Department of Computer Science

MSc Data Science and Analytics

Academic Year 2023-2024

CLUSTERING OF PATIENTS PROFILE BASED ON MUTATED GENE, HISTOLOGY, ETHINICITY AND CANCER TYPE AND CLASSIFICATION OF CANCER TYPE BASED ON MUTATED GENE

Student Name: Nazeema Begum Rahman Basha

Student Id: 2351211

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

Brunel University
Department of Computer Science
Uxbridge, Middlesex UB8 3PH
United Kingdom
Tel: +44 (0) 1895 203397
Fax: +44 (0) 1895 251686

ABSTRACT

Endometrial cancer is one of the most common gynaecological cancers, with considerable prognosis and treatment response differences depending on genetic alterations, histological subtypes, ethnicity, and disease type. Advances in genomics and bioinformatics have allowed for the investigation of molecular changes that drive the genesis and course of this malignancy. This dissertation focusses on clustering patient profiles based on genetic mutations, histology, ethnicity, and cancer type, as well as identifying endometrial cancer types using mutation profiles. By doing so, the study aims to bring knowledge about the genetic landscape of endometrial cancer, as well as its relationship to disease prognosis and personalised therapy choices.

The primary goal of the study is to categorise endometrial cancer patients based on critical clinical and genetic variables and categorise disease subtypes using machine learning algorithms. To achieve this, the study begins with a thorough literature analysis that investigates the present understanding of molecular and genetic profiles in endometrial cancer, revealing research gaps that this dissertation intends to fill. The role of ethnicity on cancer progression is highlighted, as socio-demographic factors may lead to differences in clinical outcomes.

The following step is data preprocessing, which ensures that the dataset is clean and ready for analysis. Missing values, duplicate entries, and anomalies in mutation data are carefully managed to ensure the data's integrity. Following that, descriptive statistics and exploratory data analysis (EDA) are used to better understand the dataset's structure, which includes visualisations of key categorical variables such as cancer types, ethnicity, and histology. This exploratory process provides useful insights into potential patterns and trends in the data.

The heart of the research is K-means clustering to classify patients based on genetic mutations, histology, ethnicity, and cancer type, with the goal of identifying subgroups with common genetic and clinical characteristics. The study aims to discover potential correlations between genetic variations and clinical factors including histology and ethnicity, which could provide new insights into disease progression.

Furthermore, Classification models are created using machine learning techniques, specifically Random Forest, to forecast cancer types based on mutation profiles. This stage is essential for understanding how genetic alterations influence cancer classification and diagnosis. The model's accuracy in identifying cancer subtypes contributes to more accurate treatment strategies for endometrial cancer patients.

Furthermore, the dissertation reveals the most important genes involved in cancer classification. The study identifies the top 10 most important genes for each cancer type using machine learning techniques, providing essential insights into the disease's molecular drivers. Visualisations are generated to show how these genes contribute to classification models, highlighting the significance of genetic alterations in understanding the cancer's molecular landscape.

Finally, the study confirms the clustering and classification results by comparing them to clinical data to assure their practical application. By addressing these objectives, this study contributes to a more refined understanding of endometrial cancer, laying the groundwork for developing personalised therapy choices based on patients' genetic and clinical characteristics.

ACKNOWLEDGEMENTS

I would like to convey my sincere appreciation and gratitude to all the people and stakeholders who aided me in the writing of this dissertation. First, my sincere appreciation and gratitude go to my academic supervisor **Dr Annette Payne** for providing me with the most pertinent suggestions, feedback and encouragement throughout the process of this research.

I also would like to thank my professors and other faculty members of Brunel University London for the knowledge and other contributions needed to conduct this research. I would like to express my gratitude to my friends, coworkers and classmates for support as well as discussions that enlightened me and enhanced my view.

I would like to express my profound gratitude to my family and friends for their support, tolerance and comprehension during the years of doing this work. I wish to take this opportunity to thank you for believing in me and standing behind me.

Lastly, the participants and the respondents, who made it possible to gather proper data and gain some valuable information for this study. Without the cooperation of participants, this work would not have been possible.

Thank you all for your unwavering support and belief in my work.

Declaration

I, Nazeema Begum Rahman Basha, hereby declare that this dissertation, titled "Clustering of Patients Profile Based on Mutated Gene, Histology, Ethnicity and Cancer Type and Classification of Cancer Type Based on Mutated Gene," is my original work submitted in partial fulfilment of the requirements for the award of Master's in Data science and Analytics at Brunel University London. I confirm that all the research, analysis, and conclusions in this dissertation are my own, and any external sources used have been properly credited. This dissertation has not been submitted, in whole or in part, for any other degree or professional qualification.

I certify that the work presented in the dissertation is my own unless referenced

Signature: Nazeema Begum Rahman Basha

Date: 11-09-2024

TOTAL NUMBER OF WORDS: 13569

Table of Contents

Table of figures

CLUSTERING OF PATIENTS PROFILE BASED ON MUTATED GENE, HISTOLOGY, ETHINICITY AND CANCER TYPE AND CLASSIFICATION OF CANCER TYPE BASED ON MUTATED GENE

CHAPTER 1 INTRODUCTION

Endometrial cancer is the most frequent tumour affecting the female reproductive tract, accounting for over 150,000 new cases worldwide each year (Okuda *et al.*, 2010). This is a malignancy arising from the lining of the uterus, is one of the most prevalent gynaecological cancers globally, with incidence rates continuing to rise. It accounts for a significant portion of cancer-related morbidity and mortality among women (Murali *et al.*, 2018). Endometrial cancer affects approximately 3% of women over their lives (Monteiro, 2009) (Bianco *et al.*, 2020). The condition is highly diverse, with a range of genetic, histological, and demographic differences, complicating diagnosis and treatment. As a result, molecular profiling of tumours to better understand the disease's biology is a developing area of interest in oncological research (Okuda *et al.*, 2010).

The molecular landscape of endometrial cancer is highly heterogeneous. Endometrial cancer falls into two types: Type I (endometrioid carcinoma) and Type II (non-endometrioid carcinoma), each having unique clinical, histological, and molecular characteristics. Type I tumours are typically estrogen-dependent and carry abnormalities in genes such as PTEN, KRAS, and PIK3CA. Type II tumours often lack oestrogen and have mutations in TP53, Her2/neu, and FBXW7(Murali *et al.*, 2018). Understanding the genetic and molecular distinctions between these two categories is crucial for improving patient outcomes and that provide more effective treatment approaches.

The PTEN gene, which regulates cell division and prevents uncontrolled growth, gets modified in about half of all endometrial malignancies. PTEN mutations are common in Type I malignancies and frequently co-occur with KRAS mutations, contributing to carcinogenesis (Okuda *et al.*, 2010). Type II tumours, on the other hand, are usually associated with mutations in TP53, a gene involved in DNA damage response and apoptosis (Bianco *et al.*, 2020). Such genetic changes drive the biological disparities across endometrial cancer subtypes, providing insight into why these tumours behave so differently.

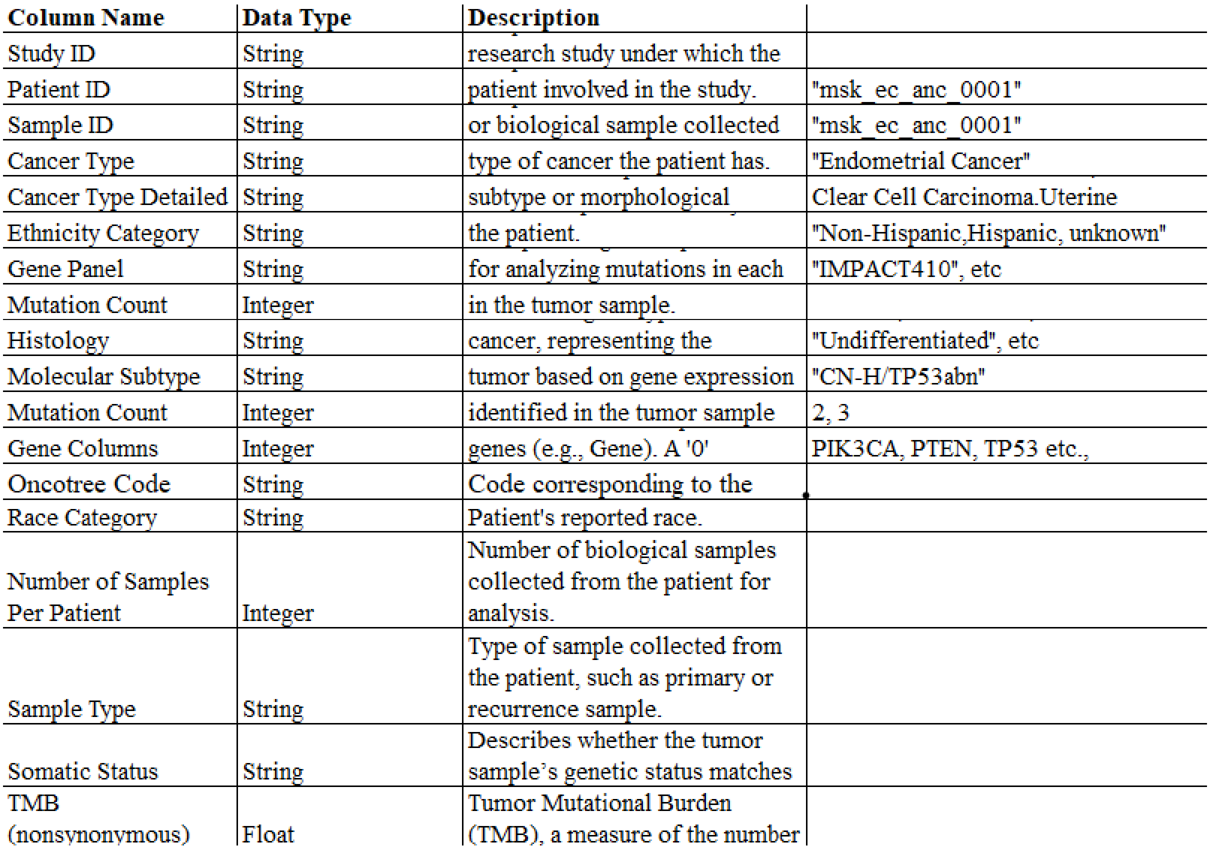
1.1 Importance of Clustering and Classification

In addition to genetic abnormalities, ethnicity has been proven to have an important influence in endometrial cancer. According to research, certain ethnic groups, such as African American women, may be more prone to aggressive forms of the disease, possibly due to underlying genetic differences (Bianco *et al.*, 2020). For example, studies have found that TP53 mutations, which are linked with a poor prognosis, are more common in endometrial malignancies in African American women than in Caucasian women (Bianco *et al.*, 2020). This shows that clustering patients based on ethnicity, in addition to genetic and clinical characteristics, could provide useful insights into health origin and help drive personalised therapy approaches (Murali *et al.*, 2018) (Bianco *et al.*, 2020)

Clustering patients based on their genetic and clinical characteristics enables the discovery of new subgroups within the endometrial cancer population, which could have significant consequences for therapy and prognosis. Researchers can better understand the links between genetic alterations, cancer aggressiveness, and patient outcomes by classifying patients based on molecular features (Murali *et al.*, 2018). Clustering, for example, may reveal that genetic alterations are more common in aggressive malignancies, allowing tailored medicines to be developed that specifically address these mutations (Bianco *et al.*, 2020). Classification algorithms, such as machine learning models, can help us better understand endometrial cancer by predicting cancer types based on genetic alterations.

Data description:

This dataset contains genetic and clinical information from patients with endometrial cancer, including mutation profiles, tumour histology, molecular subtype, and patient demographics such as ethnicity and race. This dataset is sourced from the **cBioPortal** [1]platform (cBioPortal for Cancer Genomics, 2024), which provides access to large-scale cancer genomics data. It includes patient-level data for endometrial cancer, covering genetic mutations, tumour characteristics, and clinical information, facilitating research into cancer genomics and personalized medicine.



Column Name	Data Type	Description	
Study ID	String	research study under which the	
Patient ID	String	patient involved in the study.	"msk_ec_anc_0001"
Sample ID	String	or biological sample collected	"msk_ec_anc_0001"
Cancer Type	String	type of cancer the patient has.	"Endometrial Cancer"
Cancer Type Detailed	String	subtype or morphological	Clear Cell Carcinoma.Uterine
Ethnicity Category	String	the patient.	"Non-Hispanic,Hispanic, unknown"
Gene Panel	String	for analyzing mutations in each	"IMPACT410", etc
Mutation Count	Integer	in the tumor sample.	
Histology	String	cancer, representing the	"Undifferentiated", etc
Molecular Subtype	String	tumor based on gene expression	"CN-H/TP53abn"
Mutation Count	Integer	identified in the tumor sample	2, 3
Gene Columns	Integer	genes (e.g., Gene). A '0'	PIK3CA, PTEN, TP53 etc.,
Oncotree Code	String	Code corresponding to the	
Race Category	String	Patient's reported race.	
Number of Samples Per Patient	Integer	Number of biological samples collected from the patient for analysis.	
Sample Type	String	Type of sample collected from the patient, such as primary or recurrence sample.	
Somatic Status	String	Describes whether the tumor sample's genetic status matches	
TMB (nonsynonymous)	Float	Tumor Mutational Burden (TMB), a measure of the number	

Figure 1 - Data Description

1.2 Research Aim

The primary aim of this study is to explore the clustering of endometrial cancer patients based on genetic mutations, histology, ethnicity, and cancer type, and to classify cancer types according to their mutation profiles. By achieving this, the research seeks to provide insights into the molecular characteristics associated with disease prognosis and treatment outcomes.

[1] Endometrial Carcinoma: https://www.cbioportal.org/study/summary?id=ucec_ancestry_cds_msk_2023

1.3 Objectives:

1.3.1 Conduct a complete literature review:

Examine current endometrial cancer research, focussing on molecular and genetic profiles, histological types, and the impact of ethnicity on disease development. This literature review will help clarify the current study and highlight gaps in existing data.

1.3.2 Data Preprocessing and Cleaning:

Ensure that the dataset is clean and ready for analysis by dealing with missing values, checking for duplicate entries, and resolving discrepancies in mutation counts. This step is critical to ensuring the data's integrity and usefulness for further study.

1.3.3 Descriptive statistics and exploratory data analysis (EDA)

Performing an EDA to better understand the dataset's structure, including visualising the distribution of key categorical variables like cancer type, ethnicity, and histology. This aids in the identification of trends and patterns that may be useful for future investigation.

1.3.4 Relationship Analysis for Categorical Variables:

Analysing the correlations between categorical variables such as cancer type, ethnicity, and histology to determine how they may be associated. This includes visualising tumour types across ethnicities and histology.

1.3.5 Clustering Analysis of Patient Profiles:

Using K-means clustering to analyse patient profiles by genetic mutations, histology, ethnicity, and cancer type. The goal is to combine individuals with similar profiles and find possible subgroups who share genetic, clinical, and demographic factors

1.3.6 Classification of Cancer Types Based on Gene Mutations:

Using machine learning (Random Forest) to predict cancer types based on gene alterations. This approach will assist in classifying endometrial cancer types based on mutation profiles, allowing for more precise cancer diagnosis and treatment planning.

1.3.7 Identifying the Most Significant Genes for Classification:

Using machine learning techniques such as Random Forest to determine the top ten most essential genes for each cancer type. This stage is vital for determining which genetic mutations play the most important role in classifying across cancer types.

1.3.8 Visualising Gene Importance in Different Cancer Types:

Creating visualisations of the most essential genes for different cancer types, demonstrating how they contribute to classification models. This highlights the importance of individual gene alterations in comprehending endometrial cancer's molecular landscape.

1.3.9 Validating Clustering and Classification Results:

Evaluate the findings by comparing the clustering and classification results to clinical data. This ensures that the identified subgroups and gene mutation patterns have practical applications in predicting patient outcomes and guiding personalised treatment methods.

By addressing these aims, this study will give a thorough examination of endometrial cancer patients, enabling for more focused and personalised therapy options based on genetic and clinical characteristics.

1.4 Research Questions:

1. Can clustering techniques identify groups of patients with similar combinations of gene mutations, while also accounting for their ethnicity, histology, and cancer type in endometrial cancer?
2. classify the patient's cancer type based on the gene mutation or mutations they have

CHAPTER 2 LITERATURE REVIEW

Endometrial carcinoma is among the most common gynaecological cancers globally, particularly in developed countries. In 1983, Bokhman introduced the dualistic paradigm for endometrial carcinoma, categorising it into two types: oestrogen-dependent endometrioid carcinomas (Type 1) and oestrogen-independent non-endometrioid carcinomas (Type 2)). This model provided a framework for understanding the distinctions in behaviour and therapy between these two categories. Endometrioid carcinomas have a better prognosis than non-endometrioid carcinomas, which are more aggressive and lead to lower results (McConechy *et al.*, 2012).

Sigurd F. Lax conducted a detailed assessment in 2003 on the molecular genetic pathways of several forms of endometrial cancer, distinguishing between Type I and Type II tumours. This review expanded on the dualistic paradigm of endometrial carcinogenesis. Type I tumours, such as endometrioid and mucinous carcinomas, are estrogenic-dependent and usually associated with mutations in genes such as PTEN, KRAS, and beta-catenin. These mutations occur early during these tumours, specifically between atypical endometrial hyperplasia and carcinoma. Type II tumours, which include serous and clear cell carcinomas, follow an estrogen-independent pathway and are distinguished by p53 mutations, which frequently result in poor outcomes (Lax, 2004)

However, by 2004, molecular investigations had confirmed this dualistic paradigm, demonstrating that PTEN mutations are present in up to 83% of endometrioid carcinomas, making them one of the most significant genetic changes for this subtype. On the other hand, p53 mutations were found in nearly 90% of serous carcinomas, which are the precursors to Type II tumours. These studies revealed the several molecular pathways involved in the development of Type I and Type II endometrial carcinomas (Lax, 2004). These findings highlighted the value of molecular profiling in accurately diagnosing endometrial carcinomas and directing treatment choices. The review also highlighted the importance of immunohistochemistry in detecting p53 mutations, which are critical in the aggressive clinical behaviour of Type II carcinomas.

By 2010, the molecular profile of endometrial cancers was studied, emphasising the dualistic paradigm of endometrial carcinoma as Type I (estrogen-dependent) or Type II (non-estrogen dependent). This divergence has important implications for understanding the pathophysiology and clinical behaviour of endometrial cancer. Type I tumours are often associated with favourable outcomes and mutations in genes such as PTEN, KRAS, and PIK3CA, whereas Type II tumours are more aggressive and frequently contain TP53 alterations (Coenegrachts *et al.*, 2015). This paradigm paved the way for targeted therapeutics centred on specific molecular pathways.

Furthermore, around 2012, great progress had been made in refining the categorisation of endometrial carcinomas through mutation profiles. McConechy et al. used targeted exon sequencing on 393 endometrial cancer samples, identifying nine important genes: ARID1A, PTEN, PIK3CA, KRAS, CTNNB1, TP53, PPP2R1A, BRAF, and PPP2R5C. A study published in The Journal of Pathology found significant molecular differences between high-grade endometrioid (EEC-3) and serous carcinomas (ESC). This study found that using molecular classifiers can improve diagnostic reliability and stratify patients for targeted therapy, highlighting the limitations of histology alone
Additionally, in 2012, the same study found two unique mutation patterns for carcinosarcomas, an aggressive subtype of endometrial cancer. These tumours had either endometrioid-type mutation (PTEN, PIK3CA, ARID1A, KRAS) or serous-type mutations

(TP53, PPP2R1A), highlighting the variability within endometrial carcinoma subtypes (McConechy *et al.*, 2012).

The Cancer Genome Atlas (TCGA) provided a molecular categorisation scheme for endometrial cancer, stratifying the disease into four separate groups: POLE ultramutated, microsatellite instability-high (MSI-H), copy-number low, and copy-number high . This classification method identifies considerable disparities in prognosis and molecular characteristics among these categories, offering a framework for understanding the heterogeneity of endometrial cancer (Intl Journal of Cancer). The identification of these subgroups, based on thorough genomic profiling, contributed to a better understanding of altered genes, particularly TP53, POLE, and MSH mutations, which are critical in determining patient outcomes.

Laas et al. used supervised clustering of immunohistochemical markers to distinguish between atypical endometrial hyperplasia (AEH) and grade 1 endometrial carcinoma. This study used markers such as MMP-9, oestrogen receptor (ER), progesterone receptor (PR), and CD44 isoforms, which were shown to be over- or under expressed in certain clusters. The study found that this clustering algorithm could identify between AEH and grade 1 EC with a misclassification rate of just 8%. This study refined our understanding of molecular changes that occur as endometrial hyperplasia progresses to cancer (Laas et al ,2014).

In 2015, Coenegrachts et al. investigated the mutation patterns of mixed endometrioid-serous endometrial carcinomas (mixed EEC-SC) and discovered that the clinical outcomes of mixed tumours were intermediate between pure endometrioid and serous carcinomas. Notably, TP53 mutations were more common in mixed EEC-SC and serous carcinomas, whereas PTEN mutations were mostly absent in serous carcinomas but found in a subset of endometrioid tumours (Coenegrachts et al., 2015). This work demonstrated the molecular heterogeneity between mixed and pure types of endometrial cancer, emphasising the importance of precise molecular diagnoses.

However, Laas et al. used unsupervised clustering of immunohistochemical markers to identify high-risk endometrial cancer patients. Their investigation discovered unique protein expression profiles linked to Type I and Type II tumours, revealing that Type I Grade 3 and Type II endometrial malignancies had significant molecular similarities, particularly in p53 expression (Laas *et al.*, 2017). This study helped to improve the classification of high-risk endometrial malignancies and improved treatment methods using immunohistochemistry profiling.

Furthermore, a study on racial disparities in young women with endometrial cancer found differences in disease presentation and survival rates by ethnicity. Despite controlling for stage and histology, black women had a 19% higher mortality rate than white women due to more aggressive and advanced-stage tumours (Mukerji et al., 2018). This finding highlighted the importance of ethnicity in the classification of endometrial cancer, as inequalities in outcomes are frequently impacted by genetic alterations and cancer type among racial groups.

further research into racial discrepancies in endometrial cancer outcomes had acquired importance. Huang et al. investigated how adherence to evidence-based care influences survival inequalities among Black and White women with endometrial cancer. Black women had a 30% lower incidence of endometrial cancer than White women, but their mortality rate was 80% greater. Despite perfect adherence to quality criteria, Black women nevertheless had higher rates of 30-day, 90-day, and 5-year mortality compared to White women (Huang et al., 2020). This study emphasised the significance of ethnicity and access to care when analysing survival rates and treatment choices for endometrial cancer.

Li et al. investigated a next-generation sequencing (NGS) strategy for simplifying molecular classification. Patients were classified into four primary categories based on their mutational profiles: POLEMUT, MSI-H, TP53WT, and TP53MUT. The TP53MUT group, primarily composed of patients with serous carcinoma, had the lowest disease-free survival (DFS), while the POLEMUT group had the best results (Li et al., 2022) This molecular classification improved prognosis prediction accuracy while also providing vital insights into the relationship between altered genes and cancer types.

