouble combining several datasets.[9]Our goal in this research proposal is to fill this knowledge gap by developing a prediction model that combines machine learning techniques with data from several biological data sources, including methylation and gene expression.

## 1.2 Aims and Objectives

Our research aim is to help the medical society to identify the cervical cancer subtypes using multi omics data of the patients.

The main objectives of our research is,

* To study the factors could be influence on the performance of each machine learning methods and create an accurate machine learning model for predicting the subgroup of the cervical cancer patients.
* To study the influence of the different features on predicting sub group of the cervical cancer.
* Studying the difference between omic data and building a more accurate machine learning model by taking advantage of using multi-omic data instead of single-omic data

## Research Scope

Our scope to develop a better model to classify cervical cancer subgroups. To achieve that task, we followed some constrains. Focused on three omics approaches known as DNA methylation, gene expression and miRNA due to dataset limitation.

# Chapter 2: LITERATURE REVIEW

## 2.1 Introduction

Although the Human Papillomavirus (HPV) infection is a necessary component of Invasive Cervical Cancer (ICC), a large percentage of infections clear spontaneously.[2] Even though numerous studies have demonstrated that HPV screening is more sensitive than the Pap test in the detection of high-grade cervical carcinoma. HPV testing is unable to differentiate between true triggers and associated factors to offer an accurate prediction, as not all subtypes of HPV infection can result in cervical cancer.[2] Considerable evidence links distinct genetic changes to cervical cancer's beginning and development [1], and molecular markers have been shown to enhance the accuracy of risk assessment, early diagnosis, and prognosis prediction.[2]

The extensive complexity of molecular alterations, including somatic aberrations, CNA, DNA methylation, and dysfunctional microRNA (miRNA), has been revealed by the genome characterization of the large number of cervical patients. These molecular alterations imply significant information about cancer cells, including malignant level, metastasis ability, and therapeutic sensitivity.[1][6] In recent years, the availability of multi-omic data has increased due to high throughput methods and steadily falling sequencing costs.[4] Omics data collected from various phases of gene and protein expression are likely to provide complementary information because different molecular processes beginning on DNA and resulting in protein products are closely interconnected.[4] The use of multi-omics data analysis has emerged as an effective way to improve patient outcomes and expand our understanding of the illness.

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

* miRNA Genome Data [6][7]
* gene expression profiling data[3]
* transcriptomic data[13]
* Genomics[14]
* DNA Methylation[9]
* Proteins[15]

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:

* DNA methylation and gene expression data : Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers[2]
* gene expression, miRNA expression, copy number alteration, and DNA methylation data : Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment [4]
* MRNA and gene expression data: The early detection of cervical cancer. The current and changing landscape of cervical disease detection[5]
* copy number variation and gene expression data: Integrated analysis of chromosome copy number variation and gene expression in cervical carcinoma [8]

Many biological processes, such as the regulation of gene expression, genomic imprinting, cell differentiation, development, and inflammation, depend on DNA methylation.[2][16] miRNAs that are only examined in advanced disease, or in relation to prognosis or therapy, or in research based on functional experiments, miRNA target genes, HPV, or miRNA methylation, or that only describe computational methods.[7] DNA methylation controls the transcription of genes without affecting to the DNA sequence. DNA methylation is relatively stable as compared to the complex variation in RNA expression between different patients.[9]

As a result, we use machine learning models to focus on cervical cancer sub-typing utilizing multi-omics data, including DNA methylation, gene expression, and miRNA data.

## 2.2 Forecasting Models & Prediction Models

The terms "prediction models" and "forecasting models" are often used interchangeably, but there is a subtle distinction between them. While both involve making predictions about future outcomes, there are differences in their applications and underlying assumptions.

### 2.2.1 Forecasting Models

Forecasting models are a subset of prediction models that specifically focus on making predictions for time series data. Time series data represents observations collected over time, where the order and timing of the data points are essential. Forecasting models aim to capture and account for the temporal patterns, trends, seasonality, and other time-related factors present in the data. These models consider the time dependencies explicitly, enabling them to make more accurate predictions for future time points.

### 2.2.2 Prediction Models

Prediction models are used to estimate or predict future values based on historical data. They aim to capture patterns and relationships in the data to make accurate predictions. Prediction models can be applied to various types of data, including time series data, but they may not explicitly consider time dependencies or trends. These models focus on generalizing patterns observed in the data to make predictions without necessarily considering the underlying temporal aspects.

For cancer prediction, a prediction model would be more appropriate than a forecasting model. Cancer prediction typically involves analyzing various factors or features, such as genetic markers, patient demographics, medical history, and test results, to assess the likelihood or risk of developing cancer. So, for this section we only added the models which were focused the classification of cervical cancer subtype. They used both conventional machine learning models in their studies.

Here are the conventional models that we identified in literatur

* Support vector machine (SVM)

Support Vector Machine (SVM) is a powerful prediction method widely used for classification or regression analysis based on statistical learning frameworks and Vapnik-Chervonenkis (VC) theory.[6] SVM offers a direct approach to binary classification, treating data as points in a high-dimensional space and finding an optimal hyperplane to separate the data into distinct classes. The SVM model utilizes the recurrence outcomes for training.[9]

One of the key advantages of SVM is its ability to handle both linear and nonlinear data. In the case of nonlinear data, SVM employs techniques like the radial base function and kernel functions. The kernel function plays a crucial role in transforming lower-dimensional data into higher-dimensional space, enabling effective differentiation between different classes.[10] Furthermore, SVM is particularly well-suited for complex small or medium-sized classification datasets. By separating data points with a hyperplane determined by the kernel, SVM can handle scenarios where a scatter plot of multiple variables fails to effectively separate two or more data classes.[10]

In this study, the researchers applied the SVM model to their dataset, utilizing Principal Components Analysis (PCA) algorithm as a feature selection technique to address high dimensionality and enhance the model's performance. The goal was to achieve optimal prediction accuracy by executing the SVM model on the constructed dataset.[3]

* Random forest (RF)

In their research on predictive models for cervical cancer using gene expression profiling data, the authors employed two machine learning algorithms, namely SVM and Random Forests (RF). The objective was to evaluate the performance of these algorithms and determine which one yielded better result. The dataset consisted of 714 features and 58 samples, and the researchers aimed to develop a predictive model for cervical cancer. Their computational analysis revealed that the RF algorithm outperformed the SVM algorithm with an accuracy of 94.21%.[3]

Random Forest (RF) is a type of ensemble learning that enhances model performance by utilizing multiple learners. By employing the RF bagging method, the algorithm reduces the impact of outliers and is suitable for both categorical and continuous data. Unlike other algorithms, RF does not require dataset scaling. The decision-making process in RF involves voting, where the majority decision across multiple decision trees is considered the final decision.[10]

The RF classifier is a supervised classification algorithm known for its adaptability and ease of implementation. Figure 4 illustrates the workflow of the RF classifier. RF construct decision trees by randomly selecting data, obtaining predictions from each tree, and aggregating the results through voting. The algorithm follows a divide and conquer approach, where decision trees are created by randomly partitioning the dataset and employing feature selection measures for each attribute. Each decision tree operates on an independent sample, and in a classification problem, the class with the highest number of votes is selected as the final prediction.[11]

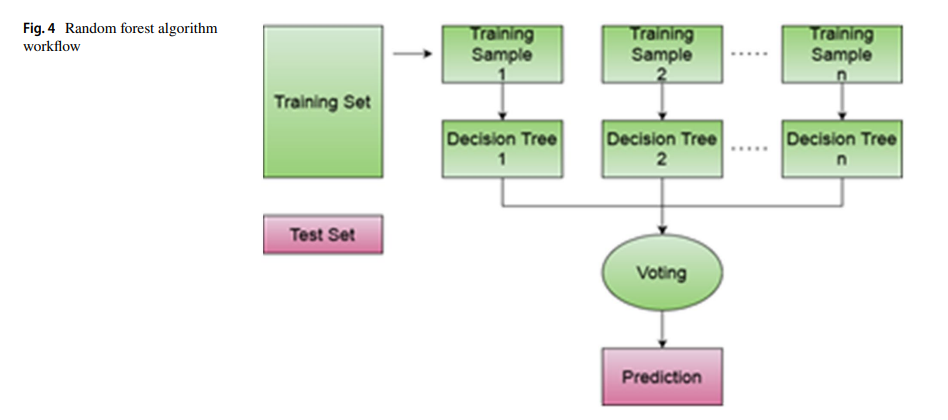


Figure 1 : Random forest algorithm workflow

* k-nearest neighbors (KNN)

In their study, the authors addressed missing numeric feature values in the clinical and proteomics datasets by employing the KNN method with k = 1.[4] They recognized the need to consider both batch effects and subtypes of cancer samples derived from individual heterogeneity in the missing values imputation process.