In 2023, Alessandrino et al. investigated the mutational landscape of uterine serous carcinoma (USC) and its relationship to metastasis and recurrence. This aggressive form of endometrial cancer is more common in Black, Asian, and Hispanic women, with the most often altered genes being TP53, ERBB2, and PIK3CA. It was discovered that ARID1A mutations were related with significantly lower overall survival, particularly when metastases originated in the liver (Alessandrino et al., 2023). This study emphasised the importance of ethnicity and mutational characteristics in understanding treatment outcomes and recurrence patterns in USC patients.

In addition to that, genetic profiling has revealed ethnic differences in the incidence of certain mutations and their effects on survival. For example, ERBB2 mutations were more common in non-Hispanic women, whereas PIK3CA mutations were associated with poorer outcomes, particularly in those with liver metastases (Alessandrino et al., 2023). This study emphasises the importance of personalised therapy approaches that take into consideration both genetic and ethnic characteristics.

By 2023, the combination of clinical sequencing and immunohistochemistry for molecular classification of endometrial carcinoma would have refined cancer classification systems. This study classified tumours using a combination of genetic and immunohistochemical markers, resulting in more accurate identification of cancer subtypes (Rios-Doria et al., 2023). The integrated classification system has high concordance with current surrogate indicators and could classify more cases of endometrial cancer than older methods. Mutations in genes including TP53, PIK3CA, and MSH can differentiate aggressive cancer subtypes and provide insight into prognosis across stages and histologic types.

In 2023, researchers used clinical sequencing and immunohistochemistry to improve the genetic categorisation of endometrial cancer. This strategy increased the number of cases identified by combining genetic data such as TP53, PIK3CA, and MSH mutations with standard histology markers. This comprehensive strategy improved prognosis for endometrial cancer across all stages and histological subtypes, leading to better tumour classification and treatment choices (Rios-Doria et al., 2023).

2.1 Research Gap

While the existing literature provides substantial insights into the molecular pathways, mutation profiles, and clinical outcomes of various endometrial cancer types, there are significant gaps when compared to the scope and objectives of this research topic, which focusses on clustering patient profiles based on mutated genes, histology, ethnicity, and cancer type, and the classification of uterine cancer types such as Uterine Undifferentiated Carcinoma, Endometrial Cancer, The following research gaps demonstrate how this research varies from previous studies:

2.1.1. Existing research focusses on certain cancer types:

Existing Research: The research summarised focusses on the dualistic paradigm of Type I (endometrioid) and Type II (serous) carcinomas (Bokhman, 1983; McConechy et al., 2012;

Lax, 2004). This dualistic classification focusses mostly on endometrioid carcinoma and serous carcinoma, with minimal attention paid to other uterine cancer subtypes such as uterine carcinosarcoma, uterine mixed endometrial carcinoma, undifferentiated carcinoma, and clear cell carcinoma.

Research Gap: This study will investigate a broader range of uterine cancer subtypes, including uterine undifferentiated carcinoma and uterine mixed endometrial carcinoma, which have not been fully explored in relation to influenced gene types as these less widely recognised cancer subtypes have distinct molecular and genomic features that warrant additional exploration.

2.1.2 Clustering by Multiple Factors (Gene, Histology, Ethnicity):

Existing Research: Laas et al. (2014) and Coenegrachts et al. (2015), for example, primarily focus on mutational profiling or histological classifications for endometrial cancers, frequently investigating individual factors such as gene mutations (PTEN, TP53) or ethnic disparities (Mukerji et al. 2018; Huang et al., 2020). However, these studies lack a complete clustering approach that considers various criteria, including altered genes, histology, and ethnicity.

Research Gap: The goal of this study is to cluster patient profiles by combining mutant genes, histology, and ethnicity, resulting in a more comprehensive clustering and categorisation. While racial differences have been identified (Mukerji et al., 2018; Huang et al., 2020), few research have systematically examined these issues. Clustering patient data based on a multifactorial approach (genes, ethnicity, histology) can offer a more detailed picture of patient outcomes and cancer progression

2.1.3 Prioritise Mutated Gene Types for Cancer Classification:

Existing research: Molecular classifications proposed by McConechy et al. (2012) and Coenegrachts et al. (2015) place a strong emphasis on detecting mutations such as TP53, PTEN, and PIK3CA in Type I and Type II tumours. However, these studies are limited to distinguishing between generic classifications such as endometrioid and serous carcinoma, rather than expanding this classification to other cancer types.

Research Gap: This study focusses on the classification of certain uterine cancer subtypes, such as uterine serous carcinoma, uterine clear cell carcinoma, and uterine mixed endometrial carcinoma, using mutant gene types. Although the molecular profiles of Type I and Type II endometrial cancers are well established, research into using mutational profiles to classify a broader range of uterine cancers is restricted. By filling this gap, this research will provide more detailed classifications, improve diagnosis accuracy, and influence personalised treatment options.

2.1.4 Genetic and ethnic disparities in less well-known cancer subtypes:

Existing Research: Racial disparities have mostly been investigated in common cancer forms such as endometrioid carcinoma and serous carcinoma, with minimal information on rarer subtypes such as uterine carcinoma sarcoma and uterine papillary serous carcinoma (Mukerji et al., 2018; Huang et al., 2020).

Research Gap: This study will look at ethnic differences in a broader spectrum of uterine cancer forms, with a focus on how mutations interact with ethnic backgrounds in rare and severe cancer subtypes such as uterine carcinosarcoma and uterine clear cell carcinoma. Ethnic differences in mutational patterns and cancer subtypes remain unexplored. This study will provide crucial insights into how ethnicity and gene variants influence cancer outcomes, allowing for better treatment protocols and addressing healthcare disparities.

This study aims to provide a multifactorial clustering approach for a wider range of uterine cancers, integrating mutated gene types, histological classification, and ethnicity. This approach will offer a more personalized and inclusive framework, better accounting for genetic and racial disparities across different cancer subtypes.

CHAPTER 3 METHODS

3.1 Data Preprocessing and Cleaning:

This section discusses the methods for data cleaning and preprocessing, with a focus on handling missing values, finding discrepancies, and removing duplicates and outliers. The dataset that I used is an endometrial cancer dataset from the CBIOS portal (cBioPortal for Cancer Genomics, 2024). These procedures prepare the data for subsequent analysis, including clustering.

3.1.1 Handling Missing Values:

Missing values are common in datasets and can arise due to insufficient data collection, human error, or inconsistent input (Kwak and Kim, 2017). In this investigation, missing values were treated using the following approach:

Missing Values Identification: The dataset was initially explored using the str() and summary() methods, which showed features with missing values. Missing numbers were addressed by using the na.rm = TRUE argument, which allows for the total of mutant gene counts while ignoring missing values.

Imputation: The technique of using na.rm = TRUE is an imputation method that excludes missing values from calculations rather than discarding entire rows. This approach ensures a complete dataset without bias from missing data (Kwak & Kim, 2017) . By imputing missing values, the analysis preserves more data, limiting the possibility of distorted conclusions caused by missing entries.

3.1.2. Identifying and Resolving Data Discrepancies:

Identifying column discrepancies is a vital step in guaranteeing data integrity. In this investigation, mutation counts for each patient were obtained by adding the relevant mutation columns and then compared to the original Mutation.Count column. This phase ensured the consistency of the dataset.

Comparison of generated and Original Mutation Count: To identify differences, the code generated mutation sums across key columns and compared them to the Mutation.Count column. Cases in which the estimated values did not match the original were flagged for more examination. Identifying and resolving inconsistencies is crucial for proper data representation and analysis (Jaeger & Banks, 2022)

3.1.3 Handling Duplicate Entries

Duplicate entries are a common problem, particularly when datasets are compiled from multiple sources or when errors occur during data collection. The methodology for detecting and handling duplicates involved:

- **Duplicate Detection**: The duplicated () function was employed to check for duplicate column names, which ensures that no redundant information is present in the dataset. Removing duplicate columns or entries helps maintain data integrity, ensuring that each data point is represented only once. Literature supports the use of field-level and record-level matching algorithms to detect such duplicates, particularly in datasets merged from multiple sources (Do, Graefe and Naughton, 2022).

- **Elimination of Duplicate Columns**: In the current dataset, the detection and removal of duplicate columns ensured that the clustering algorithms could function correctly without encountering redundant or conflicting data

3.1.4. Outlier Detection and Removal:

Outliers can skew statistical analysis and lower the accuracy of machine learning models like clustering algorithms. The process for dealing with outliers involves:

Outlier Detection and Removal: The code removed row 1832 from the dataset (endometrial_cleaned <- endometrial_dataset[-1832,]) because it was flagged as problematic or containing incorrect data. Outliers can influence clustering algorithms, resulting in erroneous results (Kwak & Kim, 2017). By deleting these rows, the dataset is more suited to precise and reliable clustering analysis.
This method of data cleaning assures that the dataset is accurate, full, and error-free, giving a solid platform for subsequent research, including the use of clustering techniques.

3.2 Exploratory data analysis (EDA).

This project's exploratory data analysis (EDA) uses descriptive statistics as well as visualisation tools to summarise and visualise the major characteristics of the endometrial cancer dataset. The procedure follows typical methodologies for data exploration, as noted by Larson (2006) focussing on categorical variables, numerical variables, and relationships among variables (Larson, 2006)

3.2.1. Descriptive Statistics of Categorical Variables

The initial stage in EDA is to calculate descriptive statistics for categorical variables such cancer type, ethnicity category, and histology. Descriptive statistics for categorical variables include frequency counts, which provide an overview of how frequently each category appears in the dataset. Larson (2006) identifies frequency statistics as the principal descriptive technique for categorical variables, offering excellent insights into data distribution within each category (Larson, 2006)

In this scenario, the table () method was used to demonstrate the distribution of cancer types, ethnicities, and histological groups. These statistics aid in analysing the dataset's overall structure and detecting any prominent categories or imbalances.

3.2.2. Visualising Categorical Variables

To show the distribution of these categorical variables, bar plots were created with ggplot2, a powerful statistical graphics configure (Wickham, 2011). Bar plots help visualise patterns across cancer Types, ethnicities, and histological categories

Cancer Type Distribution: A bar plot depicted the frequency of each specific cancer types. This enabled the identification of common cancer types in the dataset.

Ethnicity Distribution: Another bar plot showed the distribution of ethnicity groupings. This is critical for interpreting demographic representations in the dataset.

Histology Distribution: A third plot depicted the distribution of histological forms of cancer, which is critical for determining the variation in cancer subtypes.

3.2.3 Investigating Relationships Between Categorical Variables

Relationships between categorical variables, such as **Cancer Type Detailed** and **Ethnicity Category**, were also explored. This was achieved by grouping the data based on these variables and calculating the percentage of each cancer type within each ethnic group. Visualization through grouped bar plots further emphasized the relationships between cancer types and ethnic categories, which aligns with Chang and Ding's (2004) work on interactive visualization to analyse categorical data relationships.

3.2.4 Summary Statistics of Numerical Variables

Summary statistics for the numerical variable Mutation Count were calculated, including mean, median, and standard deviation. These statistics provide insight into the central tendency and distribution of mutation counts across the dataset, which is critical for understanding genomic changes associated with cancer types. Wickham (2011) emphasises the importance of numerical summaries in the EDA process

3.2.5 Visualisation of numerical variables

A histogram was built to visualise the distribution of mutation counts, demonstrating how frequently different types of mutations occur. Histograms are important for assessing distribution shape, identifying outliers, and establishing whether the data is skewed or regularly distributed, as highlighted by the chosen bin width guarantees that a balance between detail and clarity in the visualization.

3.2.6 Analysis of Relationships Between Cancer Type and Ethnicity

Grouping and percentage calculation technique used to understand tumour distribution across different ethnicities. Visualisation of categorical data helps draw conclusions about cancer prevalence in various demographic groups. Bar plot visualization illustrates relationships in a clear and interpretable manner. Use of percentage labelling within grouped bar plots helps understand variations across ethnic groups (Chang and Ding, 2005).

3.2.7 Analysing Mutation Counts

Histograms used to analyse the distribution of mutation counts. Wickham (2011) emphasizes the importance of appropriate bin widths for histograms to avoid oversimplification. Analysis helps understand genetic variations linked to cancer subtypes and detect outliers or anomalies. Combination of bar plots and histograms provides a comprehensive exploratory analysis.

3.2.8 Correlations and Pairwise Relationships

This section uses scatter plots and correlation matrices to analyse pairwise correlations between quantitative variables, with a special emphasis on mutation counts and their relationships to cancer types. Franzese and Iuliano (2019) emphasise the importance of correlation analysis in determining correlations between continuous variables like mutation counts and cancer outcomes
The boxplot depicts the distribution of mutation counts among cancer types, offering insight into any noteworthy differences across categories. Meanwhile, the correlation matrix quantifies the linear correlations between numeric variables, which is critical for understanding how mutation counts relate to other parameters in the dataset. Using this methodology, researchers can detect patterns and outliers that may impact clustering and classification procedures methodology, researchers can detect patterns and outliers that may impact clustering and classification procedures (Franzese and Iuliano, 2019).

3.3 Cluster analysis of patient's profiles

The cluster analysis method utilised here is based on patient profiles that include histology, ethnicity, cancer type, and mutation count, and it adheres to the k-means clustering principles defined by Jaeger and Banks (2022). The elbow approach is used to discover the appropriate number of clusters, ensuring that clusters are meaningful by minimising within-cluster variance while increasing between-cluster separation. Following the suggested approach, categorical variables for each patient were one-hot encoded, while numerical features were standardised to address any feature scale imbalances. The patients were divided into clusters using the k-means method, which revealed substantial connections between genetic alterations and clinical characteristics such as histology and ethnicity. (Jaeger and Banks, 2022)

3.4 Classification of Cancer Types Based on Gene Mutations:

The study of cancer type classification based on gene mutations is founded on the recognition that molecular genetic pathways play an important role in the development and classification of various endometrial carcinomas. According to Sigurd F. Lax (2004), the differentiation between type-I and type-II carcinomas is mostly determined by the genetic changes detected in each subtype. Mutations in the PTEN, K-ras, genes, as well as microsatellite instability (MIN), are common in type-I carcinomas, including endometrioid carcinoma. These mutations contribute to the development of molecular classification models, which are then used to forecast cancer types based on gene mutations.

In the Random Forest classification model, the cancer type is predicted as a categorical variable using a series of gene mutation profiles. The computer learns patterns in the mutation data by training the model on a dataset with cancer types labelled according to known mutations, such as the presence or absence of mutations in important genes such as PTEN and K-ras. According to Lax (2004), these mutations have a crucial role in predicting endometrial cancer progression. This is consistent with the application of Random Forest algorithms, which can handle complex interactions between factors (e.g., the existence of many mutations) and provide insights into which gene mutations are most predictive of specific cancer types.

3.4.1 Identifying the Most Significant Genes for Classification:

Baker and Kramer's method for selecting genes that significantly contribute to effective classification in microarray data informs the analysis of the visualisation for the top ten genes associated in specific cancer types. A major strategy in the implementation is to loop through the many cancer types and extract the top genes that contribute to each classification. This method for finding key genes is consistent with the approach utilised in their study, in which classification rules are constructed with a focus on a small number of genes that provide strong predictive signals, decreasing noise from less relevant genes.

As Baker and Kramer discovered, focussing on a few genes rather than a huge number is crucial for assuring robust classification. This technique selects genes with a high frequency of occurrence across categorisation models, making them more likely to contribute to understanding cancer biology (Baker and Kramer, 2006). By using this technique, not only visualise the function of certain genes across cancer types, but also highlight those that are repeatedly implicated in the categorisation process, which can help refine future research.

3.4.2 Visualising Gene Importance in Different Cancer Types:

The visualisation techniques used in the implementation is to plot the top genes by cancer types, such as "Uterine Serous Carcinoma" and others, are in line with current approaches to healthcare data visualisation. According to the review by Abudiyab et al. (2022), effective visualisation approaches are critical for understanding complicated healthcare data, particularly in genomic investigations where mutation patterns are convoluted and involve numerous variables. Visualisation simplifies data interpretation and supports clinical decision-making processes (Abudiyab and Alanazi, 2022.

The use of distinct colours for each cancer types, such as pink for "uterine serous carcinoma" and yellow for "uterine mixed endometrial carcinoma," improves the clarity of the results. This approach is consistent with the narrative review's conclusion that visualisation should enable stakeholders, including healthcare practitioners, to easily evaluate data and make evidence-based decisions.

CHAPTER 4 RESULTS

4.1 Visualize the distribution of key categorical variables using bar plots.

4.1.1 Distribution of Cancer Types vs Number of Patients:

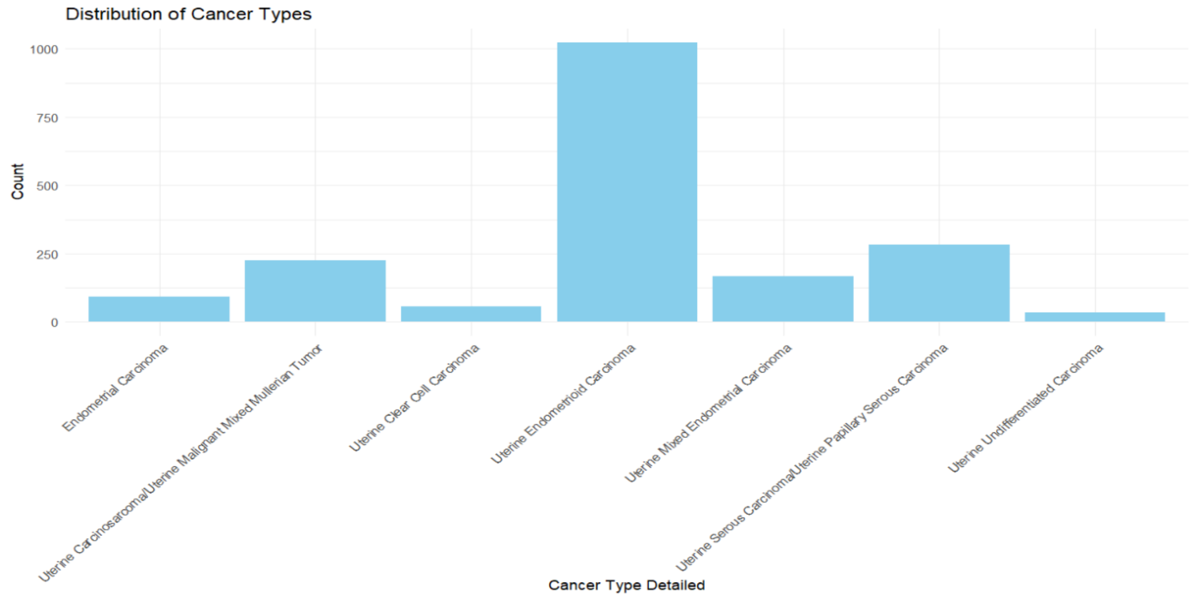


Figure 2 - Distribution of cancer type detailed Vs Number of patients

Fig 2 shows that "Uterine Endometrioid Carcinoma" is the most common cancer type in the sample, with a much greater prevalence than "Uterine Mixed Endometrial Carcinoma" and "Uterine Serous Carcinoma." This concentration represents a possible focus for cancer-specific research or treatment options. Research suggests that understanding the distribution of cancer types helps improve clustering models by identifying common genetic alterations (Unwin, 2018).

4.1.2 Distribution of Ethnicity vs Number of Patients:

Fig 3 shows that the majority of patients are "Non-Hispanic," with smaller proportions in the "Hispanic" and "Unknown" categories. This skewed distribution suggests that ethnicity may have an important influence in mutation patterns and cancer outcomes. (Salanti, Ades and Ioannidis, 2011)

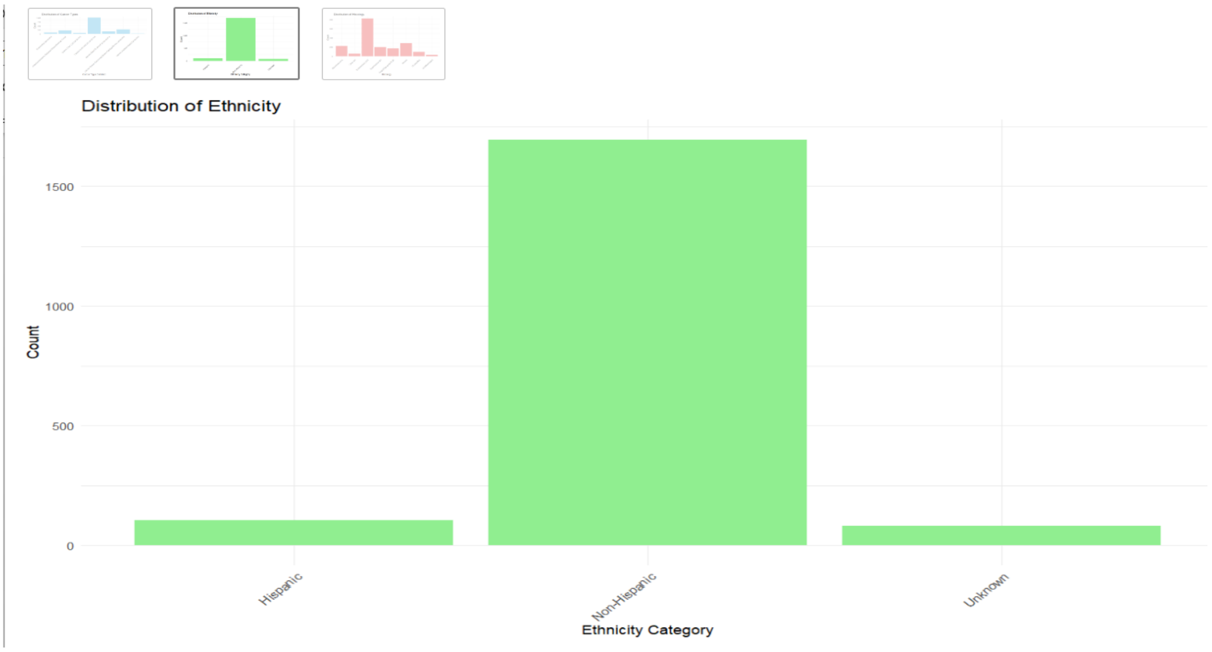


Figure 3 - Ethnicity category Vs Number of patients

4.1.3 Distribution of Histology Vs Number of patients:

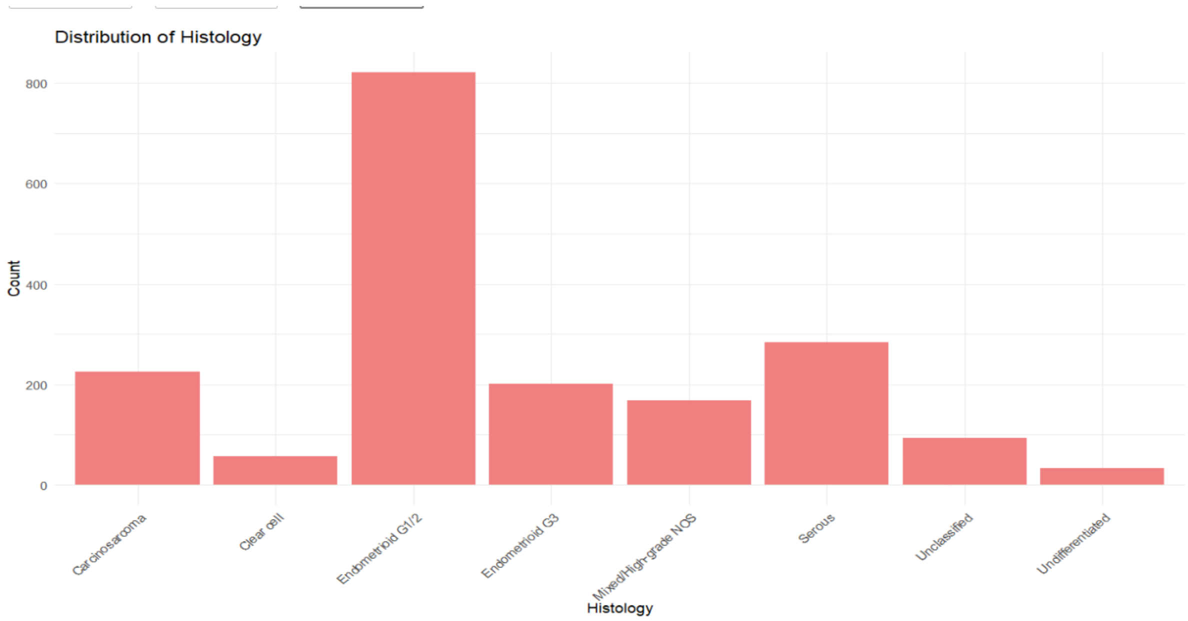


Figure 4 - Histology vs Number of patients

The histology plot (Fig 4) reveals that "Endometrioid G1/2" dominates the dataset, followed by "Serous" and "Carcinosarcoma." Cancer treatment relies heavily on histological classification since different types frequently have diverse genetic alterations and treatment responses. The literature supports combining histology and genomic data to improve predictive accuracy in cancer classification algorithms (Unwin, 2018).

4.2 Investigate relationships between categorical variables

4.2.1 Cancer type detailed vs Ethnicity

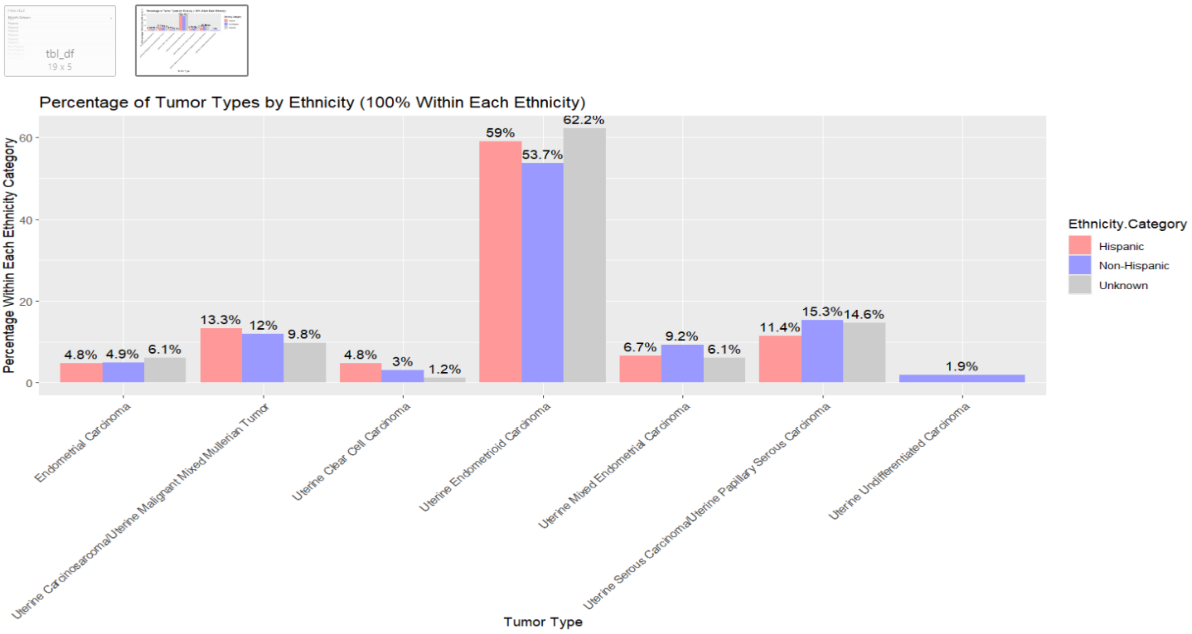


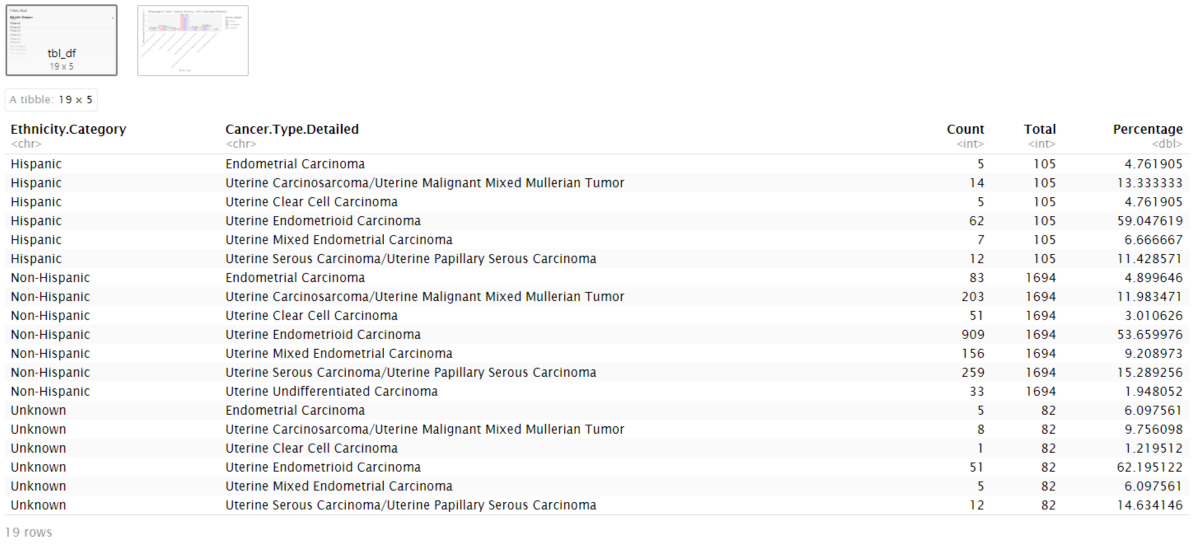
Figure 5 - Cancer type detailed vs Ethnicity

The graphic representation (Fig 5) shows the link between cancer types and ethnicity. This graph depicts how various cancer types are distributed across Hispanic, non-Hispanic, and unknown ethnic groups. The statistics show that uterine endometrioid carcinoma is most common in all groups, but especially among non-Hispanics (62.2%), with significant presence in the Hispanic (59%) and Unknown (53.7%) categories as well. The insights from this distribution are consistent with the findings of Higgins et al. (2023), who addressed how students struggled to evaluate categorical data with numerous levels. Similarly, the interpretation in this case necessitates comparing percentages within each group to draw relevant conclusions.

Hispanics have a significantly greater incidence of uterine carcinosarcoma or uterine malignant mixed Mullerian tumour (13.3%) than non-Hispanics (9.8%). This finding might suggest an ethnicity-based likelihood to this type of cancer. This categorical breakdown resembles the conceptual challenges identified by Higgins et al., where the complexity of multi-level category variables creates difficulties in capturing such patterns. (Higgins *et al.*, 2023)

Non-Hispanics (15.3%) have a higher incidence of uterine serous carcinoma/uterine papillary serous carcinoma than Hispanics (11.4%) or Unknown ethnicities (14.6%). This distribution highlights the importance of looking at patterns at many levels within the dataset, as suggested by Higgins et al. (2023). Understanding these correlations requires conditioning based on both ethnicity and cancer type.

This table in Fig 6 shows the distribution of various cancer types across different ethnic categories (Hispanic, Non-Hispanic, and Unknown) and their respective percentage.



A tibble: 19 × 5

Ethnicity.Category	Cancer.Type.Detailed	Count	Total	Percentage
<chr>	<chr>	<int>	<int>	<dbl>
Hispanic	Endometrial Carcinoma	5	105	4.761905
Hispanic	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor	14	105	13.333333
Hispanic	Uterine Clear Cell Carcinoma	5	105	4.761905
Hispanic	Uterine Endometrioid Carcinoma	62	105	59.047619
Hispanic	Uterine Mixed Endometrial Carcinoma	7	105	6.666667
Hispanic	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma	12	105	11.428571
Non-Hispanic	Endometrial Carcinoma	83	1694	4.899646
Non-Hispanic	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor	203	1694	11.983471
Non-Hispanic	Uterine Clear Cell Carcinoma	51	1694	3.010626
Non-Hispanic	Uterine Endometrioid Carcinoma	909	1694	53.659976
Non-Hispanic	Uterine Mixed Endometrial Carcinoma	156	1694	9.208973
Non-Hispanic	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma	259	1694	15.289256
Non-Hispanic	Uterine Undifferentiated Carcinoma	33	1694	1.948052
Unknown	Endometrial Carcinoma	5	82	6.097561
Unknown	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor	8	82	9.756098
Unknown	Uterine Clear Cell Carcinoma	1	82	1.219512
Unknown	Uterine Endometrioid Carcinoma	51	82	62.195122
Unknown	Uterine Mixed Endometrial Carcinoma	5	82	6.097561
Unknown	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma	12	82	14.634146

19 rows

Figure 6 - Percentage table of ethnicity by Cancer type detailed

4.2.2 Correlations and Pairwise Relationships

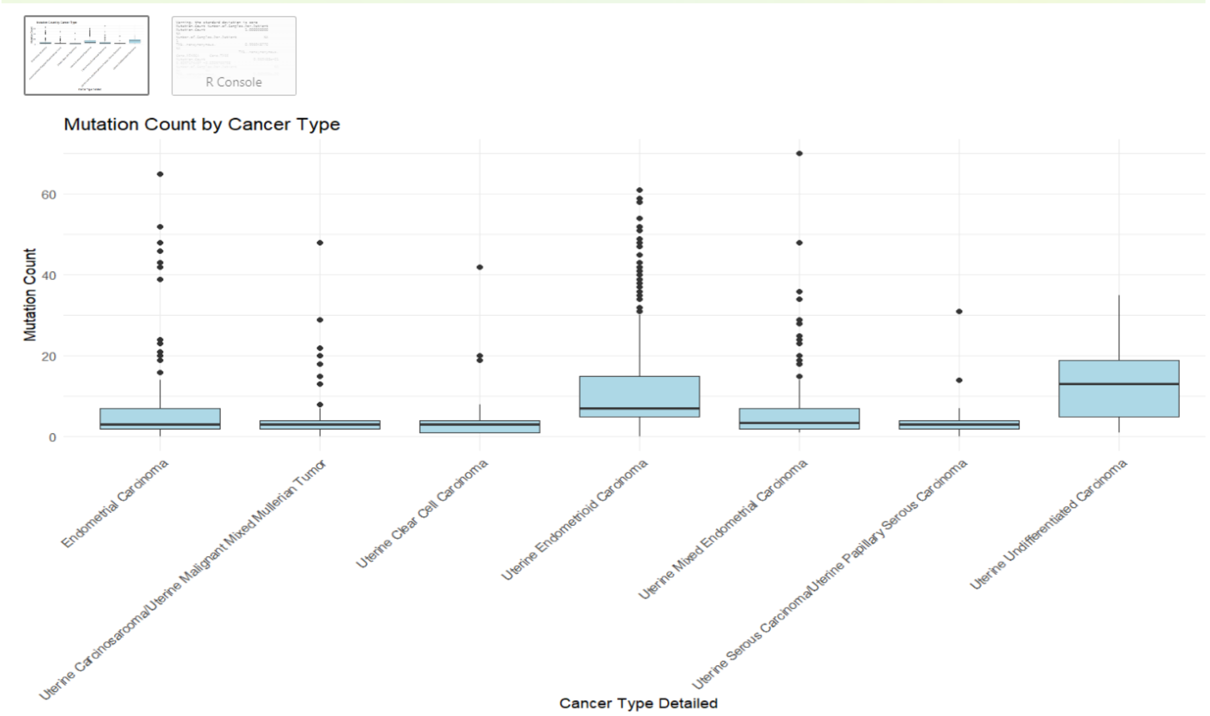


Figure 7 - Mutation Count Vs Cancer type detailed

The box plot (Fig 7) of mutation count by cancer type depicts the distribution of mutations across various uterine cancers, including endometrial carcinoma, uterine malignant mixed Mullerian tumour, uterine clear cell carcinoma, and others. The boxes reflect the interquartile range (IQR), while the lines inside denote the median mutation count.

The outliers in each cancer type, depicted by dots beyond the whiskers, are tumours with an extremely high number of mutations. For example, "Uterine Undifferentiated Carcinoma" has a wide range of mutation counts and numerous outliers, indicating greater mutation variability.

This change in mutation count can be essential in understanding the heterogeneity among cancer subtypes, validating prior results that mutation load can differ greatly between tumour types (Zhang & Wang, 2017).

4.3 cluster analysis of patient’s profile based on histology, ethnicity and cancer type detailed, and mutation count

Determine the optimal number of clusters using the elbow method:

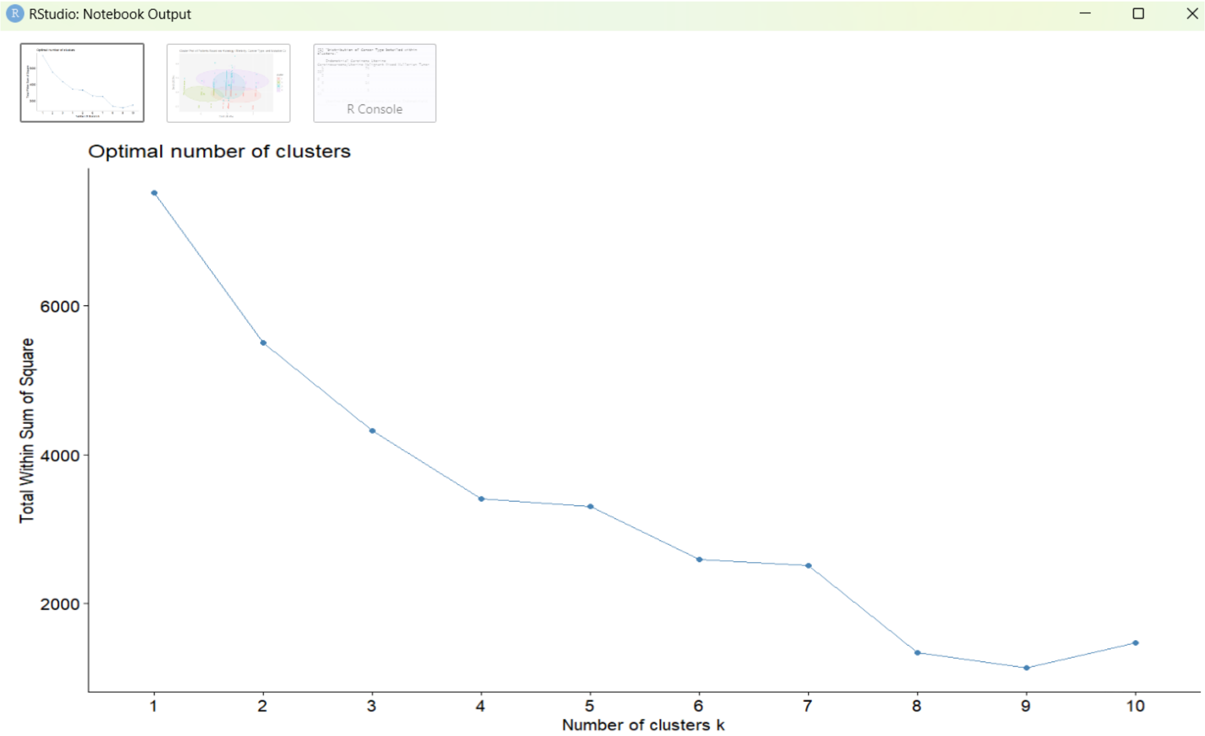


Figure 8 - Optimal Number of clusters

K-means clustering is a popular unsupervised learning algorithm that divides data into discrete groups (or clusters) based on similarity. It operates by creating a fixed number of clusters (K) and iteratively allocating each data point to the cluster with the closest centroid (centre). The algorithm repeats this procedure, updating the centroids and reassigning points until the centroids stabilise or there are no more changes to the cluster assignment (Cui, 2020).

The process starts with the selection of K initial cluster centres, which can be done randomly or based on some criterion. Euclidean distance is commonly used to measure the similarity of data points, with those with shorter distances being considered comparable and assigned to the same cluster. The goal is to reduce the sum of squared distances between each location and its cluster centroid. One of the most difficult aspects of K-means is determining the ideal number of clusters. This is often handled using the Elbow Method, which entails running the algorithm for various K values and displaying the Within-Cluster Sum-of-Squares (WCSS) vs the number of clusters. The "elbow" point on the graph, where the rate of fall in WCSS drops significantly, represents the optimal number of clusters. At this point, increasing the number of clusters gives diminishing returns in terms of clustering performance.

By adjusting the K value at the elbow, the algorithm may effectively strike a compromise between underfitting (too few clusters) and overfitting. This strategy assures that the clusters created are meaningful and representational of the data structure (Cui, 2020).

In the elbow plot (Fig 8), the WSS generally decreases as the number of clusters increases, because more clusters will naturally reduce the distance between data points and their assigned cluster centroids. The "elbow" is the point at which adding more clusters doesn’t significantly decrease the WSS, indicating diminishing returns in terms of data compactness.

The plot depicts how the elbow approach is used to determine the ideal number of clusters in a dataset.

The X-axis (Number of clusters k) represents the number of clusters, which ranges from 1 to 10 in this case. As we increase the value of k, the clustering algorithm attempts to partition the dataset into a greater number of groups.

The Y-axis (Total Within Sum of Squares) reflects the total within-cluster sum of squares (WSS), which is a measure of variance or how evenly distributed the points are within each cluster. A lower value of WSS indicates that the data points within each cluster are closer together.

The elbow technique calculates the ideal number of clusters by graphing the WSS against the number of clusters. In this graph, as k increases, the WSS reduces, but at some point, the decline is small. The "elbow" or point at which the curve bends is regarded the optimal amount of clusters, as adding more clusters does not appreciably reduce the WSS beyond this point. In this scenario, the elbow appears at k=4, implying that 4 clusters may be the best choice.

This technique and analysis are described in numerous clustering papers, including efficient implementations like the filtering algorithm for k-means clustering, which can assist minimise the time complexity in high-dimensional datasets (Kanungo *et al.*, 2002)

4.4 Cluster plot of patients based on Histology, ethnicity, cancer type and mutation count:

The cluster plot (Fig 9) of patients by histology, ethnicity, cancer type, and mutation count depict how different clusters group comparable patients, with each cluster representing a unique combination of these variables. The technique uses k-means clustering to divide the data into four groups (represented by various shapes and colours).

Each cluster can be read as follows.

Cluster 1 (red): Contains patients with similar cancer types and low mutation numbers, possibly suggesting less dangerous forms of cancer. The red cluster may correspond to distinct histological categories that have similar mutation patterns.

Cluster 2 (green): This cluster may reflect a population with moderate mutation numbers, presumably associated with specific ethnic groups or histological types exhibiting moderate genetic changes.

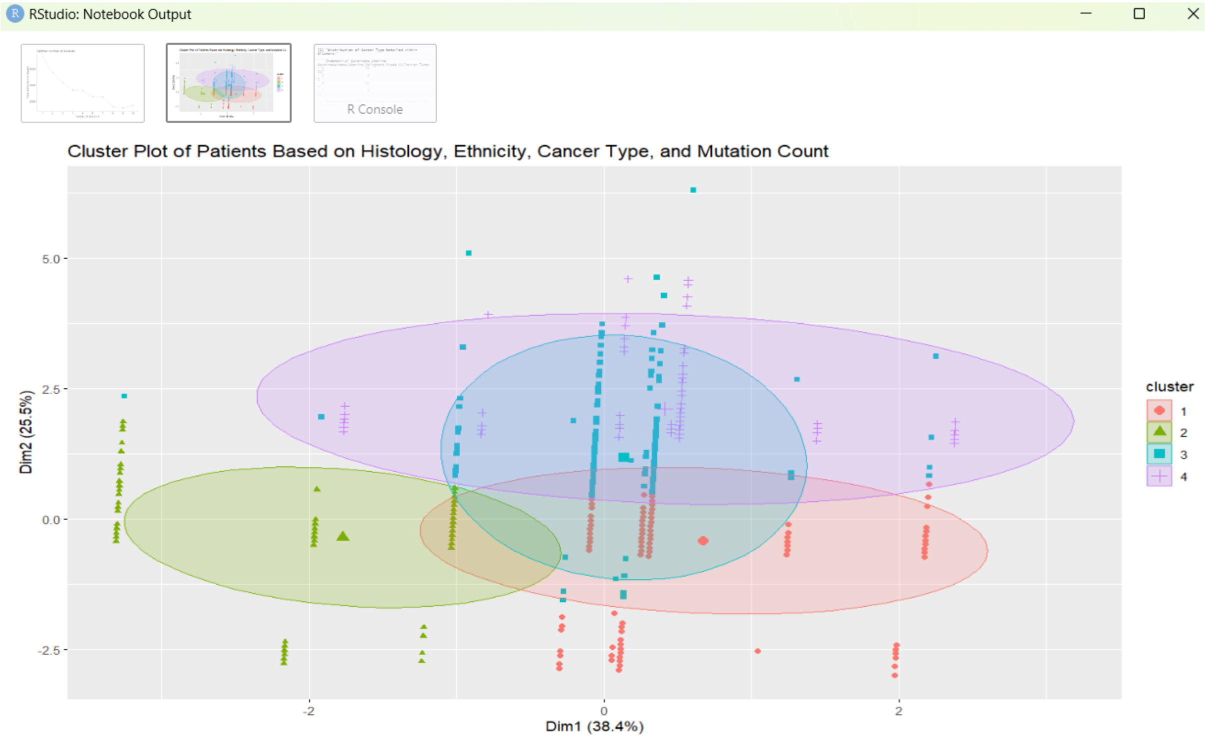


Figure 9 - Clustering of patients profile based on histology, Ethnicity, cancer type and mutation count

Cluster 3 (blue): This cluster looks to have greater mutation counts, indicating patients with more aggressive or advanced tumours, such as Uterine Serous Carcinoma, which is recognised for its high mutation rate.
Cluster 4 (purple): It may reflect a more genetically heterogeneous group of individuals, with mutations spanning various gene types. This could indicate variability in cancer progression or response to therapy. The use of k-means clustering for categorical (histology, ethnicity) and numeric (mutation count) data provides insights into how these characteristics interact across patient groups.

The points in the cluster figure are vertically aligned because some of the clustering factors (most likely category ones like Cancer Type, Ethnicity, and Histology) were numerically encoded. In one-hot encoding or numerical conversion of categorical data, each unique category is assigned a numerical value, which may lead the points to align vertically if these variables dominate the clustering or are handled as continuous numeric variables in the algorithm.

Because these category variables have only a few possible values, the plot points will frequently align with specific vertical axes. This means that for any unique value of a categorical variable, all data points with that value will be plotted along the same vertical line for that variable, resulting in vertical alignment of points.
In contrast, if numeric variables such as Mutation Count were more diverse, the points would spread out more horizontally or diagonally. This dynamic handling of mixed categorical and

numeric data in clustering is discussed in detail by Ahmad and Dey (2007) (Ahmad and Dey, 2007).