To handle the missing values, the researchers performed two independent steps. Firstly, KNN imputation was applied within each batch, where KNN were determined based on physical units.[6] If a specific batch did not contain a sufficient number of samples of a certain subtype, the K nearest neighbors were selected from that particular batch. After the removal of batch effects using the quantile normalization algorithm, missing values were again replaced using KNN imputation, this time within each molecular subtype. The data was further processed through log2 transformation and feature scaling using the Z-score algorithm.[6]

In the field of cervical malignancy risk factor exploration, different machine learning algorithms including Decision Tree, Logistic Regression, SVM, K-NN, and Multilayer Perceptron were utilized, resulting in accuracies of 77.97%, 82.78%, 79.25%, 82.93%, and 83.16% respectively. In another study, Nithya et al. developed a cervical disease detection system using various algorithms, including SVM, C5.0, r-part, RS, and K-NN, with accuracies of 97%, 96.9%, 96%, 88%, and 88% respectively.[11] They selected 25 features for K-NN, 17 features for the decision tree, and 11 features for the random forest classifier, with K-NN exhibiting the highest precision and an AUC of 0.82.[11]

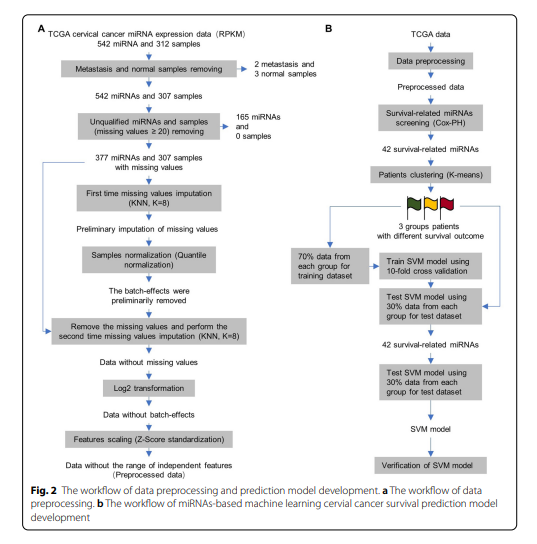


Figure 2 : The workflow of data preprocessing and prediction model development. A-The workflow of data preprocessing. B- The workflow of miRNAs based machine learning cervical survival prediction model development

* Decision tree algorithm (DT)

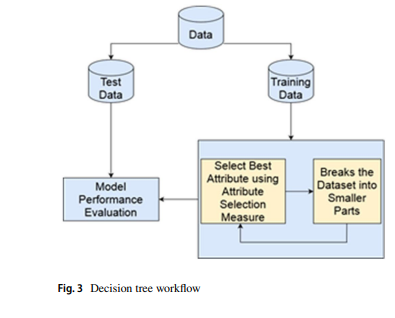


Figure 3 : Decision tree workflow

The decision tree algorithm, known as DT, is a supervised learning classification algorithm widely used in various domains.[11] The workflow of the decision tree algorithm can be observed in Figure 3, illustrating its hierarchical tree structure. Each inner node represents a feature, branches depict decision conditions, and leaf nodes represent outcomes. The root node serves as the starting point, where attribute selection measures (ASM) are employed to identify the best split for the records. The dataset is then divided into smaller subsets based on these splits, and the tree construction process continues recursively for each child until specific conditions are met.[11]

Decision trees offer an intuitive flowchart-like representation that closely resembles human-level reasoning. The interpretability and ease of decoding are advantageous features of decision trees. Moreover, decision trees can handle both classification and regression problems. The Classification and Regression Tree (CART) algorithm, also referred to as DT, exhibits a tree-like structure resembling branches, mirroring the growth of an actual tree. The decision nodes in the tree correspond to different decision conditions, allowing the algorithm to make informed choices based on these conditions.[10]

## 2.3 Performance Analysis

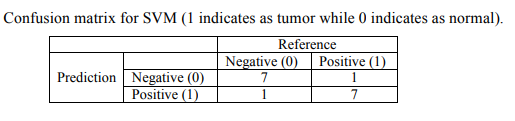
In this study, our primary focus is on analyzing the performance methods and evaluating the effectiveness of the models discussed previously. The performance analysis results of these models are presented in Table 1, allowing for a clear comparison of the performance of conventional models.