In this cluster plot, the **X-axis (Dim1)** and **Y-axis (Dim2)** represent the principal components obtained from a dimensionality reduction technique, such as Principal Component Analysis (PCA). These dimensions are combinations of the original features (i.e., cancer type, ethnicity, histology, and mutation count) and are used to explain the maximum variance in the dataset.

- **X-axis (Dim1)** represents the first principal component. It captures the largest amount of variance (38.4% in this case) from the high-dimensional data. Essentially, it is a weighted combination of the original variables that explains the most significant differences between the data points (patients) based on the features selected.
- **Y-axis (Dim2)** represents the second principal component, which captures the second-largest amount of variance (25.5%). Like Dim1, it is a weighted combination of the original variables, but it explains a different aspect of the data variability that is not captured by Dim1.

These components allow the data to be visualized in two dimensions, making it easier to interpret patterns and groupings (clusters). Each point represents a patient profile, and their position along the X and Y axes shows how their combination of cancer type, ethnicity, histology, and mutation count compares to others in the dataset.

The clusters formed in the plot show patients that have similar profiles based on the selected features. The ellipses around each cluster help to visually differentiate the clusters and give an idea of the spread or variability within each group.

4.4.2 Distribution of patients based on cancer type, ethnicity, and histology within four distinct clusters:

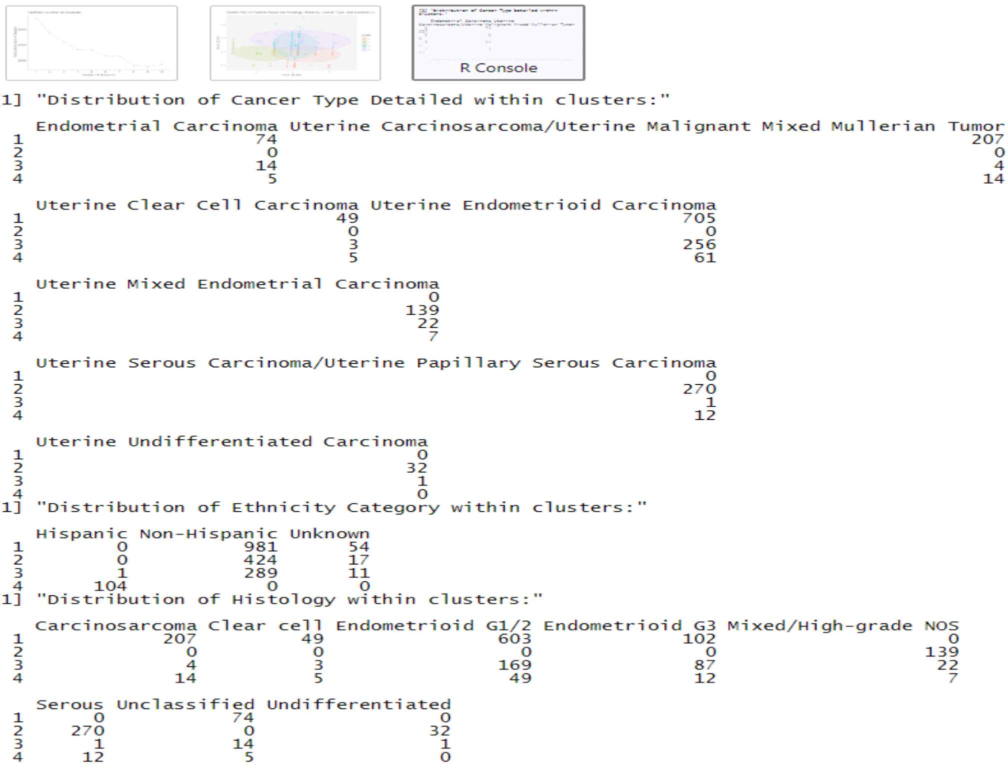
The table in Fig 10 indicate the distribution of patients based on cancer type, ethnicity, and histology into four separate groups. Here's a breakdown of the main findings in each category:

4.4.2.1 Cancer Type Distribution:

Cluster 1 primarily includes individuals with " Uterine Endometroid Carcinoma" (705) and "Carcinosarcoma/Malignant Mixed Mullerian Tumour" (207).

Cluster 2 has the majority of its patients diagnosed with "Uterine Mixed Endometrial Carcinoma" (139) and “Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma” (270).

Cluster 3 consists largely of individuals with "Uterine Clear Cell Carcinoma Uterine Endometrioid Carcinoma" (256)



```
[1] "Distribution of Cancer Type Detailed within clusters:"

  Endometrial Carcinoma Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
1                    74                                                             207
2                     0                                                               0
3                    14                                                               4
4                     5                                                              14

  Uterine Clear Cell Carcinoma Uterine Endometrioid Carcinoma
1                           49                            705
2                            0                              0
3                            3                            256
4                            5                             61

  Uterine Mixed Endometrial Carcinoma
1                                   0
2                                 139
3                                  22
4                                   7

  Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
1                                                           0
2                                                         270
3                                                           1
4                                                          12

  Uterine Undifferentiated Carcinoma
1                                    0
2                                   32
3                                    1
4                                    0
[1] "Distribution of Ethnicity Category within clusters:"

  Hispanic Non-Hispanic Unknown
1        0          981      54
2        0          424      17
3        1          289      11
4      104            0       0
[1] "Distribution of Histology within clusters:"

  Carcinosarcoma Clear cell Endometrioid G1/2 Endometrioid G3 Mixed/High-grade NOS
1            207         49               603             102                    0
2              0          0                 0               0                  139
3              4          3               169              87                   22
4             14          5                49              12                    7

  Serous Unclassified Undifferentiated
1      0           74                0
2    270            0               32
3      1           14                1
4     12            5                0
```

Figure 10 - Distribution of patients based on Histology, ethnicity, cancer type detailed

Cluster 4 is quite small, with a mix of cases from various types, including some Endometrial Carcinoma, Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumour, Uterine Clear Cell Carcinoma (5), Uterine Endometrioid Carcinoma (61).

4.4.2.2 Distribution of Ethnicity Category within Clusters:

Cluster 1 consists of 981 non-Hispanic individuals and 54 patients of unknown ethnicity. There are no Hispanic patients in this cluster.

Cluster 2: This cluster has 424 non-Hispanic patients, 17 unknown patients, and 0 Hispanic patients.

Cluster 3 has 289 non-Hispanic, 11 unknown, and 1 Hispanic patients. It is the only cluster with a Hispanic patient; however the number is relatively small.

Cluster 4: This small cluster has just 104 Hispanic patients, with no Non-Hispanic or Unknown ethnicity patients.

4.4.2.3 Distribution of Histology within Clusters:

Cluster 1: This cluster has a high prevalence of **Endometrioid G1/2** (603 cases), followed by **Carcinosarcoma** (207 cases) and **Clear Cell** (49 cases). There are also 102 cases of **Endometrioid G3**, Unclassified (74) and while there are no cases of **Mixed/High-grade NOS**, serous and undifferentiated.

Cluster 2: This cluster is composed of patients with **Mixed/High-grade NOS** (139 cases) and **serous** (270).

Cluster 3: The largest representation in this cluster is **Endometrioid G1/2** (169 cases), followed by **Endometrioid G3** (87 cases). There are smaller numbers of **Clear Cell** (3 cases) and **Mixed/High-grade NOS** (22 cases), and smaller number of other types aswell.

Cluster 4: This smaller cluster shows a mixture of histology, including **Endometrioid G1/2** (49 cases), **Carcinosarcoma** (14 cases), and **Clear Cell** (5 cases), with a few cases of others.

4.5 cluster analysis of patients profile based on Histology, Ethnicity, Cancer Type Detailed and all gene columns

The elbow point (Fig 11) on represents the appropriate number of clusters at which the drop in WSS begins to slow. According to this plot, the elbow point is around four clusters, indicating that this is likely the best number of clusters to use for this dataset. After this point, adding more clusters does not appreciably reduce the WSS, implying that the improvement in clustering quality fades.

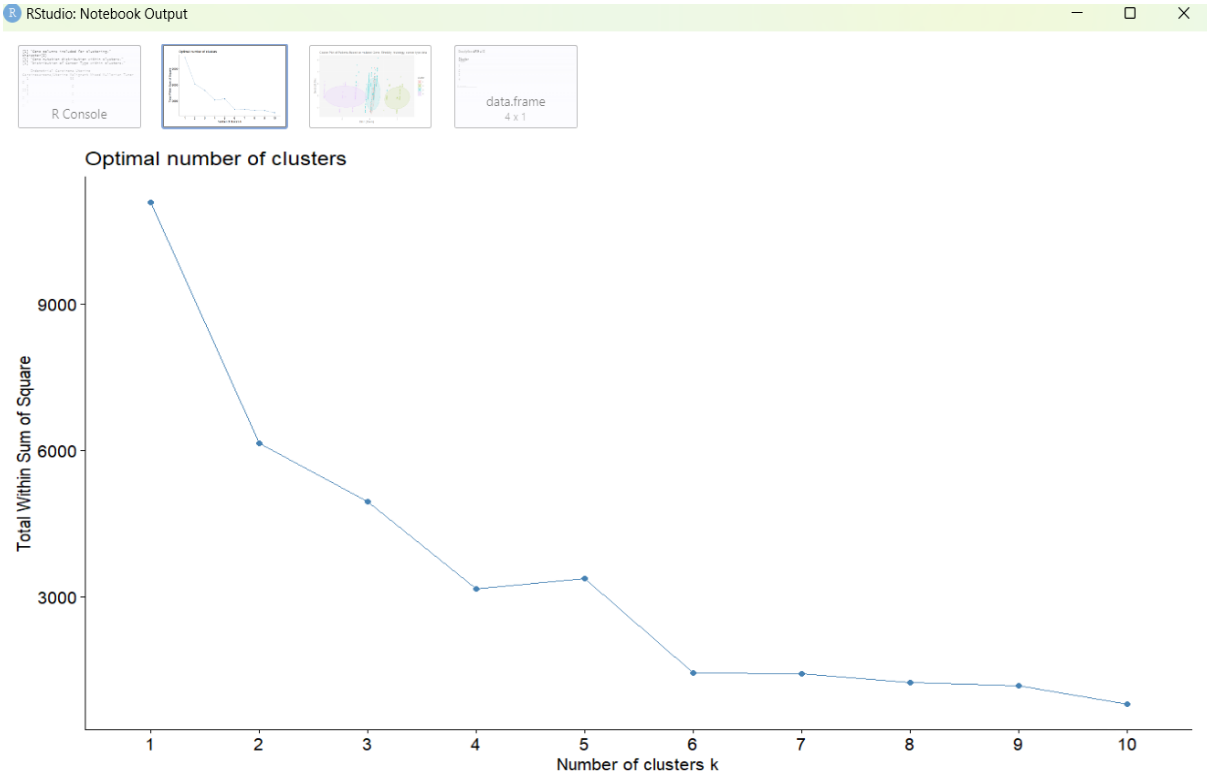


Figure 11 - Elbow Chart

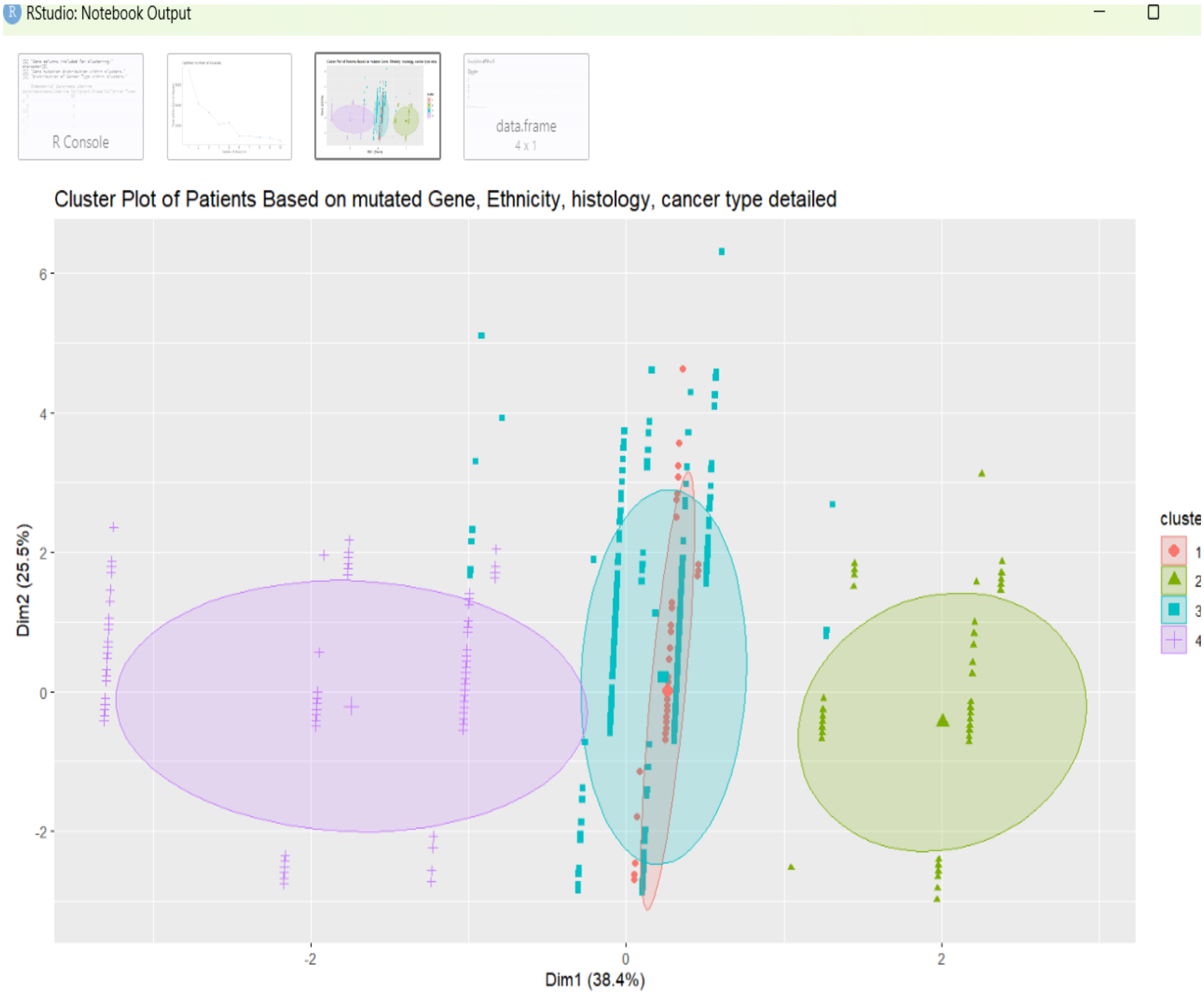
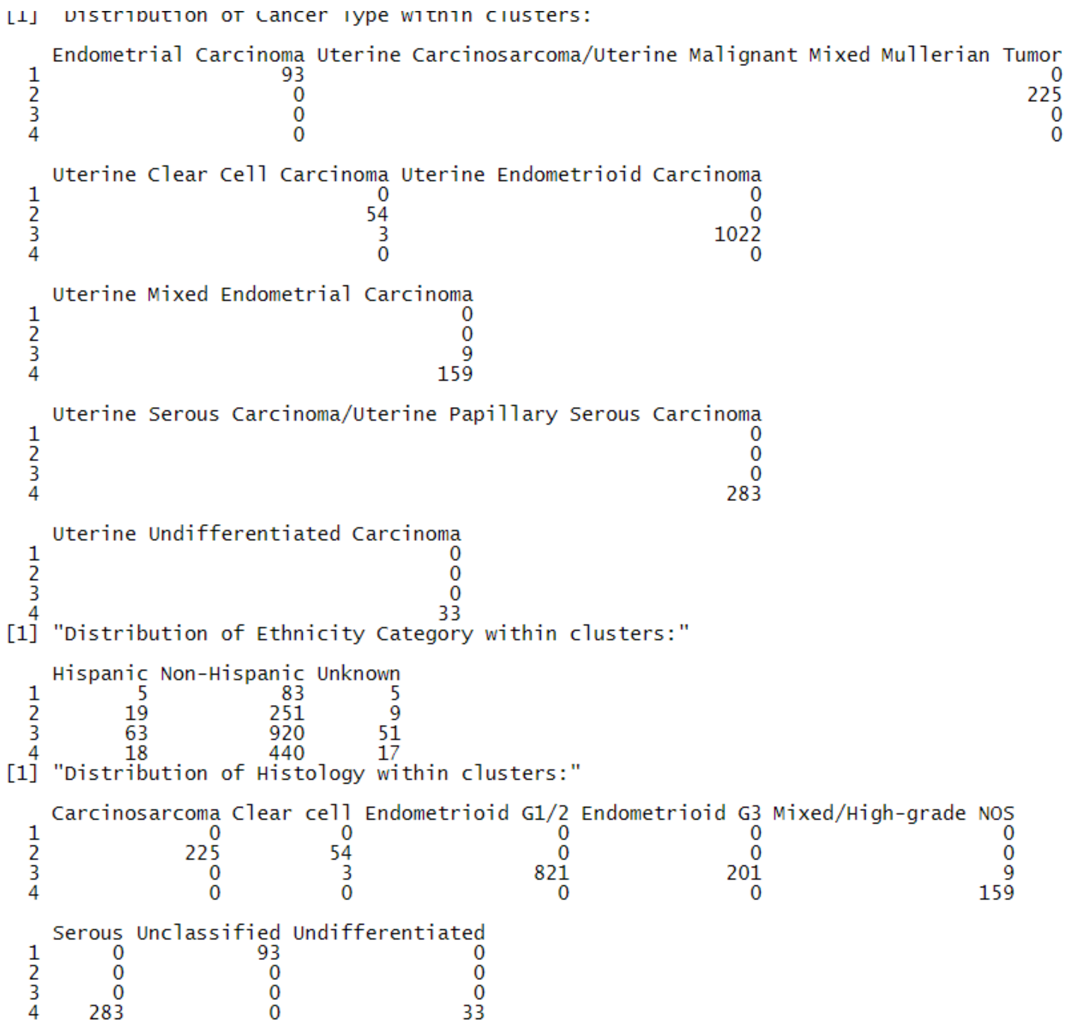


Figure 12 - Cluster of patients profile based on mutated gene, Ethnicity, Histology ana cancer type detailed

Fig 12 clusters patients based on gene mutation data as well as clinical characteristics such as cancer types, ethnicity, and histology. The tightness and overlap of the ellipses indicate how similar the individuals in each cluster are in terms of gene mutations and clinical data. Some clusters overlap, indicating shared characteristics or similarities among patient groups.



```
[1]  Distribution of Cancer Type within clusters:

  Endometrial Carcinoma Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
1                    93                                                              0
2                     0                                                            225
3                     0                                                              0
4                     0                                                              0

  Uterine Clear Cell Carcinoma Uterine Endometrioid Carcinoma
1                            0                              0
2                           54                              0
3                            3                           1022
4                            0                              0

  Uterine Mixed Endometrial Carcinoma
1                                   0
2                                   0
3                                   9
4                                 159

  Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
1                                                           0
2                                                           0
3                                                           0
4                                                         283

  Uterine Undifferentiated Carcinoma
1                                  0
2                                  0
3                                  0
4                                 33
[1] "Distribution of Ethnicity Category within clusters:"

  Hispanic Non-Hispanic Unknown
1        5           83       5
2       19          251       9
3       63          920      51
4       18          440      17
[1] "Distribution of Histology within clusters:"

  Carcinosarcoma Clear cell Endometrioid G1/2 Endometrioid G3 Mixed/High-grade NOS
1              0          0                 0               0                    0
2            225         54                 0               0                    0
3              0          3               821             201                    9
4              0          0                 0               0                  159

  Serous Unclassified Undifferentiated
1      0           93                0
2      0            0                0
3      0            0                0
4    283            0               33
```

Figure 13 - Distribution of patients based on Gene mutated, Histology, ethnicity, cancer type detailed

Fig 13 provides data on several forms of uterine cancer and their corresponding clusters. The breakdown of each cluster is mentioned below:

4.5.1 Cancer Type Distribution:

Cluster 1: This cluster is dominated by cases of **endometrial carcinoma** (93). No other types of cancer are observed in this cluster.

Cluster 2: This cluster mainly contains **uterine carcinosarcoma/uterine malignant mixed Mullerian tumor** (225) and **uterine clear cell carcinoma** (54). No other cancer types are present in this cluster.

Cluster 3: This cluster contains small numbers of **uterine clear cell carcinoma** (3), **uterine endometrioid carcinoma** (1022), and **uterine mixed endometrial carcinoma (9).** No other types of cancer are present in this cluster.

Cluster 4: This cluster is primarily composed of uterine mixed endometrial carcinoma (159), uterine serous carcinoma/uterine papillary serous carcinoma (283), and uterine undifferentiated carcinoma (33). There are no additional forms of cancer in this cluster.

4.5.2 Distribution of Ethnicity Category within Clusters:

Cluster 1: This cluster is predominantly **Non-Hispanic** (83), with very few **Hispanic** (5) and **Unknown** (5) ethnicities.

Cluster 2: This cluster also has a majority of **Non-Hispanic** individuals (251), with a smaller number of **Hispanic** (19) and **Unknown** (9) ethnicities.

Cluster 3: This is the largest cluster overall, and it is predominantly **Non-Hispanic** (920), with a moderate number of **Hispanic** individuals (63) and a fair number of **Unknown** (51)

Cluster 4: This cluster is primarily **Non-Hispanic** (440), with a small number of **Hispanic** (18) and **Unknown** (17) ethnicities.

Across all clusters, the **Non-Hispanic** category dominates, with varying proportions of **Hispanic** and **Unknown** individuals. Cluster 3 stands out as being much larger than the others, with higher representation across all categories.

4.5.3 Distribution of Histology within Clusters:

Cluster 1: This cluster contains **93 cases of undifferentiated carcinoma**, with no other histology types present.

Cluster 2: This cluster contains 225 cases of carcinosarcoma and 54 cases of clear cell carcinoma. No other histology types are present in this cluster.

Cluster 3: This is the largest cluster, consisting largely of 821 Endometrioid G1/2 carcinoma, 201 Endometrioid G3 carcinoma, and 9 Mixed/High-grade NOS carcinomas. There are three cases of clear cell carcinoma. There are no other histological types present.

Cluster 4: This cluster contains **283 cases of serous carcinoma**, **159 cases of Mixed/High-grade NOS carcinoma**, and **33 cases of undifferentiated carcinoma**. No other histology types are present in this cluster.

Differences Between and both the clusters:

Variables used for clustering: Fig12 uses gene mutation data along with Histology, Ethnicity, Cancer Type , Fig 9 primarily focuses on clinical features like Histology, Ethnicity, Cancer Type, and Mutation Count. This difference in input data leads to variations in how patients are grouped**.**

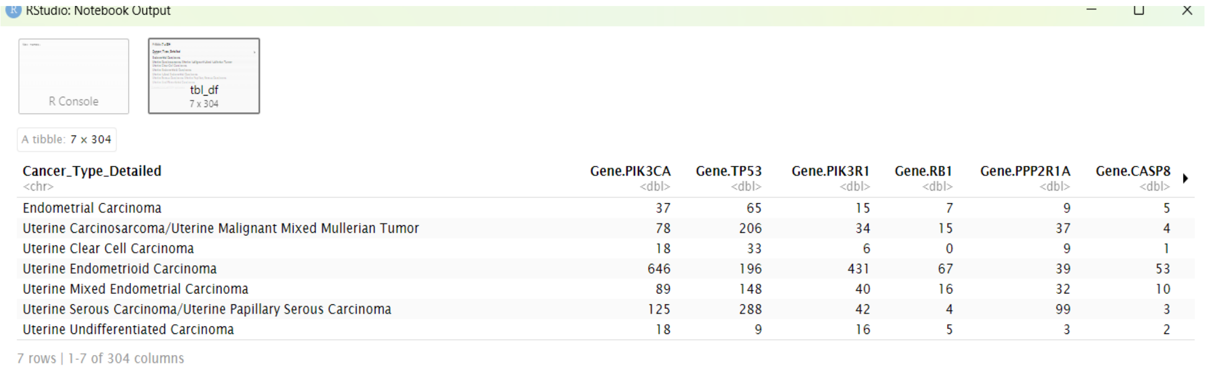
Shape of clusters: The clusters in the Fig 12 are more closely clustered around specific data ranges, which is most likely due to gene mutation trends. The Fig 9, on the other hand, displays larger or differently formed clusters because of the emphasis on clinical aspects.

4.6 classifying the patient's cancer type based on the gene mutation:

"A Random Forest (RF) is an ensemble of decision trees that creates several decision trees and then aggregates them to provide a final forecast. RF employs the bagging technique and random feature selection, making it ideal for lowering data dimensionality and boosting generalisation performance. It has been successfully employed in several fields, including disease prediction and genetics" (Yin *et al.*, 2019).

- **Random Forest** is useful in classifying cancer types because of its ability to handle high-dimensional data, its resistance to overfitting, and its feature importance metrics.

- For gene mutation data, it can highlight the most relevant genes and classify the type of cancer with high accuracy, making it a valuable tool in medical research and clinical diagnosis.



A tibble: 7 x 304

Cancer_Type_Detailed <chr>	Gene.PIK3CA <dbl>	Gene.TP53 <dbl>	Gene.PIK3R1 <dbl>	Gene.RB1 <dbl>	Gene.PPP2R1A <dbl>	Gene.CASP8 <dbl>
Endometrial Carcinoma	37	65	15	7	9	5
Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor	78	206	34	15	37	4
Uterine Clear Cell Carcinoma	18	33	6	0	9	1
Uterine Endometrioid Carcinoma	646	196	431	67	39	53
Uterine Mixed Endometrial Carcinoma	89	148	40	16	32	10
Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma	125	288	42	4	99	3
Uterine Undifferentiated Carcinoma	18	9	16	5	3	2

7 rows | 1-7 of 304 columns

Figure 14 - Cancer types Vs Number of times each gene mutated

The table in the Fig 14 provides a summary of how often certain genes are mutated in different types of uterine cancer. It highlights:

- The differences in gene mutation profiles between different cancer types.
- Which genes (like **TP53** or **PIK3CA**) have more frequent mutations in certain cancers (like Endometrioid or Serous carcinomas).

This knowledge is useful for understanding the molecular characteristics of certain cancers, and it may help guide targeted therapy based on specific gene alterations.

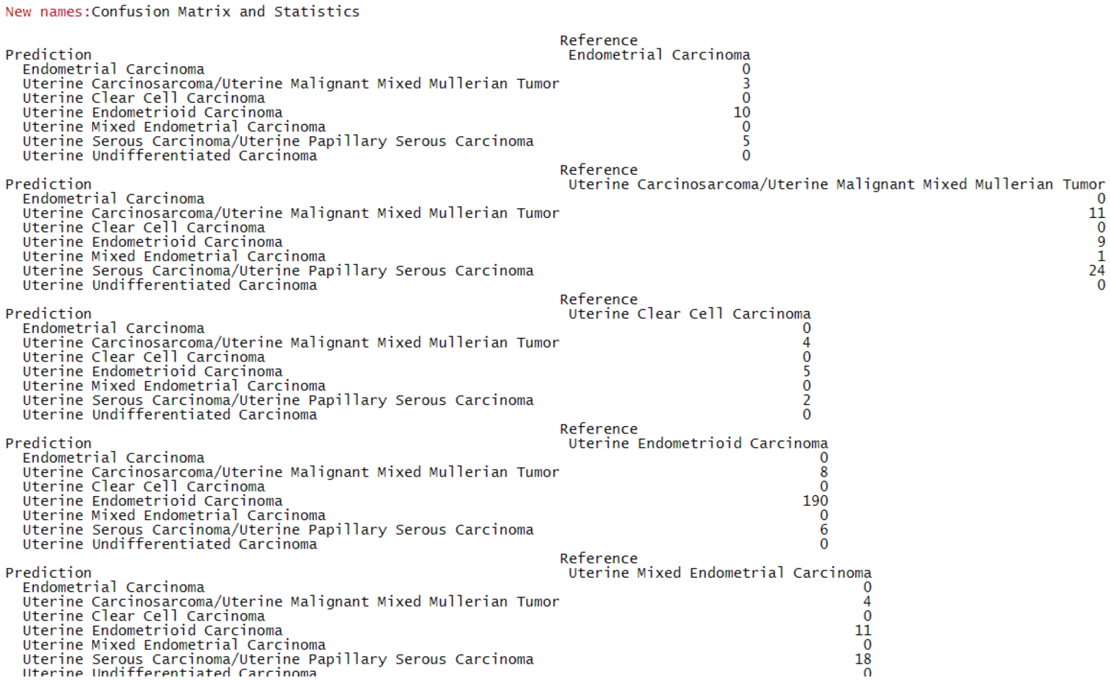
1. Confusion Matrix and Statistics:

The confusion matrix evaluates the Random Forest model's performance in diagnosing various forms of uterine cancer based on gene mutations. Each **row** corresponds to the **predicted class**, whereas each **column** represents the actual **class (reference)**. The model's purpose is to accurately classify each form of cancer using genetic data.

For example:

In the first block under Reference in Fig 15 **Endometrial Carcinoma**, the model **did not correctly predict any cases** of **Endometrial Carcinoma**. Instead, it incorrectly predicted:

- 3 cases of Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumour.
- 10 cases of Uterine Endometrioid Carcinoma.
- 5 cases of Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma.



```
New names:Confusion Matrix and Statistics

                                                                Reference
Prediction                                                       Endometrial Carcinoma
  Endometrial Carcinoma                                                              0
  Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor                     3
  Uterine Clear Cell Carcinoma                                                       0
  Uterine Endometrioid Carcinoma                                                    10
  Uterine Mixed Endometrial Carcinoma                                                0
  Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma                        5
  Uterine Undifferentiated Carcinoma                                                 0
                                                                Reference
Prediction                                                       Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
  Endometrial Carcinoma                                                                                                       0
  Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor                                                             11
  Uterine Clear Cell Carcinoma                                                                                                0
  Uterine Endometrioid Carcinoma                                                                                              9
  Uterine Mixed Endometrial Carcinoma                                                                                         1
  Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma                                                                24
  Uterine Undifferentiated Carcinoma                                                                                          0
                                                                Reference
Prediction                                                       Uterine Clear Cell Carcinoma
  Endometrial Carcinoma                                                                     0
  Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor                            4
  Uterine Clear Cell Carcinoma                                                              0
  Uterine Endometrioid Carcinoma                                                            5
  Uterine Mixed Endometrial Carcinoma                                                       0
  Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma                               2
  Uterine Undifferentiated Carcinoma                                                        0
                                                                Reference
Prediction                                                       Uterine Endometrioid Carcinoma
  Endometrial Carcinoma                                                                       0
  Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor                              8
  Uterine Clear Cell Carcinoma                                                                0
  Uterine Endometrioid Carcinoma                                                            190
  Uterine Mixed Endometrial Carcinoma                                                         0
  Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma                                 6
  Uterine Undifferentiated Carcinoma                                                          0
                                                                Reference
Prediction                                                       Uterine Mixed Endometrial Carcinoma
  Endometrial Carcinoma                                                                            0
  Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor                                   4
  Uterine Clear Cell Carcinoma                                                                     0
  Uterine Endometrioid Carcinoma                                                                  11
  Uterine Mixed Endometrial Carcinoma                                                              0
  Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma                                     18
  Uterine Undifferentiated Carcinoma                                                               0
```

Figure 15 - Confusion Matrix

Similarly, in the second block it Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumour. The model correctly predicted **11 cases** of **Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumour**.

However, it also misclassified some cases:

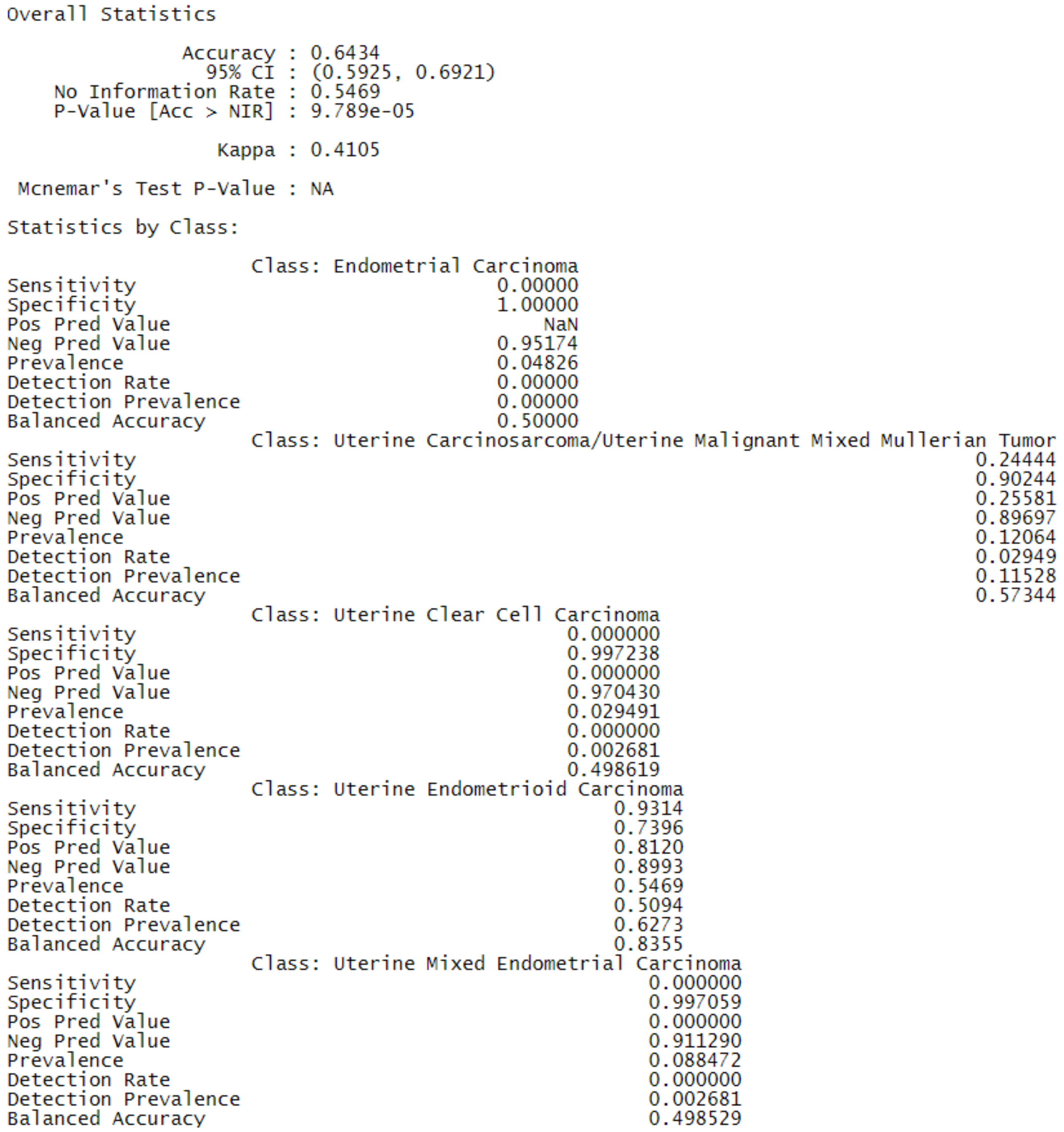
- **9 cases** were predicted as **Uterine Endometrioid Carcinoma**.
- **24 cases** were predicted as **Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma**.
- **1 case** was predicted as **Uterine Mixed Endometrial Carcinoma**.

This suggests that the model has some difficulty distinguishing particular tumour type for instance **Endometrial Carcinoma** from other types of uterine cancers, particularly **Uterine Endometrioid Carcinoma** and **Uterine Serous Carcinoma** etc., This is because some of the following reasons.

1. Gene mutations may overlap across cancer types. Many uterine malignancies, including uterine serous carcinoma and uterine carcinosarcoma, have comparable mutation patterns in genes such as TP53, PIK3CA, and PPP2R1A. The overlap in genetic traits can mislead the model, resulting in inaccurate predictions. If two cancer types have similar mutation patterns, the model may be unable to classify them correctly.

2. If the dataset is imbalanced, with some cancer types having much more data points than others, the model may be biassed towards predicting the more common types. For example, if there are far more cases of Uterine Serous Carcinoma than Uterine Carcinosarcoma, the model may likely to predict Uterine Serous Carcinoma more frequently, even when encountering cases of Uterine Carcinosarcoma.

3. Some genes may be more important for identifying some cancer forms than others. For example, if the model is largely reliant on a specific gene mutation found in numerous types of cancer, it may misclassify cases in which the gene mutation is shared by multiple types. The importance of genes such as TP53 (which is frequently altered in numerous cancer types) can lead to inaccurate predictions if they do not efficiently distinguish between classes.



```
Overall Statistics

               Accuracy : 0.6434
                 95% CI : (0.5925, 0.6921)
    No Information Rate : 0.5469
    P-Value [Acc > NIR] : 9.789e-05

                  Kappa : 0.4105

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Endometrial Carcinoma
Sensitivity                               0.00000
Specificity                               1.00000
Pos Pred Value                                NaN
Neg Pred Value                            0.95174
Prevalence                                0.04826
Detection Rate                            0.00000
Detection Prevalence                      0.00000
Balanced Accuracy                         0.50000
                     Class: Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Sensitivity                                                                        0.24444
Specificity                                                                        0.90244
Pos Pred Value                                                                     0.25581
Neg Pred Value                                                                     0.89697
Prevalence                                                                         0.12064
Detection Rate                                                                     0.02949
Detection Prevalence                                                               0.11528
Balanced Accuracy                                                                  0.57344
                     Class: Uterine Clear Cell Carcinoma
Sensitivity                                     0.000000
Specificity                                     0.997238
Pos Pred Value                                  0.000000
Neg Pred Value                                  0.970430
Prevalence                                      0.029491
Detection Rate                                  0.000000
Detection Prevalence                            0.002681
Balanced Accuracy                               0.498619
                     Class: Uterine Endometrioid Carcinoma
Sensitivity                                         0.9314
Specificity                                         0.7396
Pos Pred Value                                      0.8120
Neg Pred Value                                      0.8993
Prevalence                                          0.5469
Detection Rate                                      0.5094
Detection Prevalence                                0.6273
Balanced Accuracy                                   0.8355
                     Class: Uterine Mixed Endometrial Carcinoma
Sensitivity                                            0.000000
Specificity                                            0.997059
Pos Pred Value                                         0.000000
Neg Pred Value                                         0.911290
Prevalence                                             0.088472
Detection Rate                                         0.000000
Detection Prevalence                                   0.002681
Balanced Accuracy                                      0.498529
```

Figure 16 - Statistics Matrix

4.6.1 Class-Level Statistics:

The **overall statistics** in Fig 16 and **per-class performance metrics** from a classification model, likely the Random Forest model used to classify uterine cancer types based on gene mutations. Here, the class endometrial cancer is explained:

1. **Accuracy**: The model's accuracy is **0.6434**, meaning the model correctly predicted the cancer type about 64.34% of the time.

2. **95% CI**: This is the confidence interval for the accuracy, ranging from **0.5925 to 0.6921**. It shows the range within which the true accuracy is likely to lie.
3. **No Information Rate (NIR)**: **0.5469** indicates the accuracy of a simple model that always predicts the most frequent class. The model performs better than this baseline.
4. **P-Value (Acc > NIR)**: **9.789e-05** indicates that the model's accuracy is significantly better than random guessing (with a very low p-value, it's a strong indication).
5. **Kappa**: **0.4105** reflects the agreement between the actual and predicted classifications. A value above 0.4 indicates moderate agreement.

Sensitivity: The proportion of true positives correctly identified (recall). A sensitivity of 0 for "Endometrial Carcinoma" indicates that the model did not correctly identify any instances of this class.

Specificity :The proportion of true negatives correctly identified. For "Endometrial Carcinoma," specificity is 1.0, meaning the model correctly identified all instances that were not "Endometrial Carcinoma."

Pos Pred Value (Precision): The proportion of positive identifications that were actually correct. In this case, it's NaN because there were no true positive predictions for "Endometrial Carcinoma."

Neg Pred Value: The probability that a non-Endometrial Carcinoma prediction is correct, which is 0.9517 for "Endometrial carcinoma".

Balanced Accuracy: The average of sensitivity and specificity. For "Endometrial Carcinoma," the balanced accuracy is 0.5, showing poor performance in classifying this type.

Similarly for the **Class: Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor**

- **Sensitivity**: **0.2444** (The model correctly identified 24.44% of true cases).
- **Specificity**: **0.9024** (The model avoided false positives in 90.24% of non-cases).
- **Pos Pred Value**: **0.2558** (The probability that a predicted case is truly Uterine Carcinosarcoma).
- **Neg Pred Value**: **0.8969** (The probability that a non-case prediction is correct).
- **Balanced Accuracy**: **0.5734** (Shows moderate ability to identify this cancer type).

Mean Decrease Accuracy denotes how much the model's overall accuracy decreases when this gene is excluded. A large negative value indicates that the gene is crucial for accurate predictions

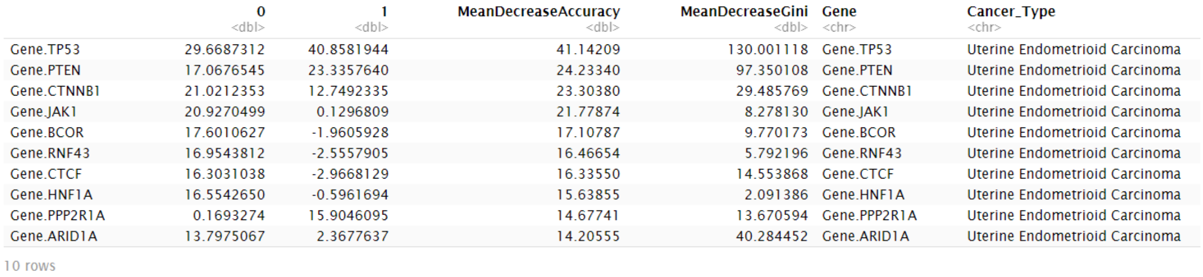
Mean Decrease Gini: A measure of the gene's importance in splitting data within the decision tree model (incase of using random forests). Larger values indicate higher importance for making predictions.

Each cancer type has its own list of genes, ranked by their importance according to these two metrics. For instance: **Uterine Undifferentiated Carcinoma**: The most important gene is **Gene.TP53**, with a Mean Decrease Accuracy of **51.35** and a high Gini importance score of 95.81.**TP53** is again a highly important gene with a Mean Decrease Accuracy of **51.35**, reflecting its significance in predicting this cancer type. Overall the model's accuracy is 64.34%

However, there are nearly 303 genes present in this dataset. So it is difficult to have to look on its contribution towards predicting each cancer type. So, training a Random Forest model for each cancer type in the dataset to identify the top 10 most important genes based on their contribution to predicting each specific cancer type. The top 10 genes are chosen because they are the most influential in improving the model’s accuracy, helping to focus on the most relevant genetic mutations for distinguishing between cancer types, thereby improving model interpretability and efficiency.

4.7 Classify the cancer type based on Top 10 most important genes

The Fig 17: below shows the results of a Random Forest model that has been trained to classify "Uterine Endometrioid Carcinoma" and extract the top 10 genes that contribute most to the classification accuracy. Let me explain each column:



	0	1	MeanDecreaseAccuracy	MeanDecreaseGini	Gene	Cancer_Type
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
Gene.TP53	29.6687312	40.8581944	41.14209	130.001118	Gene.TP53	Uterine Endometrioid Carcinoma
Gene.PTEN	17.0676545	23.3357640	24.23340	97.350108	Gene.PTEN	Uterine Endometrioid Carcinoma
Gene.CTNNB1	21.0212353	12.7492335	23.30380	29.485769	Gene.CTNNB1	Uterine Endometrioid Carcinoma
Gene.JAK1	20.9270499	0.1296809	21.77874	8.278130	Gene.JAK1	Uterine Endometrioid Carcinoma
Gene.BCOR	17.6010627	-1.9605928	17.10787	9.770173	Gene.BCOR	Uterine Endometrioid Carcinoma
Gene.RNF43	16.9543812	-2.5557905	16.46654	5.792196	Gene.RNF43	Uterine Endometrioid Carcinoma
Gene.CTCF	16.3031038	-2.9668129	16.33550	14.553868	Gene.CTCF	Uterine Endometrioid Carcinoma
Gene.HNF1A	16.5542650	-0.5961694	15.63855	2.091386	Gene.HNF1A	Uterine Endometrioid Carcinoma
Gene.PPP2R1A	0.1693274	15.9046095	14.67741	13.670594	Gene.PPP2R1A	Uterine Endometrioid Carcinoma
Gene.ARID1A	13.7975067	2.3677637	14.20555	40.284452	Gene.ARID1A	Uterine Endometrioid Carcinoma

10 rows

Figure 17 - Top 10 genes predicting the Uterine endometroid carcinoma

This Fig 17 depicts the feature relevance of various genes in categorising uterine endometrioid carcinoma using a Random Forest model. The table covers genes and their contributions to model performance using two essential metrics: MeanDecreaseAccuracy and MeanDecreaseGini.

Gene: The genes listed are those employed by the Random Forest model to classify uterine endometrioid carcinoma. These include TP53, PTEN, CTNNB1, and JAK1.

0 and 1: These columns appear to indicate numeric values or coefficients relating to the model's internal workings for these specific genes, but without additional context, it's unclear how these numbers directly connect to the important ratings. They may represent intermediate or gene-specific metrics in the model. MeanDecreaseAccuracy: This column indicates how much the overall accuracy of the Random Forest model lowers when a given gene is eliminated. Higher numbers suggest that the gene is critical to the model's accuracy.

- TP53 has the highest MeanDecreaseAccuracy of 41.14, indicating that it is the most critical gene for keeping the model accurate.
- Other relevant genes are PTEN (24.23) and CTNNB1 (23.30), which both contribute significantly to the model's accuracy.

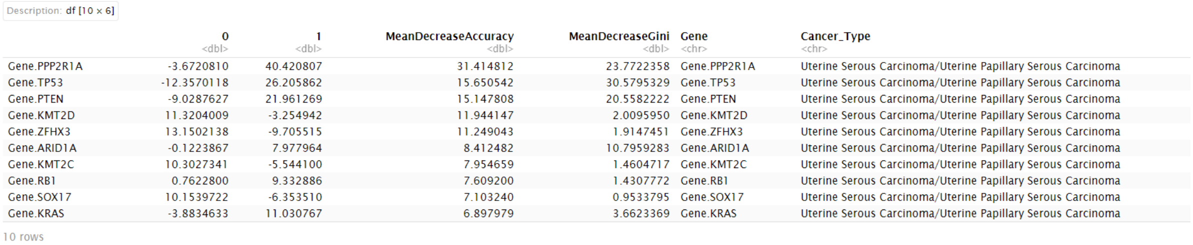
MeanDecreaseGini: This column indicates how much the gene contributes to increasing the "purity" of decision trees in the Random Forest model. Higher numbers suggest that the gene is critical for developing more distinct classifications (i.e., effectively distinguishing between cancer types).