Table 1 : Performance analysis of models in study [3]

|  |  |  |  |
| --- | --- | --- | --- |
| Study | ML model | Performance Analysis | |
| **Method** | **Performance** |
| [3] | SVM | Accuracy | 87.50% |
| Kappa | 75% |
| Recall | 87.50% |
| RF | Accuracy | 94.12 % |
| Kappa | 87.59% |
| Recall | 100.00% |

These coefficients are defined and illustrated as a confusion matrix Table 2

Table 2 : Confusion matrix for SVM (1 indicates as tumor while 0 indicates as normal)



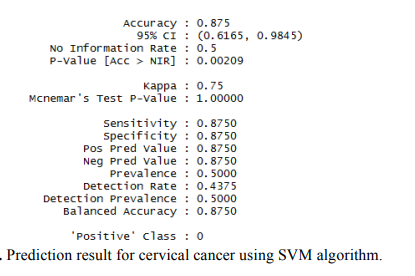
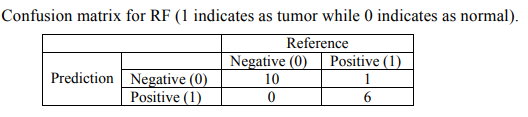


Table 3 : Confusion matrix for RF (1 indicates as tumor while 0 indicates as normal)



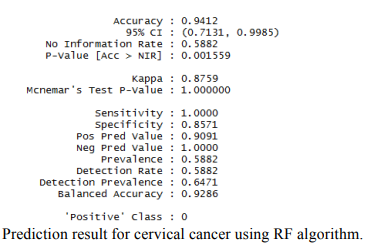


Table 4 : Performance analysis of models in study [11]

|  |  |  |  |
| --- | --- | --- | --- |
| Study | ML model | Performance Analysis | |
| **Method** | **Performance** |
| [11]  Error! Reference source not found. | Decision Tree | Accuracy | 93.33% |
| Precision | 80.00% |
| Recall | 100.00% |
| F1-score | 89.00% |
| Random forest | Accuracy | 93.33% |
| Precision | 100.00% |
| Recall | 75.00% |
| F1-score | 86.00% |
| XGBoost | Accuracy | 93.33% |
| Precision | 00.00% |
| Recall | 00.00% |
| F1-score | 00.00% |
| Naïve Bayes | Accuracy | 91.67% |
| Logistic regression | Accuracy | 87.50% |

Table 5 : Performance analysis of models in study [10]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Study | ML model | Negative class (0)/ Positive class (1) | Performance Analysis | |
| **Method** | **Performance** |
| [10] | Logistic Regression | 0 | Precision | 98% |
| Recall | 100% |
| F1 scores | 99% |
| 1 | Precision | 100% |
| Recall | 77% |
| F1 scores | 87% |
| Accuracy | 98% |
| SVM | 0 | Precision | 99% |
| Recall | 100% |
| F1 scores | 100% |
| 1 | Precision | 100% |
| Recall | 92% |
| F1 scores | 96% |
| Accuracy | 99% |
| Random Forest | 0 | Precision | 100% |
| Recall | 100% |
| F1 scores | 100% |
| 1 | Precision | 100% |
| Recall | 100% |
| F1 scores | 100% |
| Accuracy | 100% |
| Decision Tree | 0 | Precision | 100% |
| Recall | 100% |
| F1 scores | 100% |
| 1 | Precision | 100% |
| Recall | 100% |
| F1 scores | 100% |
| Accuracy | 100% |
| Adaptive Boosting | 0 | Precision | 100% |
| Recall | 100% |
| F1 scores | 100% |
| 1 | Precision | 100% |
| Recall | 100% |
| F1 scores | 100% |
| Accuracy | 100% |
| KNN | 0 | Precision | 95% |
| Recall | 100% |
| F1 scores | 97% |
| 1 | Precision | 100% |
| Recall | 31% |
| F1 scores | 47% |
| Accuracy | 95% |

TN is the number of true negative,

FP is defined as the number of false positive,

TP is defined as number of true positive

FN is the number of false negative.

To obtain the classification report, the following Equations (1) – (4) are used.[10]

(1)

Recall/sensitivity: Positivity is represented by the ratio of accurate to inaccurate predictions. It is written in mathematical notation as follows:

(2)

(3)

(4)

## 2.4 Available Databases

The mutant MAF file of cervical cancer.[1]

* Downloaded using the R package TCGA bio links
* Contains the mutation results of 297 samples

Single-nucleotide polymorphism (SNP) 6 copy number segment 287 datasets, and 299 methylation chip data of cervical cancer samples.[1]

* Downloaded from FireBrowse ([http://firebrowse.org/)](http://firebrowse.org/)%20) with Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (platform for Illumina 450K chip).