- TP53 once again stands out with the greatest MeanDecreaseGini of 130.00, indicating that it is the most essential gene in the model for distinguishing between classes.
- PTEN and CTNNB1 make considerable contributions to the model's decision-making process, with Gini scores of 97.35 and 29.48, respectively.

Cancer_Type: This table contains top 10 genes related with uterine endometrioid carcinoma. Based on the mutation data, the model classifies this cancer type using only these genes.

TP53, **PTEN**, and **CTNNB1** are the most important genes, contributing the most to both accuracy and decision-tree splitting in the model. Other genes like **JAK1** and **BCOR** also play notable roles but to a lesser extent.

Similarly for all the types of cancers the top10 genes were predicted mentioned on the figures below.



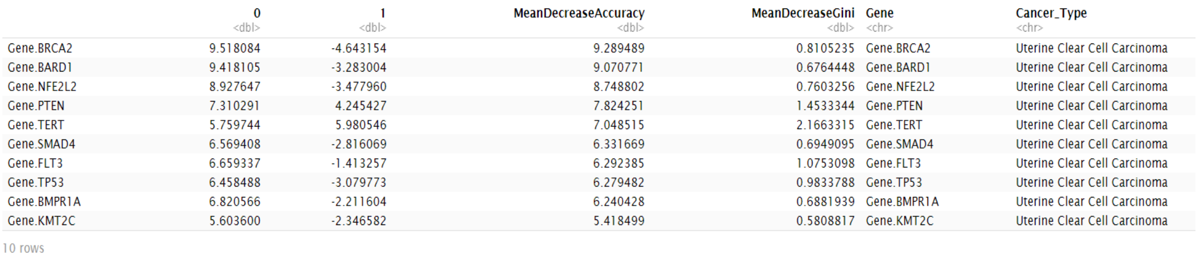
Description: df [10 x 6]

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini	Gene	Cancer_Type
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
Gene.PPP2R1A	-3.6720810	40.420807	31.414812	23.7722358	Gene.PPP2R1A	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.TP53	-12.3570118	26.205862	15.650542	30.5795329	Gene.TP53	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.PTEN	-9.0287627	21.961269	15.147808	20.5582222	Gene.PTEN	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.KMT2D	11.3204009	-3.254942	11.944147	2.0095950	Gene.KMT2D	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.ZFHX3	13.1502138	-9.705515	11.249043	1.9147451	Gene.ZFHX3	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.ARID1A	-0.1223867	7.977964	8.412482	10.7959283	Gene.ARID1A	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.KMT2C	10.3027341	-5.544100	7.954659	1.4604717	Gene.KMT2C	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.RB1	0.7622800	9.332886	7.609200	1.4307772	Gene.RB1	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.SOX17	10.1539722	-6.353510	7.103240	0.9533795	Gene.SOX17	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma
Gene.KRAS	-3.8834633	11.030767	6.897979	3.6623369	Gene.KRAS	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma

10 rows

Figure 18 - Top 10 genes predicting the Uterine serous carcinoma

Fig 18 shows the feature importance scores of various genes for classifying **Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma** using a Random Forest model. Key genes such as **PPP2R1A**, **TP53**, and **PTEN** have the highest **MeanDecreaseAccuracy** and **MeanDecreaseGini**, indicating that they contribute significantly to the model's classification accuracy and decision purity. **PPP2R1A** stands out as the most important gene, followed by **TP53** and **PTEN**, which are critical for accurate predictions. Other genes, like **KMT2D** and **ZFHX3**, also play important but lesser roles.

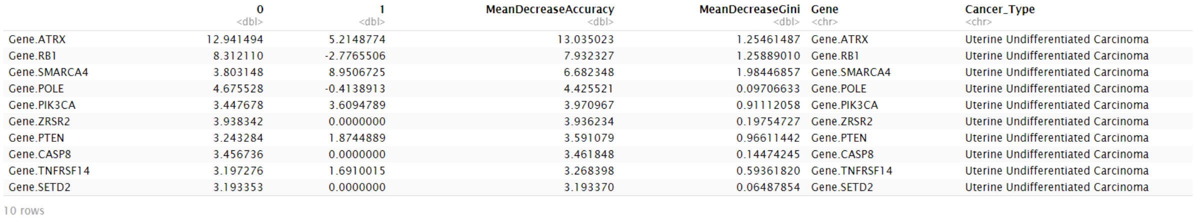


	0	1	MeanDecreaseAccuracy	MeanDecreaseGini	Gene	Cancer_Type
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
Gene.BRCA2	9.518084	-4.643154	9.289489	0.8105235	Gene.BRCA2	Uterine Clear Cell Carcinoma
Gene.BARD1	9.418105	-3.283004	9.070771	0.6764448	Gene.BARD1	Uterine Clear Cell Carcinoma
Gene.NFE2L2	8.927647	-3.477960	8.748802	0.7603256	Gene.NFE2L2	Uterine Clear Cell Carcinoma
Gene.PTEN	7.310291	4.245427	7.824251	1.4533344	Gene.PTEN	Uterine Clear Cell Carcinoma
Gene.TERT	5.759744	5.980546	7.048515	2.1663315	Gene.TERT	Uterine Clear Cell Carcinoma
Gene.SMAD4	6.569408	-2.816069	6.331669	0.6949095	Gene.SMAD4	Uterine Clear Cell Carcinoma
Gene.FLT3	6.659337	-1.413257	6.292385	1.0753098	Gene.FLT3	Uterine Clear Cell Carcinoma
Gene.TP53	6.458488	-3.079773	6.279482	0.9833788	Gene.TP53	Uterine Clear Cell Carcinoma
Gene.BMPR1A	6.820566	-2.211604	6.240428	0.6881939	Gene.BMPR1A	Uterine Clear Cell Carcinoma
Gene.KMT2C	5.603600	-2.346582	5.418499	0.5808817	Gene.KMT2C	Uterine Clear Cell Carcinoma

10 rows

Figure 19 - Top 10 genes predicting the Uterine clear cell carcinoma

This table displayed in Fig 19 the top ten genes involved in the classification of uterine clear cell carcinoma based on Mean Decrease Accuracy and Mean Decrease Gini. The gene BRCA2 had the greatest influence, followed by BARD1 and NFE2L2, demonstrating its importance in predicting this cancer type.

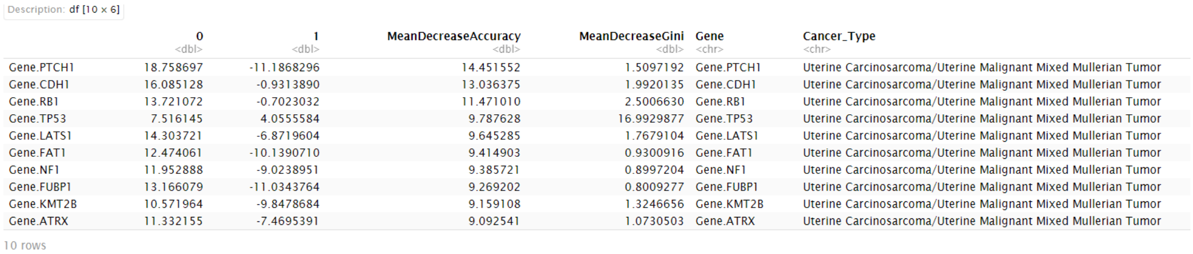


	0	1	MeanDecreaseAccuracy	MeanDecreaseGini	Gene	Cancer_Type
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
Gene.ATRX	12.941494	5.2148774	13.035023	1.25461487	Gene.ATRX	Uterine Undifferentiated Carcinoma
Gene.RB1	8.312110	-2.7765506	7.932327	1.25889010	Gene.RB1	Uterine Undifferentiated Carcinoma
Gene.SMARCA4	3.803148	8.9506725	6.682348	1.98446857	Gene.SMARCA4	Uterine Undifferentiated Carcinoma
Gene.POLE	4.675528	-0.4138913	4.425521	0.09706633	Gene.POLE	Uterine Undifferentiated Carcinoma
Gene.PIK3CA	3.447678	3.6094789	3.970967	0.91112058	Gene.PIK3CA	Uterine Undifferentiated Carcinoma
Gene.ZRSR2	3.938342	0.0000000	3.936234	0.19754727	Gene.ZRSR2	Uterine Undifferentiated Carcinoma
Gene.PTEN	3.243284	1.8744889	3.591079	0.96611442	Gene.PTEN	Uterine Undifferentiated Carcinoma
Gene.CASP8	3.456736	0.0000000	3.461848	0.14474245	Gene.CASP8	Uterine Undifferentiated Carcinoma
Gene.TNFRSF14	3.197276	1.6910015	3.268398	0.59361820	Gene.TNFRSF14	Uterine Undifferentiated Carcinoma
Gene.SETD2	3.193353	0.0000000	3.193370	0.06487854	Gene.SETD2	Uterine Undifferentiated Carcinoma

10 rows

Figure 20 - Top 10 genes predicting the Uterine Undifferentiated carcinoma

The table in Fig 20 shows the top ten genes that contribute to the classification of uterine undifferentiated carcinoma based on Mean Decrease Accuracy and Mean Decrease Gini. Gene ATRX has the greatest impact on classification, followed by RB1 and SMARCA4. These genes are critical in identifying this cancer type, with ATRX playing a prominent role. Other genes, such as POLE and PIK3CA, also participate, but to a lesser extent.



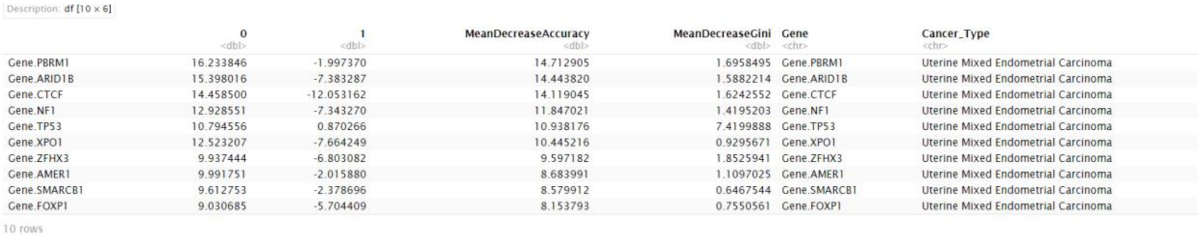
Description: df [10 x 6]

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini	Gene	Cancer_Type
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
Gene.PTCH1	18.758697	-11.1868296	14.451552	1.5097192	Gene.PTCH1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.CDH1	16.085128	-0.9313890	13.036375	1.9920135	Gene.CDH1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.RB1	13.721072	-0.7023032	11.471010	2.5006630	Gene.RB1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.TP53	7.516145	4.0555584	9.787628	16.9929877	Gene.TP53	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.LATS1	14.303721	-6.8719604	9.645285	1.7679104	Gene.LATS1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.FAT1	12.474061	-10.1390710	9.414903	0.9300916	Gene.FAT1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.NF1	11.952888	-9.0238951	9.385721	0.8997204	Gene.NF1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.FUBP1	13.166079	-11.0343764	9.269202	0.8009277	Gene.FUBP1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.KMT2B	10.571964	-9.8478684	9.159108	1.3246656	Gene.KMT2B	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor
Gene.ATRX	11.332155	-7.4695391	9.092541	1.0730503	Gene.ATRX	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor

10 rows

Figure 21 - Top 10 genes predicting the Uterine carcinosarcoma

Fig 21 depicts the top ten genes that help to classify Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumour using Mean Decrease Accuracy and Mean Decrease Gini. Genes PTCH1 and CDH1 had a major impact on categorisation accuracy, followed by RB1 and TP53. The significance of these genes reveals their essential role in identifying this cancer type, whereas other genes, such as LATS1 and FAT1, contribute to the decision-making process to a lesser level.



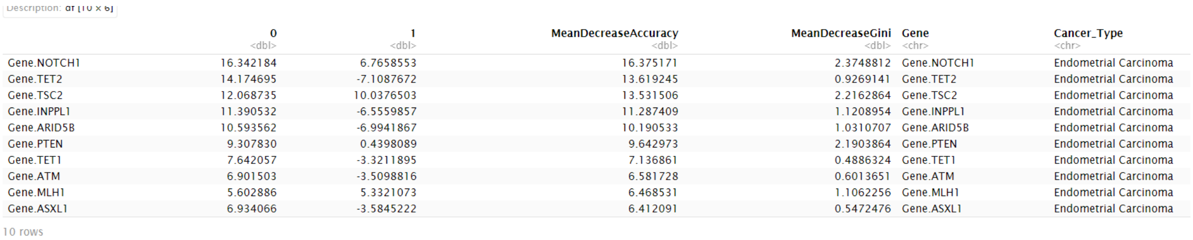
Description: df [10 x 6]

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini	Gene	Cancer_Type
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
Gene.PBRM1	16.233846	-1.997370	14.712905	1.6958495	Gene.PBRM1	Uterine Mixed Endometrial Carcinoma
Gene.ARID1B	15.398016	-7.383287	14.443820	1.5882214	Gene.ARID1B	Uterine Mixed Endometrial Carcinoma
Gene.CTCF	14.458500	-12.053162	14.119045	1.6242552	Gene.CTCF	Uterine Mixed Endometrial Carcinoma
Gene.NF1	12.928551	-7.343270	11.847021	1.4195203	Gene.NF1	Uterine Mixed Endometrial Carcinoma
Gene.TP53	10.794556	0.870266	10.938176	7.4199888	Gene.TP53	Uterine Mixed Endometrial Carcinoma
Gene.XPO1	12.523207	-7.664249	10.445216	0.9295671	Gene.XPO1	Uterine Mixed Endometrial Carcinoma
Gene.ZFHX3	9.937444	-6.803082	9.597182	1.8525941	Gene.ZFHX3	Uterine Mixed Endometrial Carcinoma
Gene.AMER1	9.991751	-2.015880	8.683991	1.1097025	Gene.AMER1	Uterine Mixed Endometrial Carcinoma
Gene.SMARCB1	9.612753	-2.378696	8.579912	0.6467544	Gene.SMARCB1	Uterine Mixed Endometrial Carcinoma
Gene.FOXP1	9.030685	-5.704409	8.153793	0.7550561	Gene.FOXP1	Uterine Mixed Endometrial Carcinoma

10 rows

Figure 22 - Top 10 genes predicting the Uterine Mixed Endometrial carcinoma

Fig 22 shows the top ten genes that contribute to the classification of uterine mixed endometrial carcinoma based on Mean Decrease Accuracy and Mean Decrease Gini. The genes PBRM1, ARID1B, and CTCF have a major impact on classification accuracy. These genes play an important role in distinguishing Uterine Mixed Endometrial Carcinoma from other cancers. Other genes, such as NF1 and TP53, also participate, but to a lower extent.



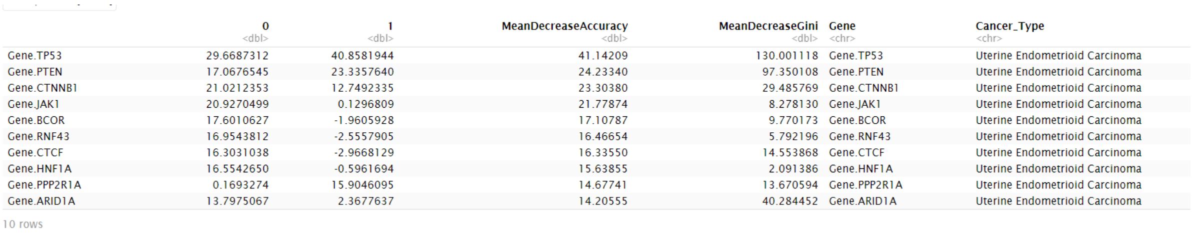
Description: df [10 x 6]

	0 <dbl>	1 <dbl>	MeanDecreaseAccuracy <dbl>	MeanDecreaseGini <dbl>	Gene <chr>	Cancer_Type <chr>
Gene.NOTCH1	16.342184	6.7658553	16.375171	2.3748812	Gene.NOTCH1	Endometrial Carcinoma
Gene.TET2	14.174695	-7.1087672	13.619245	0.9269141	Gene.TET2	Endometrial Carcinoma
Gene.TSC2	12.068735	10.0376503	13.531506	2.2162864	Gene.TSC2	Endometrial Carcinoma
Gene.INPPL1	11.390532	-6.5559857	11.287409	1.1208954	Gene.INPPL1	Endometrial Carcinoma
Gene.ARID5B	10.593562	-6.9941867	10.190533	1.0310707	Gene.ARID5B	Endometrial Carcinoma
Gene.PTEN	9.307830	0.4398089	9.642973	2.1903864	Gene.PTEN	Endometrial Carcinoma
Gene.TET1	7.642057	-3.3211895	7.136861	0.4886324	Gene.TET1	Endometrial Carcinoma
Gene.ATM	6.901503	-3.5098816	6.581728	0.6013651	Gene.ATM	Endometrial Carcinoma
Gene.MLH1	5.602886	5.3321073	6.468531	1.1062256	Gene.MLH1	Endometrial Carcinoma
Gene.ASXL1	6.934066	-3.5845222	6.412091	0.5472476	Gene.ASXL1	Endometrial Carcinoma

10 rows

Figure 23 - Top 10 genes predicting the Endometrial carcinoma

The top ten genes that contribute to the classification of endometrial carcinoma are shown in this Fig 23. NOTCH1, TET2, and TSC2 are the most important genes in terms of both Mean Decrease Accuracy and Mean Decrease Gini, suggesting their importance in accurately categorising this cancer type. Other genes, such as INPPL1 and PTEN, play important functions but have a slightly lower influence. Because of their mutation profiles, these genes aid in distinguishing Endometrial Carcinoma from other cancer forms.

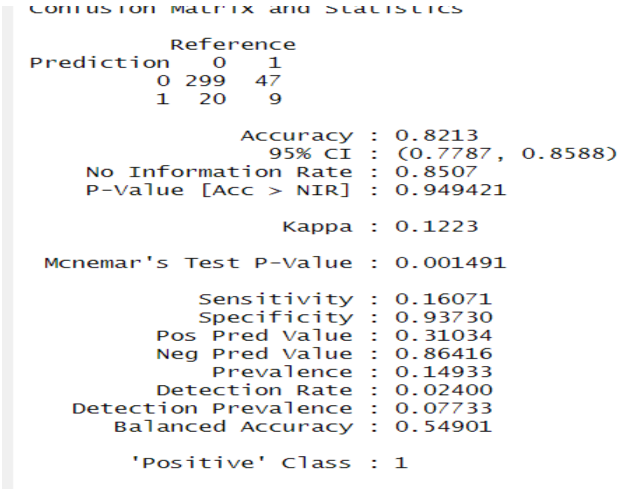


	0 <dbl>	1 <dbl>	MeanDecreaseAccuracy <dbl>	MeanDecreaseGini <dbl>	Gene <chr>	Cancer_Type <chr>
Gene.TP53	29.6687312	40.8581944	41.14209	130.001118	Gene.TP53	Uterine Endometrioid Carcinoma
Gene.PTEN	17.0676545	23.3357640	24.23340	97.350108	Gene.PTEN	Uterine Endometrioid Carcinoma
Gene.CTNNB1	21.0212353	12.7492335	23.30380	29.485769	Gene.CTNNB1	Uterine Endometrioid Carcinoma
Gene.JAK1	20.9270499	0.1296809	21.77874	8.278130	Gene.JAK1	Uterine Endometrioid Carcinoma
Gene.BCOR	17.6010627	-1.9605928	17.10787	9.770173	Gene.BCOR	Uterine Endometrioid Carcinoma
Gene.RNF43	16.9543812	-2.5557905	16.46654	5.792196	Gene.RNF43	Uterine Endometrioid Carcinoma
Gene.CTCF	16.3031038	-2.9668129	16.33550	14.553868	Gene.CTCF	Uterine Endometrioid Carcinoma
Gene.HNF1A	16.5542650	-0.5961694	15.63855	2.091386	Gene.HNF1A	Uterine Endometrioid Carcinoma
Gene.PPP2R1A	0.1693274	15.9046095	14.67741	13.670594	Gene.PPP2R1A	Uterine Endometrioid Carcinoma
Gene.ARID1A	13.7975067	2.3677637	14.20555	40.284452	Gene.ARID1A	Uterine Endometrioid Carcinoma

10 rows

Figure 24 - Top 10 genes predicting Uterine Endometrioid Carcinoma

Fig 24 shows the top ten genes that contribute to the classification of uterine endometrioid carcinoma. The most significant genes are TP53, PTEN, and CTNNB1, with TP53 having the most impact on model accuracy and decision-making, as evidenced by the highest Mean Decrease Accuracy (41.14) and Gini score (130.00). Other genes, including as JAK1, BOCR, and RNF43, play a role, albeit to a lesser extent. Based on their mutation profiles, these top genes help to identify Uterine Endometrioid Carcinoma from other cancer types.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 299  47
         1  20   9

               Accuracy : 0.8213
                 95% CI : (0.7787, 0.8588)
    No Information Rate : 0.8507
    P-Value [Acc > NIR] : 0.949421

                  Kappa : 0.1223

 Mcnemar's Test P-Value : 0.001491

            Sensitivity : 0.16071
            Specificity : 0.93730
         Pos Pred Value : 0.31034
         Neg Pred Value : 0.86416
             Prevalence : 0.14933
         Detection Rate : 0.02400
   Detection Prevalence : 0.07733
      Balanced Accuracy : 0.54901

       'Positive' Class : 1
```

Figure 25 - Confusion matrix and associated statistics for predicting Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma

Key performance metric:

Accuracy 82.13%: The fraction of successfully predicted instances (True Positives + True Negatives) among all predictions. This is a relatively high level of accuracy (Fig 25)
95% CI: The confidence interval for the accuracy is 77.87% to 85.88%, indicating that the true accuracy is within this range with 95% confidence.
No Information Rate: 85.07%, which is the accuracy achieved by always guessing the most frequent class (class 0). The model performs somewhat worse than the No Information Rate, as evidenced by the P-Value of 0.9494, implying that its performance is not significantly better than random chance based on the majority class.

]Kappa (0.1223): A measure of the agreement between actual and expected classifications after correcting for chance. A result of 0.1223 suggests just minor agreement above random chance.

McNemar's Test P-Value (0.001491): There is a significant difference between false positives and false negatives, indicating that the model's errors do not occur randomly.

Sensitivity (16.07%): The model identifies just 16.07% of actual class 1 instances, demonstrating a difficulty predicting class 1.

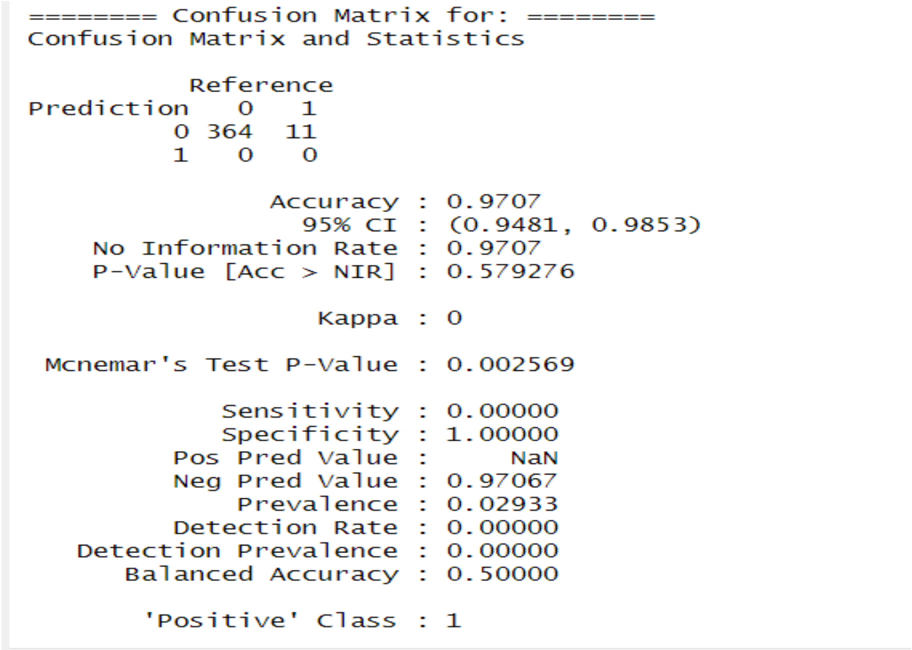
Specificity (93.73%): The model correctly predicts 93.73% of class 0 instances, indicating that it works effectively for the majority of cases.

Positive Predictive Value (31.03%): Out of the projected class 1 cases, only 31.03% are correct, suggesting low precision.

Negative Predictive Value (86.41%): Of the projected class 0 cases, 86.41% are correct, indicating a high prediction rate for class 0.

Prevalence (14.93%): Class 1 makes up 14.93% of the total dataset. Balanced Accuracy

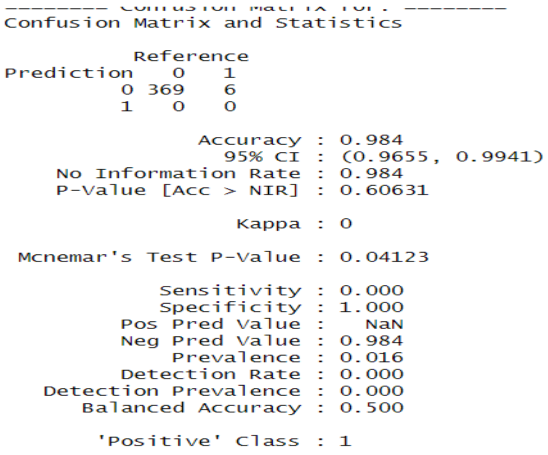
(54.91%): The model’s average ability to correctly identify both classes is slightly better than random (50%).