304 messenger RNA (mRNA) expression profile data and 307 miRNA expression profile data of cervical cancer samples.[1]

* Downloaded from the National Cancer Institute Genomic Data Commons Data Portal (<https://portal.gdc.cancer.gov/>).

Cervical cancer fusion genes.[1]

* downloaded from the Tumor Fusion Gene Data Portal ([https://tumorfusions.org/ PanCanFusV2/database](https://tumorfusions.org/%20PanCanFusV2/database))

Illumina HumanMethylation450K array data and RNA-seq expression profiles from cervical cancers dataset.[2]

* Obtained from The Cancer Genome Atlas (TCGA).
* Data from 307 cervical tumors and 3 associated normal tissues

Illumina 450K methylation datasets.[2]

* Downloaded from the GEO (Gene Expression Omnibus database) database.
* GSE38266: Methylation data from 21 HPV-positive and 21 HPV-negative tumors
* GSE46306: Data from 20 normal cervical samples (HPV-negative) and 6 cervical cancer tissues (HPV-positive)
* GSE68339: Methylation profiles from a discovery cohort of 149 cervical cancer patients (GSE68339\_149) and a validation cohort of 121 cancer patients (GSE68339\_121)

The gene expression profiling data for classification of cervical cancer [3]

* obtained from the Gynecologic Oncology Group Tissue Bank (PA, USA)

MiRNAs expression data used as features for the development of Cervical Cancer Subtyping Prediction Model.[6]

* Obtained from The Cancer Genome Atlas (TCGA).
* The dataset included 542 miRNAs in 312 cervical cancer samples.

DNA methylation datasets, RNA expression datasets, and clinical information for patients with cervical cancer dataset.[9]

* Downloaded from TCGA (https://portal.gdc.cancer.gov/) using the FireBrowse Data Portal (http://firebrowse.org/).
* The DNA methylation profiles were obtained using Infinium HumanMethylation450 BeadChip sequencing technology (Illumina, San Diego, CA, USA) with a total of 312 profiles.
* The mRNA-seq profiles were analyzed using HiSeq2000 (Illumina) with a total of 307 profiles.

Cervical Cancer Risk Factors for Biopsy dataset [10]

* Contributed to the dataset by the UCI repository.
* The dataset comprises 858 instances, each with 32 properties.

Cervical Cancer Behavior Risk dataset [11]

* Obtained from The UCI (University of California, Irvine) Machine Learning Repository.
* Includes 32 risk factors and 4 target variables of the clinical history of 858 patients.

## 2.5 Research Gap

Their objective was to develop a novel molecular features-based machine learning cervical cancer survival prediction model (CCSPM) with high performance. They have chosen only one omic data which is MiRNA as features obtained from the TCGA dataset.

In our study, we are going to use multi-omics data to predict the subgroups of cervical cancer. So we can get higher accuracy than them because we are using multi-omics data instead of using single omic data.

# Chapter 3: Methodology and Research Plan

## 3.1 Methodology in Brief



Figure 4 : Overview of the Methodology

## 3.2 Detailed Methodology

The methodology can be divided into six steps, as illustrated in Figure 4. The following section provides a detailed explanation of each step along with relevant examples.

### 3.2. Data selection of cervical cancer

For our research, we selected The Cancer Genome Atlas (TCGA) data repository as our primary data source. TCGA is a comprehensive collection of data on 33 different cancer types, including detailed patient phenotypic information.[4] Specifically, we focused our study on cervical cancer.

The phenotype data within the TCGA repository contains a vast array of clinical information related to the patients. Among the numerous variables available, we identified two key factors of interest:

1. Cancer stage: This variable provides crucial information about the extent and progression of the cancer in each patient.
2. Cancer subgroups: We considered the classification of cervical cancer into two major subtypes, namely adenosquamous and cervical squamous cell carcinoma.

Within the dataset specific to cervical cancer, we identified various molecular data types that would be relevant for our study. These include:

1. Copy number (gene-level): This data captures alterations in gene copy numbers, providing insights into genomic amplifications or deletions.
2. DNA methylation: These data provide information about epigenetic modifications in DNA, which can influence gene expression patterns.
3. Gene expression RNAseq: These data capture the overall gene expression profiles, allowing us to analyze differential expression patterns.
4. miRNA mature strand expression RNAseq: This dataset focuses on the expression levels of mature microRNA strands, which play a role in gene regulation.
5. Protein expression RPPA: These data provide information about protein expression levels, allowing us to examine protein-based biomarkers.
6. Somatic mutation (SNP and INDEL): This dataset captures somatic mutations in single-nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs), shedding light on genetic alterations specific to cervical cancer.