```
======== Confusion Matrix for: ========
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 364  11
         1   0   0

               Accuracy : 0.9707
                 95% CI : (0.9481, 0.9853)
    No Information Rate : 0.9707
    P-Value [Acc > NIR] : 0.579276

                  Kappa : 0

 Mcnemar's Test P-Value : 0.002569

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.97067
             Prevalence : 0.02933
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 1
```

Figure 26 - confusion matrix and associated statistics for predicting Uterine Clear Cell Carcinoma

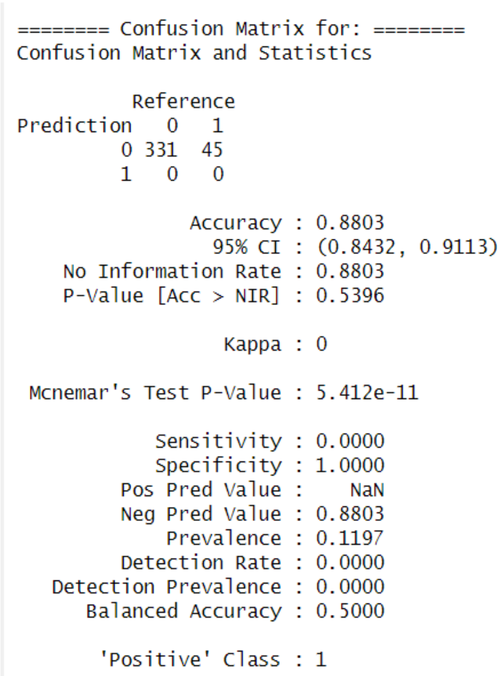
This confusion matrix of Uterine Clear Cell Carcinoma in Fig 26 has a high accuracy (97.07%) but a low sensitivity (0.0%), indicating that the model does not recognise any positive cases. All real positive cases were projected as negatives, resulting in a Kappa of zero, showing no agreement above chance. Specificity is perfect (1.00), indicating that the model accurately detects all true negatives. The McNemar's test P-value (0.002569) reveals a substantial discrepancy between false negatives and false positives, indicating a problem with the model's classification performance, particularly in the positive class.



```
======== Confusion Matrix for: ========
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 369   6
         1   0   0

               Accuracy : 0.984
                 95% CI : (0.9655, 0.9941)
    No Information Rate : 0.984
    P-Value [Acc > NIR] : 0.60631

                  Kappa : 0

 Mcnemar's Test P-Value : 0.04123

            Sensitivity : 0.000
            Specificity : 1.000
         Pos Pred Value :   NaN
         Neg Pred Value : 0.984
             Prevalence : 0.016
         Detection Rate : 0.000
   Detection Prevalence : 0.000
      Balanced Accuracy : 0.500

       'Positive' Class : 1
```

Figure 27 - Confusion matrix and associated statistics for predicting undifferentiated carcinoma

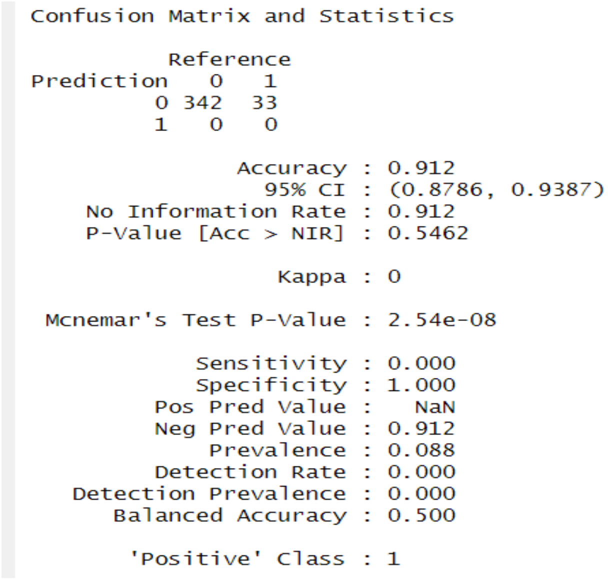
The confusion matrix (Fig 27) for uterine undifferentiated carcinoma has a high accuracy (98.4%), but a sensitivity of 0%, indicating that the model does not predict any positive cases. Specificity is 100 percent, which means that all negative cases were correctly predicted. The Kappa value is zero, indicating no agreement beyond chance. The McNemar's test P-value (0.04123) indicates a substantial imbalance in the misclassification of positives and negatives. The positive predictive value (precision) is not applicable (NaN) due to the lack of real positives, and the balanced accuracy is 50%, suggesting that the model is only successful in negative circumstances.



```
======== Confusion Matrix for: ========
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 331  45
         1   0   0

               Accuracy : 0.8803
                 95% CI : (0.8432, 0.9113)
    No Information Rate : 0.8803
    P-Value [Acc > NIR] : 0.5396

                  Kappa : 0

 Mcnemar's Test P-Value : 5.412e-11

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.8803
             Prevalence : 0.1197
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

Figure 28 - Confusion matrix and associated statistics for predicting Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian carcinoma

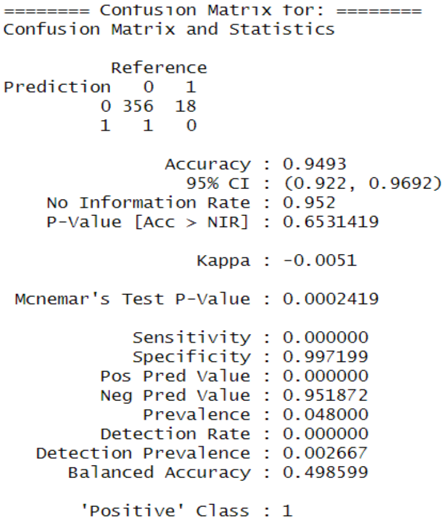
The model's accuracy for Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian carcinoma Fig 28 is 88.03%, which means that 88.03% of predictions are right. However, the sensitivity is 0%, indicating that the model does not discover any positive cases (true positives). Specificity is 100%, which means that all true negatives are anticipated correctly. The Kappa value is zero, indicating no agreement beyond chance. The McNemar's test P-value is extremely low (5.412e-11), indicating a significant difference between false positive and false negative errors. The model cannot produce any valid predictions for the positive class (NaN precision), and the negative predictive value is 88.03%.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 342  33
         1   0   0

               Accuracy : 0.912
                 95% CI : (0.8786, 0.9387)
    No Information Rate : 0.912
    P-Value [Acc > NIR] : 0.5462

                  Kappa : 0

 Mcnemar's Test P-Value : 2.54e-08

            Sensitivity : 0.000
            Specificity : 1.000
         Pos Pred Value :   NaN
         Neg Pred Value : 0.912
             Prevalence : 0.088
         Detection Rate : 0.000
   Detection Prevalence : 0.000
      Balanced Accuracy : 0.500

       'Positive' Class : 1
```

Figure 29 - confusion matrix and associated statistics for predicting uterine mixed endometrial carcinoma.

The model in Fig 29 has a 91.2% accuracy rate for uterine mixed endometrial carcinoma, which means that 91.2% of predictions are right. The sensitivity is 0%, which means that the model does not detect any true positives. Specificity is 100%, indicating that all true negatives are correctly identified. The Kappa value of 0 indicates that there is no agreement beyond random chance. The McNemar's test P-value of 2.54e-08 indicates a substantial difference between false positive and false negative errors. Positive Predictive Value (PPV) cannot be obtained (NaN), but Negative Predictive Value (NPV) is 91.2%, indicating that the model accurately predicts negative cases for 91.2% of the time. The prevalence of positive cases is low, at 8.8%, and the detection rate is zero, indicating that no positive cases were found.

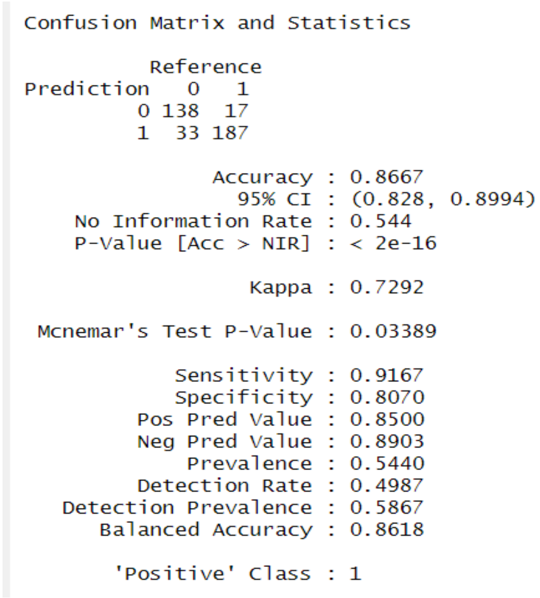


```
======== Confusion Matrix for: ========
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 356  18
         1   1   0

               Accuracy : 0.9493
                 95% CI : (0.922, 0.9692)
    No Information Rate : 0.952
    P-Value [Acc > NIR] : 0.6531419

                  Kappa : -0.0051

 Mcnemar's Test P-Value : 0.0002419

            Sensitivity : 0.000000
            Specificity : 0.997199
         Pos Pred Value : 0.000000
         Neg Pred Value : 0.951872
             Prevalence : 0.048000
         Detection Rate : 0.000000
   Detection Prevalence : 0.002667
      Balanced Accuracy : 0.498599

       'Positive' Class : 1
```

Figure 30 - confusion matrix and associated statistics for predicting Endometrial Carcinoma.

The model's accuracy for endometrial carcinoma (Fig 30) is 94.93%, indicating that it accurately categorised about 95% of the cases. The sensitivity (true positive rate) is 0%, suggesting that the model detected no true positives in this class. The specificity (true negative rate) is 99.72 percent, indicating that nearly all genuine negatives were accurately identified. The Kappa score is -0.0051, suggesting that no agreement exists beyond chance.

The McNemar's test P-value of 0.0002419 indicates an important difference between false positives and false negatives. The Positive Predictive Value (PPV) is 0%, which means there are no valid predictions for positive situations. The Negative Predictive Value (NPV) is 95.19%, showing that the model correctly predicts negatives.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 138  17
         1  33 187
                                          
               Accuracy : 0.8667          
                 95% CI : (0.828, 0.8994)
    No Information Rate : 0.544           
    P-Value [Acc > NIR] : < 2e-16         
                                          
                  Kappa : 0.7292          
                                          
 Mcnemar's Test P-Value : 0.03389         
                                          
            Sensitivity : 0.9167          
            Specificity : 0.8070          
         Pos Pred Value : 0.8500          
         Neg Pred Value : 0.8903          
             Prevalence : 0.5440          
         Detection Rate : 0.4987          
   Detection Prevalence : 0.5867          
      Balanced Accuracy : 0.8618          
                                          
       'Positive' Class : 1               
```

Figure 31 - confusion matrix and associated statistics for predicting uterine endometrioid carcinoma.

The model's accuracy for uterine endometrioid carcinoma Fig 31 is 86.67%, indicating that it accurately classified the majority of cases. The sensitivity is 91.67%, which means the model accurately detected 91.67% of positive cases. Specificity is lower (80.70%), indicating that some negative cases were misclassified. The PPV (positive predictive value) is 85%, while the NPV (negative predictive value) is 89.03%. The Kappa score is 0.7292, indicating a significant agreement between predictions and actual classifications.

4.8 Visualizing the Top 10 Genes contributes for the prediction of cancer types

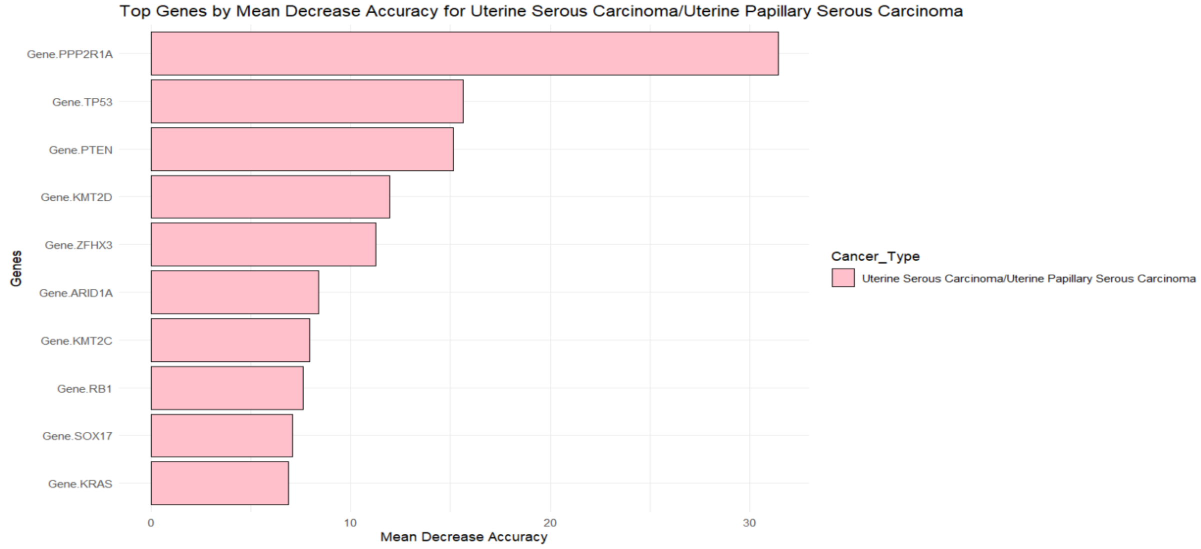


Figure 32 - Top 10 Genes by Mean Decrease Accuracy for Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma

This bar chart in Fig 32 depicts the top genes that contribute to the classification of Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma in a Random Forest model based on Mean Decrease Accuracy. PPP2R1A, TP53, and PTEN are the three most important genes, with PPP2R1A having the greatest impact on model accuracy.

Similarly for all the cancer types the plots were shown below:

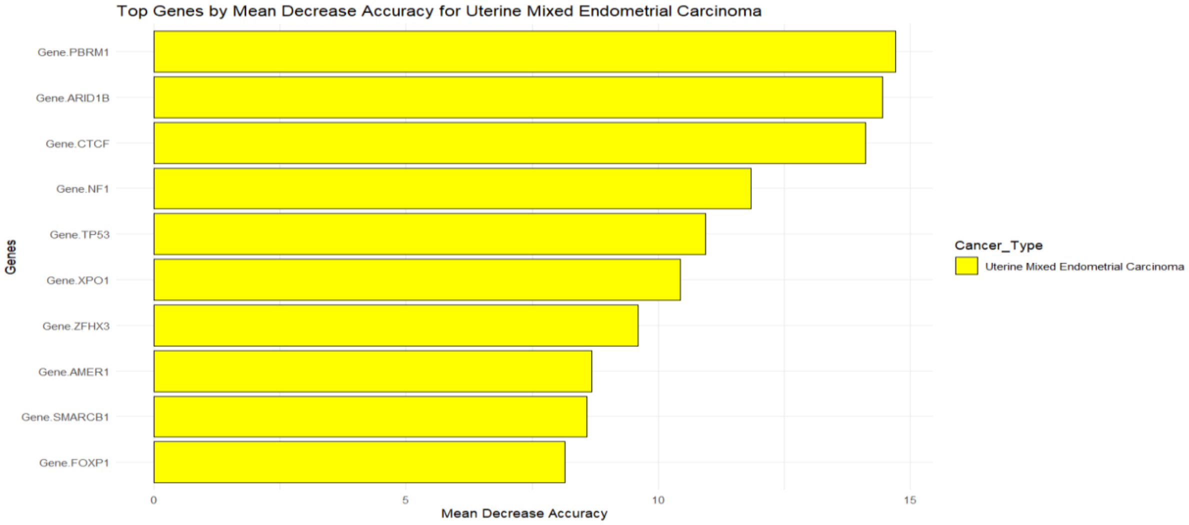


Figure 33 - Top 10 Genes by Mean Decrease Accuracy for Uterine Mixed Endometrial Carcinoma

Fig 33 depicts the top genes contributing to the classification of Uterine Mixed Endometrial Carcinoma based on Mean Decrease Accuracy. Genes PBRM1 and ARID1B had the greatest impact on prediction accuracy, followed by CTCF and NF1, demonstrating their importance in differentiating this cancer type.

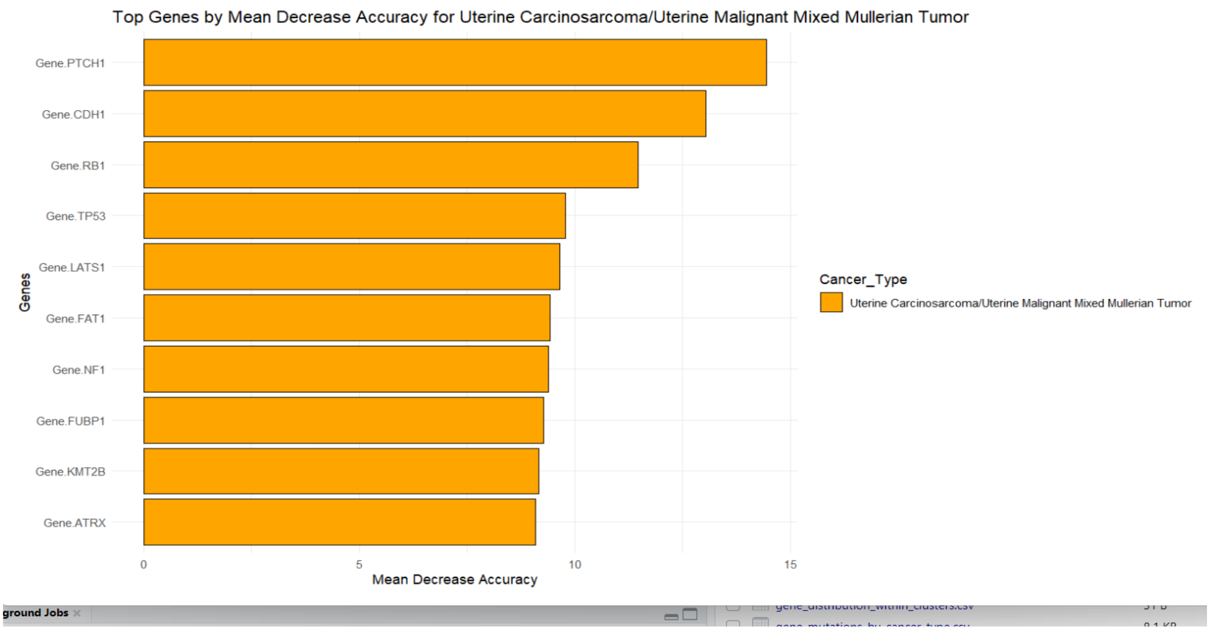


Figure 34 - Top 10 Genes by Mean Decrease Accuracy for Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumour

This chart 34 shows the top genes influencing the classification of Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumour, with Gene PTCH1 having the highest impact on model accuracy. Genes CDH1 and RB1 also play significant roles in enhancing the predictive performance for this cancer type.

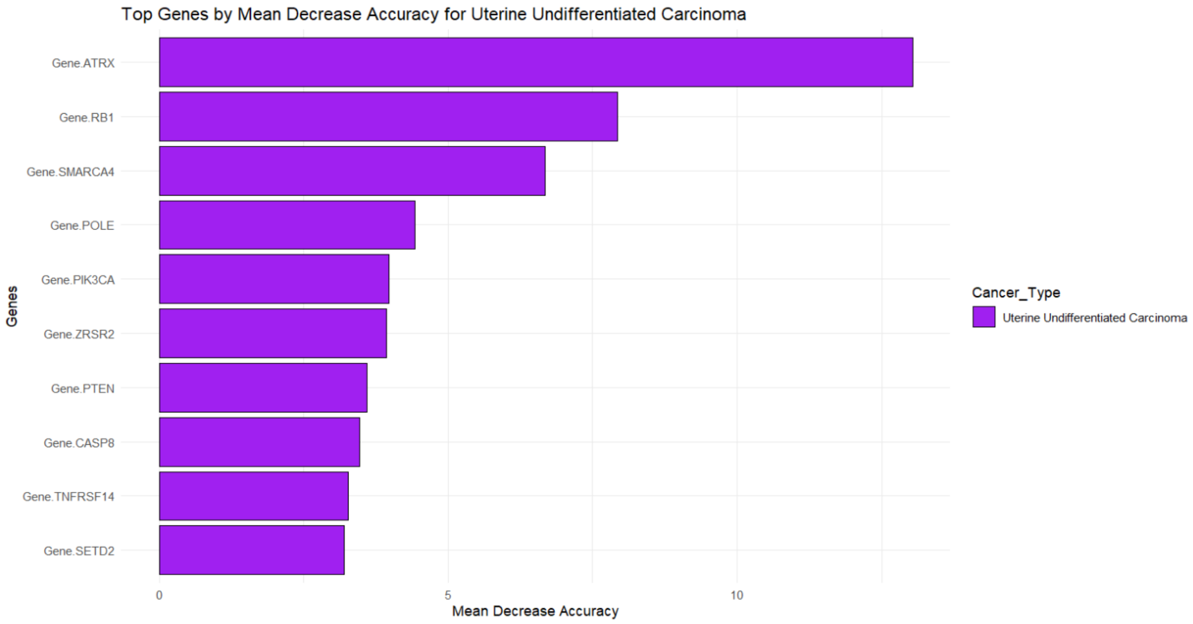


Figure 35 - Top 10 Genes by Mean Decrease Accuracy for Uterine Undifferentiated Carcinoma

This bar chart Fig 35 depicts the top genes that contribute to the classification of uterine undifferentiated carcinoma, with ATRX and RB1 being the most significant. These genes have the greatest Mean Decrease in Accuracy, demonstrating their critical importance in accurately predicting this cancer type.