### 3.2.2 Data preprocessing

Data pre-processing is an important step in the data mining process. It is implemented on the dataset because the dataset obtained from microarray analysis contains irrelevant, unreliable and redundant of data or noise present in the dataset.[3] The data preprocessing was first performed; data preprocessing plays important roles for statistical analysis of big data, including elimination of the impact from unquantified samples and features, missing values and outliers, reduction of batch-effects and experimental deviation, and normalization of the range of independent features.[6]

In machine learning, data preprocessing techniques are essential for preparing raw data before it can be used for model training and analysis. Several key techniques are commonly employed in this process, including data cleaning, data integration, data transformation, data reduction, and data discretization (Figure 5). Let's explore each technique in detail which we hope to use for our research are given below:

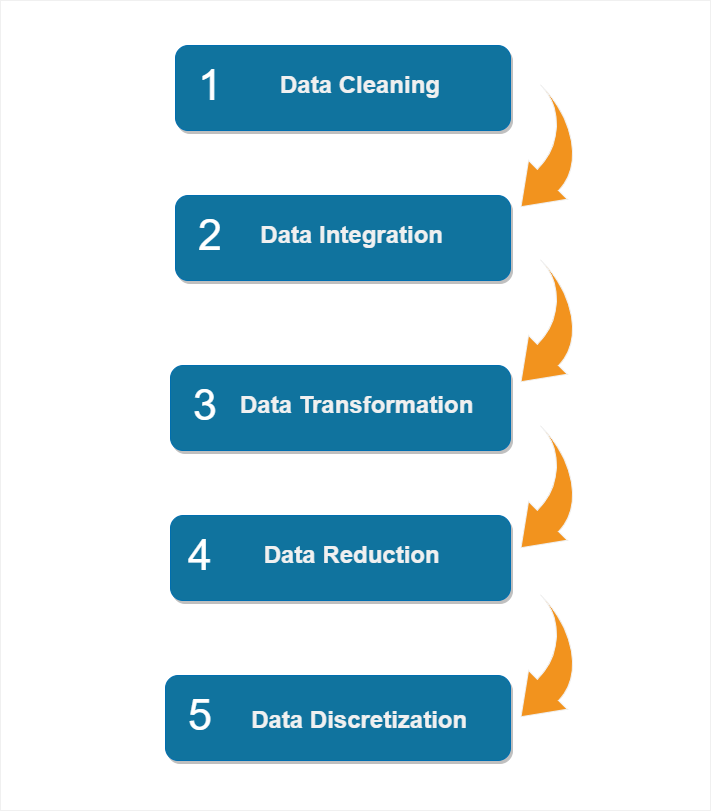


Figure 5 : Data preprocessing workflow

1. Data Cleaning:

* Handling missing values
* Dealing with outliers
* Correcting inconsistencies in data formats
* Resolving conflicting or duplicate entries

1. Data Integration:

* Combining data from multiple sources
* Resolving schema and attribute conflicts
* Harmonizing data representations
* Handling data with different structures or levels of granularity

1. Data Transformation:

* Normalization, scaling, or standardization of numerical features
* Applying logarithmic or power transformations
* Addressing skewness or nonlinear relationships in the data

1. Data Reduction:

* Dimensionality reduction techniques (e.g., PCA, feature selection)
* Eliminating redundant or irrelevant features
* Enhancing computational efficiency
* Preventing overfitting in machine learning models

1. Data Discretization:

* Converting continuous numerical variables into categorical or ordinal variables
* Dividing value ranges into intervals or bins
* Simplifying data representation
* Enabling the use of algorithms designed for categorical data

### 3.2.3 Feature Selection

Feature selection is a process in which the features that contribute more to the estimated predictor variable are automatically selected from the data. Feature selection (FS) methods can be used in data pre-processing to accomplish effective data reduction and this is suitable for finding accurate data models. Selecting appropriate features in the data are important, since irrelevant features can decrease the accuracy of many models. We need not use every feature present in the data for creating an algorithm. We can train our algorithm with those features that are certainly important and it will authorize improved results than using complete set of features for the same algorithm.[10] [12]

Advantages of using feature selection

• Allows the ML procedure to train the model more rapidly

• Reduces model complexity with an ease of interpretation

• Advances the precision of a model when the precise subset is selected

• Decreases overfitting

**Filter methods**

Filter methods are commonly employed as a preliminary step in data preprocessing within the context of research papers. These methods involve the selection of features based on their correlation with the outcome variable, determined through statistical tests. The significance of features is evaluated by examining their correlation with the dependent variable. Figure 6 illustrates the process of feature selection using filter methods. Compared to wrapper methods, filter methods are known for their significantly faster execution.[12]



Figure 6 : Filter methods workflow

In feature selection, there are three types of methods: filter methods, wrapper methods, and embedded methods. For our research, we have chosen to use wrapper methods. Wrapper methods are a type of feature selection approach that evaluates the performance of a model with different subsets of features. They involve iterative selection or elimination of features based on the model's performance. Two common examples of wrapper methods are forward feature selection and backward feature elimination.

• Correlation Matrix - Pearson correlation:

The Pearson correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where -1 represents a perfect negative linear correlation, 0 indicates no correlation, and +1 indicates a perfect positive linear correlation.

• Forward feature selection:

Forward feature selection is an iterative process that starts with an empty set of features. In each iteration, it adds one feature to the existing set and evaluates the model's performance. The feature that contributes the most to the performance improvement is selected and added to the feature set. This process continues until a desired number of features is reached or no further improvement is observed.

• Backward feature elimination:

Backward feature elimination is the reverse of forward feature selection. It begins with all features included and iteratively removes one feature at a time. After removing each feature, the model's performance is evaluated, and the feature with the least impact on the performance is eliminated. This process continues until a desired number of features is reached or further elimination does not significantly affect the model's performance.

### 3.2.4 Apply machine learning methods

The research project involves the application of machine learning (ML) algorithms for prediction purposes. Four types of ML algorithms are commonly used: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. For this research, the focus is on supervised learning, specifically classification-based algorithms. From the available algorithms in this category, the three most popular ones are selected for the project based on their performance in previous studies. These algorithms are:

* Random forest algorithm[3][11]
* Support Vector Machine (SVM) [3]
* K-Nearest Neighbor (KNN) algorithm[10][6][4]
* Decision Tree[11]

### 3.2.5 Compare performance

In our research, we aim to evaluate the effectiveness of multiple machine learning (ML) algorithms by creating test datasets and assessing their performance using train datasets. Our objectives include,

* selecting the best ML algorithm and
* determining the subtype

That can be classified more efficiently. To achieve these goals, we will utilize specific performance measurements, namely:

* Confusion matrix:

The confusion matrix is a valuable technique for summarizing the performance of a classification algorithm. By calculating a confusion matrix, we can gain insights into the model's accuracy and identify the types of errors it may be making.

* Area under the ROC curve (AUC):

The AUC-ROC curve is a widely-used performance measurement for classification problems. It represents the degree of separability between classes and provides an indication of how well the model can distinguish between different classes. A higher AUC value signifies a better model at correctly predicting 0 classes as 0 and 1 classes as 1. In other words, it reflects the model's ability to differentiate between patients with and without the disease.

* Accuracy:

Accuracy is a commonly used metric for evaluating classification models. It quantifies the fraction of correct predictions made by the model. More formally, accuracy is defined as the number of correctly predicted instances divided by the total number of predictions. And

## 3.3 Timeline

.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weeks | Semester 06 | | | | | Semester 07 | | | | | Semester 08 | | | |
| Tasks | 1-3 | 4-6 | 7-9 | 10-12 | 13-15 | 1-3 | 4-6 | 7-9 | 10-12 | 13-15 | 1-2  3-4 | 3-4 | 5-6 | 7-8 |
| **1** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **2** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **3** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **4** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **5** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **6** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **7** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **8** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **9** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1 - Literature review

2- Bibliography writing

3 - Proposal writing

4 - Data collection

5 - Data preparation

6 - Finalize the model

7 - Model implementation

8 - Report writing

9 - Research paper writing

# Chapter 4: PROGRESS TO DATE

## 4.1 Literature Review

In the progress to date, we have extensively reviewed a range of resources including research articles, books, and educational websites related to our research topic. A comprehensive review of over 16 articles has been conducted, and annotated bibliographies have been prepared for 10 selected articles. This literature review serves as a foundation for our ongoing research and will continue to be expanded and refined throughout the course of the study.

## 4.2 Database Collection

### 4.2.1 Phenotype Data selection

In the progress made to date, we focused on the selection of phenotype data for our research. Specifically, we extracted information related to the stage of cancer and the subtype of the cancer from the clinical data of the patients. This information will contribute to the development of our research objectives and modeling processes.

### 4.2.2 Data set selection

In the progress to date, our project focuses on the classification of cervical cancer subtypes using three omics approaches. Specifically, we have downloaded three datasets from The Cancer Genome Atlas (TCGA) to analyze and classify the two subtypes of cervical cancer. These datasets were obtained exclusively from TCGA, ensuring consistency and reliability in our analysis. Our current progress involves the preprocessing and exploration of these datasets to gain insights into the molecular characteristics and classification of cervical cancer subtypes.

## 4.3 Database Preparation

In the progress to date, our focus has shifted to the dataset analysis phase. Since our study involves multi-omic classification, we are working with multiple datasets. These datasets have been downloaded separately and stored in individual files. To facilitate our analysis, we have utilized Python code to merge each dataset with its corresponding phenotype data. This includes combining genomic data with cancer subgroups, cancer stages, and other relevant patient information. By leveraging patient ID as a common field, we have successfully integrated these datasets to create a comprehensive dataset for further preprocessing and analysis.

**References**

1. Y. Xu, H. Luo, Q. Hu, and H. Zhu, “Identification of potential driver genes based on multi-genomic data in cervical cancer,” Frontiers, 07-Jan-2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2021.598304/full. [Accessed: 29-Apr-2023].
2. W. Xu, M. Xu, L. Wang, W. Zhou, R. Xiang, Y. Shi, Y. Zhang, and Y. Piao, “Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers,” Nature News, 13-Dec-2019. [Online]. Available: https://www.nature.com/articles/s41392-019-0081-6. [Accessed: 29-Apr-2023].
3. A. A. Abdullah, “Development of predictive models for cervical cancer based on gene expression profiling data ,” Iop science. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/557/1/012003. [Accessed: 29-Apr-2023].
4. H. J. Cho, “Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment, Academic.oup.com ”. [Online]. Available: https://academic.oup.com/bioinformatics/article-abstract/39/4/btad113/7067742. [Accessed: 29- April-2023].
5. A. Shiraz, “The early detection of cervical cancer. The current and changing landscape of cervical disease detection, Wiley Online Library”. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cyt.12835. [Accessed: 29- April-2023].
6. D. Ding, “Machine learning-based prediction of survival prognosis in cervical cancer - BMC Bioinformatics, BioMed Central”. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04261-x. [Accessed: 29- April-2023].
7. B. Pardini, “MicroRNAs as markers of progression in cervical cancer: A systematic review - BMC cancer, BioMed Central”. [Online]. Available: https://bmccancer.biomedcentral.com/articles/10.1186/s12885018-4590-4. [Accessed: 29- April-2023].
8. D. Yan, “Integrated analysis of chromosome copy number variation and gene expression in cervical carcinoma, Oncotarget”. [Online]. Available: https://www.oncotarget.com/article/22403/text/ [Accessed: 29- April-2023].
9. J.H. Ma, “An 8-gene DNA methylation signature predicts the recurrence risk of cervical cancer, journals sage pub”. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/03000605211018443 [Accessed: 29- April-2023].
10. Al Mudawi, N. and Alazeb, A. (2022) A model for predicting cervical cancer using machine learning algorithms, MDPI. Available at: https://www.mdpi.com/1424-8220/22/11/4132 (Accessed: 28 May 2023).
11. Akter, L. et al. (2021) Prediction of cervical cancer from behavior risk using machine learning techniques - SN computer science, SpringerLink. Available at: https://link.springer.com/article/10.1007/s42979-021-00551-6 (Accessed: 15 May 2023).
12. Nithya, B. and Ilango, V. (2019) Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction - SN Applied Sciences, SpringerLink. Available at: https://link.springer.com/article/10.1007/s42452-019-0645-7 (Accessed: 30 May 2023).
13. Le, T.M. et al. (2022) Transcriptomic immune profiles can represent the tumor immune microenvironment related to the tumor budding histology in uterine cervical cancer, MDPI. Available at: https://www.mdpi.com/2073-4425/13/8/1405 (Accessed: 30 May 2023).
14. Hao, Y. et al. (2023) Construction of a prognostic model based on eight ubiquitination-related genes via machine learning and potential therapeutics analysis for cervical cancer, Frontiers in genetics. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10043205/ (Accessed: 30 May 2023).
15. Choi, C.H. et al. (2020) Chemoradiotherapy response prediction model by proteomic expressional profiling in patients with locally advanced cervical cancer, Gynecologic oncology. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7217751/ (Accessed: 30 May 2023).
16. El-Zein, M. and Cheishvili, D. (2020) Genome-wide DNA methylation profiling identifies two novel genes in cervical neoplasia- wiley online library, wiley online library. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.32880 (Accessed: 30 May 2023).