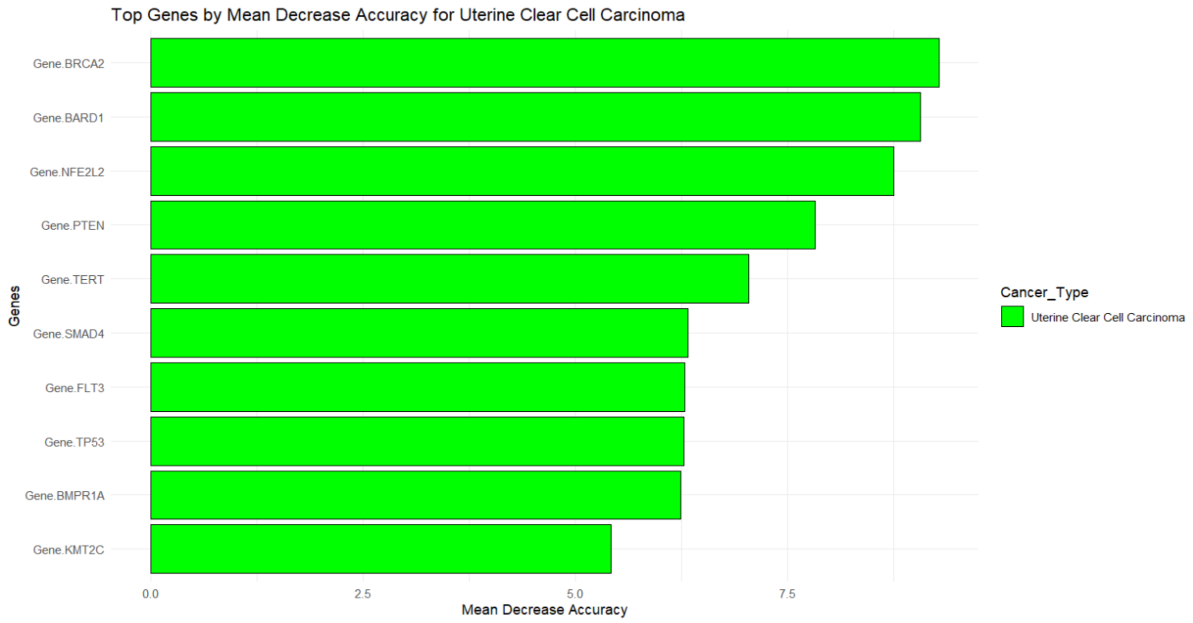


Figure 36 - Top 10 Genes by Mean Decrease Accuracy for Uterine Clear Cell Carcinoma

The Fig 36 shows the top genes that influence the classification of **Uterine Clear Cell Carcinoma based on Mean Decrease Accuracy. BRCA2, BARD1, and NFE2L2 are the most important genes, with BRCA2 having the greatest influence on the model's accuracy in predicting this cancer type.

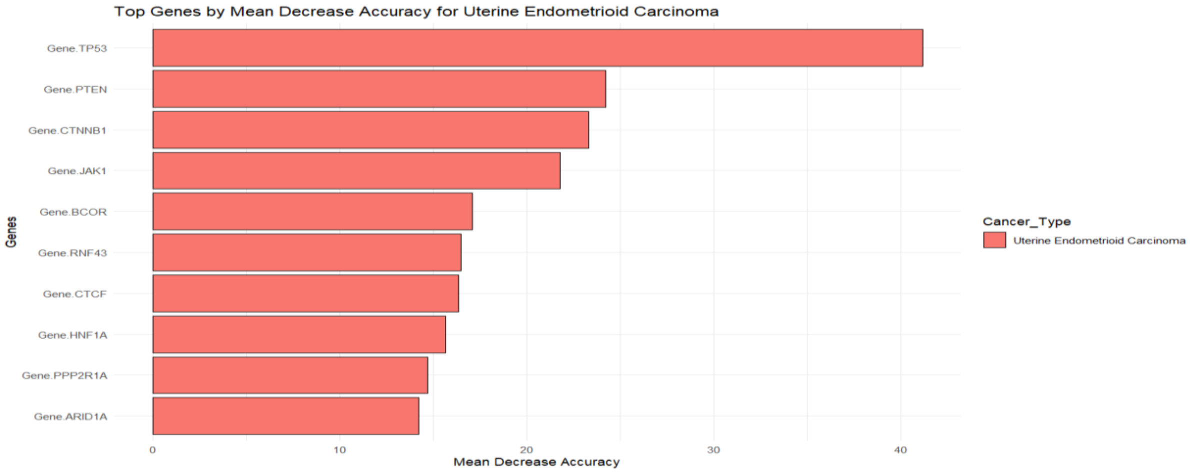


Figure 37 - Top 10 Genes by Mean Decrease Accuracy for Uterine Endometrioid Carcinoma

The graphic in Fig 37 shows the top genes that help classify Uterine Endometrioid Carcinoma based on Mean Decrease Accuracy. TP53 is the most important gene, followed by PTEN and CTNNB1, demonstrating its significance in proper classification of this cancer type.

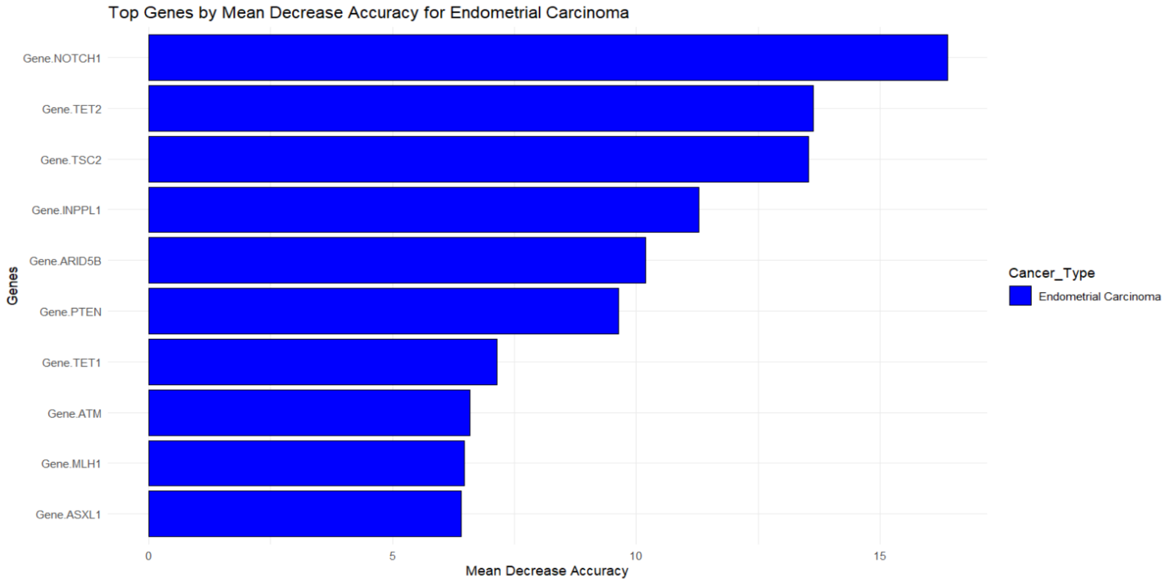


Figure 38 - Top 10 Genes by Mean Decrease Accuracy for Endometrial Carcinoma

The figure 38 depicts the top genes that help to classify Endometrial Carcinoma based on Mean Decrease Accuracy. NOTCH1 is the most significant gene, followed by TET2 and TSC2, which are critical for classifying this cancer types.

CHAPTER 5 DISCUSSION

The primary goal of this work was to investigate how clustering algorithms can be used to identify groups of patients with similar combinations of gene mutations while considering their ethnicity, histology, and cancer type, as well as to classify endometrial cancer types based on gene mutation. This investigation was guided by two research questions:

Can clustering techniques identify groups of patients with similar combinations of gene mutations, while also accounting for their ethnicity, histology, and cancer type in endometrial cancer?
classify the patient's cancer type based on the gene mutation or mutations they have gene alterations they carry?

This study made use of a dataset from the cBioPortal for Cancer Genomics, which includes patient profiles with information such as gene mutations, ethnicity, histology, and cancer type. The investigation used clustering and classification approaches to gather insights into patterns within the dataset and predict cancer types based on genetic profiles.

The results from the clustering analysis showed that gene mutation profiles could indeed be used to identify subgroups of patients who share similar genetic and clinical characteristics. By applying the K-means algorithm and using the elbow method to determine the optimal number of clusters, the patient data were divided into several meaningful clusters. The clusters revealed significant relationships between the cancer types and the genetic mutations. Specifically, patients with similar histology types or cancer types exhibited closely related mutation patterns, consistent with findings in the literature.

Furthermore, the clustering study revealed the potential impact of ethnicity on cluster structure. For example, certain clusters had a higher representation of specific ethnic groups, implying that ethnicity may play a role in mutation patterns, which is consistent with previous research that shows genetic and ethnic variations can influence cancer susceptibility and outcomes. This is especially important in the context of personalised medicine, because identifying ethnic variations in gene mutations can help guide focused treatments.

5.1 Clustering by Cancer Type and Mutation Count

The exploratory research and subsequent clustering of patients based on cancer type, ethnicity, histology, and mutation count revealed patterns that would not have been obvious from a purely clinical standpoint. Clustering revealed that specific cancer subtypes, such as uterine serous carcinoma and uterine mixed endometrial carcinoma, have distinct mutation counts and genetic profiles. The contrast between these clusters offers practical insights into how genetic variation emerges across various histological types. This discovery corresponds with the literature on cancer genomics and personalised treatment.

Interestingly, the cluster analysis revealed outliers, such as patients with a large number of gene mutations across multiple cancer types. These outliers indicate a level of genetic complexity in certain people that traditional diagnostic criteria may fail to detect. This finding confirms the growing emphasis in cancer research on mutation burden as a predictive factor.

5.2 Limitations of Clustering for Ethnicity

One significant constraint of the clustering analysis was the under-representation of some ethnic groups, particularly Hispanics, in the dataset. The unequal distribution of ethnic groups most likely influenced cluster formation, limiting the generalisability of the findings on ethnicity and mutation shifts. While others were identified such as a greater number of specific cancer types among non-Hispanic patients, the small sample numbers for certain ethnic groups made it impossible to draw robust conclusions about ethnicity-based differences in mutation patterns. Future research should include a more diverse and balanced sample to properly examine the impact of ethnicity on genetic mutations in endometrial cancer.

5.3 Classifying Cancer Types Based on Gene Mutations

The classification assignment, which used the Random Forest model, revealed important information about the relationship between gene mutations and cancer type classification. The model has an overall accuracy of 64.34%, with significant differences in performance among cancer types. For example, the model was quite effective at predicting uterine endometrioid carcinoma, due to the strong contributions of genes such as TP53, PTEN, and CTNNB1. These genes, especially TP53, have been identified in studies as important causes of endometrial cancer.

However, the classification algorithm struggled to distinguish between particular cancer forms, such as endometrial carcinoma and uterine serous carcinoma, most likely due to overlapping gene alterations. Many uterine malignancies contain mutations in genes such as TP53 and PIK3CA, making it challenging for the model to distinguish subtypes based solely on these genes. This difficulty emphasises the complexity of cancer classification using genetic data and the need for more refined models that can account for the small differences across cancer subtypes.

5.4 Importance of Gene Selection for Classification

One of the study's primary results is the need of identifying the most significant genes for accurate cancer classification. The Random Forest model was able to attain a higher level of accuracy for specific types of cancer by focussing on the top ten genes contributing to each. For example, in uterine serous carcinoma, genes such as PPP2R1A and TP53 were found as the most critical for categorisation, which is consistent with their recognised roles in the cause of aggressive cancer types.

The method of picking a small number of highly informative genes was found to be beneficial in decreasing noise and enhancing model performance. This method is especially crucial in the case of high-dimensional genomic data, because include too many irrelevant or weakly informative genes may reduce a model's ability to generate accurate predictions. By identifying and visualising the most relevant genes for each cancer type, this study contributes to the growing body of research on gene-based cancer classification.

5.5 Visualizing Gene Importance Across Cancer Types

The visualisations of the top genes involved in the classification of various cancer types gave a simple approach to evaluate the Random Forest model's findings. These graphs clearly showed the varying importance of specific genes in predicting each cancer types. For example, the gene PPP2R1A was found to have a significant impact on the classification of uterine serous carcinoma, whereas genes such as TP53 and PTEN were more essential for uterine endometrioid carcinoma. The ability to visualise these gene contributions not only helps to explain the model's decisions, but it also provides insights into the molecular causes of each cancer types, which may guide future research into specific treatments. **The interesting observation not only TP53, PTEN but also different gene contributed in the top position for predicting the each cancer types.**

5.6 Limitations and Challenges in Cancer Type Classification

While the Random Forest model performed well in predicting cancer types based on gene mutations it did have certain limitations. One challenge was the model's low sensitivity for certain cancer forms, particularly endometrial carcinoma, which made it difficult to recognise genuine positives. This low sensitivity could be attributed to the overlap in gene alterations between cancer types, as well as the possibility of noise in genomic data. Furthermore, the model's performance was limited by the dataset's imbalance, with cancer types under-represented. This imbalance most likely contributed to the model's bias towards more common cancer types, resulting in misclassifications of rarer cancer types.

Furthermore, using gene mutations as the sole predictor of cancer type may have limits, because cancer categorisation is influenced by a variety of factors other than genetic alterations, such as epigenetic modifications and environmental factors. Future models should consider incorporating more forms of data, such as gene expression profiles and epigenetic markers, to increase classification accuracy and provide a more complete picture of cancer subgroups.

5.7 Implications for Clinical Practice and Personalized Medicine

This study's findings have various implications for clinical practice and the development of personalised medicine approaches to endometrial cancer. The ability to cluster individuals based on genetic profiles and classify cancer types with good precision brings up potential for customised treatment options. Identifying the major genes associated with distinct cancer types allows clinicians to develop a better understanding of the disease's biological processes and potentially create medicines that target specific genetic mutations. For example, patients with Uterine Serous Carcinoma who have PPP2R1A mutations may benefit from medicines that target this pathway. Patient clustering based on ethnicity, histology, and gene mutations can identify subgroups receptive to specific medications, enhancing therapeutic decision-making and improving patient outcomes.

.

5.8 Future Research Directions

Future research should explore the application of deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in cancer categorization. These models can identify complex patterns from genetic data, potentially increasing cancer classification accuracy. The study suggests expanding the dataset to include diverse patient populations, integrating genomic data like gene expression and methylation profiles, and combining clinical data with genetic data to create more robust models for predicting cancer outcomes and tailoring therapies. This approach aligns with the growing emphasis on precision medicine in cancer care.

CHAPTER 6 CONCLUSON

This study sought to investigate two key questions about endometrial cancer: (1) whether clustering techniques could effectively identify groups of patients with similar combinations of gene mutations while accounting for ethnicity, histology, and cancer type, and (2) whether machine learning models could accurately classify a patient's cancer type based on their specific gene mutations. Using both clustering and classification approaches, the study revealed substantial insights into the genetic and clinical variables that distinguish different subtypes of endometrial cancer. This section will summarise the key findings, explore their ramifications, and suggest prospective areas for future research.

Clustering techniques, notably k-means clustering, were used successfully to identify groups of patients with similar characteristics based on gene mutations, histology, ethnicity, and cancer type. By incorporate these variables in the clustering algorithm, meaningful patterns developed, dividing individuals into clusters with distinct clinical and genetic characteristics. The elbow technique was used to estimate the most effective number of clusters, ensuring that patient clustering maximised both within-group similarity and between-group differences.

The findings revealed that specific gene mutations were more common in various clusters, implying possible links between genetic changes and clinical characteristics including ethnicity and cancer histology. For example, uterine serous carcinoma and uterine carcinosarcoma, two aggressive forms of endometrial cancer, commonly emerged in clusters with significant mutation loads in genes such as TP53 and PPP2R1A. Patients with less aggressive cancer forms, such as uterine endometrioid carcinoma, tended to cluster together, indicating lower mutation rates in important oncogenes. These findings highlight the importance of clustering as a tool for understanding the heterogeneity of endometrial cancer, implying that certain genetic profiles may predispose individuals to specific cancer subtypes.

Furthermore, the addition of ethnicity in the clustering study shed light on potential demographic factors that promote cancer development. Although most patients in the sample were not Hispanic, the distribution of cancer types among ethnic groups suggested that some subtypes, such as uterine carcinosarcoma, may be more common in Hispanic communities. This conclusion emphasises the importance of demographic considerations in cancer research and treatment planning.

The second study question asked if gene mutation profiles could be utilised to determine a patient's cancer types. Using Random Forest, a machine learning method well-suited for high-dimensional genetic data, the study successfully created classification models that identified cancer type based on gene mutation. The results showed that the Random Forest model worked reasonably well, with an overall accuracy of approximately 64.34% and higher accuracy for several specific cancer subtypes.

Importantly, the study revealed the most important genes in determining the classification of each cancer subtype. For example, TP53, a gene renowned for its role in cancer suppression, was a significant predictor of uterine serous carcinoma, but genes such as PTEN and CTNNB1 were required to predict uterine endometrial carcinoma. These findings are consistent with previous research, which has long recognised the relevance of specific gene mutations in causing various endometrial cancer subtypes.

The algorithm struggled to accurately forecast less prevalent cancer subtypes like uterine clear cell carcinoma and uterine undifferentiated carcinoma due to overlapping gene mutation patterns and a dataset imbalance, which likely confused the model and biased its predictions.

The study demonstrates that clustering and machine learning techniques can aid in understanding genetic and clinical factors contributing to endometrial cancer. It identifies patient groups and genes for cancer categorization, paving the way for personalized treatment regimens. Although challenges persist, this represents a significant advancement in precision oncology.

CHAPTER 7 REFERENCE

Ahmad, A. and Dey, L. (2007) ‘A k-mean clustering algorithm for mixed numeric and categorical data’, *Data & Knowledge Engineering*, 63(2), pp. 503–527. Available at: https://doi.org/10.1016/j.datak.2007.03.016.

Alessandrino, F. *et al.* (2023) ‘Uterine serous carcinoma: assessing association between genomics and patterns of metastasis’, *Frontiers in Oncology*, 13. Available at: https://doi.org/10.3389/fonc.2023.1066427.

Baker, S.G. and Kramer, B.S. (2006) ‘Identifying genes that contribute most to good classification in microarrays’, *BMC Bioinformatics*, 7(1). Available at: https://doi.org/10.1186/1471-2105-7-407.

Bianco, B. *et al.* (2020) ‘Endometrial cancer: a genetic point of view’, *Translational Cancer Research*, 9(12). Available at: https://doi.org/10.21037/tcr-20-2334.

cBioPortal for Cancer Genomics (2024) Cbioportal.org. Available at: https://www.cbioportal.org/study/summary?id=ucec_ancestry_cds_msk_2023

Chang, C.-H. and Ding, Z.-K. (2005) ‘Categorical data visualization and clustering using subjective factors’, *Data & Knowledge Engineering*, 53(3), pp. 243–262. Available at: https://doi.org/10.1016/j.datak.2004.09.001.

Cui, M. (2020) ‘Introduction to the K-Means Clustering Algorithm Based on the Elbow Method’. Available at: https://doi.org/10.23977/accaf.2020.010102.

Do, T., Graefe, G. and Naughton, J. (2022) ‘Efficient Sorting, Duplicate Removal, Grouping, and Aggregation’, *ACM Transactions on Database Systems*, 47(4), pp. 1–35. Available at: https://doi.org/10.1145/3568027.

Franzese, M. and Iuliano, A. (2019) ‘Correlation analysis’, *Encyclopedia of Bioinformatics and Computational Biology*, 1, pp. 706–721. Available at: https://doi.org/10.1016/b978-0-12-809633-8.20358-0.

Granger, IN: ISDSA Press. ISBN: 978-1-946728-01-2.

Higgins, T. *et al.* (2023) ‘Students’ approaches to exploring relationships between categorical variables’, *Teaching Statistics* [Preprint]. Available at: https://doi.org/10.1111/test.12331.

Huang, A.B. *et al.* (2020) ‘Impact of quality of care on racial disparities in survival for endometrial cancer’, *American Journal of Obstetrics and Gynecology*, 223(3), pp. 396.e1–396.e13. Available at: https://doi.org/10.1016/j.ajog.2020.02.021.

Jaeger, A. and Banks, D. (2022) ‘Cluster analysis: A modern statistical review’, *WIREs Computational Statistics* [Preprint]. Available at: https://doi.org/10.1002/wics.1597.

Jaeger, A. and Banks, D. (2022) ‘Cluster analysis: A modern statistical review’, *WIREs Computational Statistics* [Preprint]. Available at: https://doi.org/10.1002/wics.1597.

Kanungo, T. *et al.* (2002) ‘An efficient k-means clustering algorithm: analysis and implementation’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), pp. 881–892. Available at: https://doi.org/10.1109/tpami.2002.1017616.

Kwak, S.K. and Kim, J.H. (2017) ‘Statistical data preparation: management of missing values and outliers’, *Korean Journal of Anesthesiology*, 70(4), pp. 407–411. Available at: https://doi.org/10.4097/kjae.2017.70.4.407.

Laas, E. *et al.* (2017) ‘Unsupervised Clustering of Immunohistochemical Markers to Define High-Risk Endometrial Cancer’, *Pathology & Oncology Research*, 25(2), pp. 461–469. Available at: https://doi.org/10.1007/s12253-017-0335-y.

Larson, M.G. (2006) ‘Descriptive Statistics and Graphical Displays’, *Circulation*, 114(1), pp. 76–81. Available at: https://doi.org/10.1161/circulationaha.105.584474.

Lax, S.F. (2004) ‘Molecular genetic pathways in various types of endometrial carcinoma: from a phenotypical to a molecular-based classification’, *Virchows Archiv*, 444(3), pp. 213–223. Available at: https://doi.org/10.1007/s00428-003-0947-3.

Li, Y. *et al.* (2022) ‘One-stop molecular classification of endometrial carcinoma using comprehensive next-generation sequencing’, *International Journal of Cancer*, 151(11), pp. 1969–1977. Available at: https://doi.org/10.1002/ijc.34241.

McConechy, M.K. *et al.* (2012) ‘Use of mutation profiles to refine the classification of endometrial carcinomas’, *The Journal of Pathology*, 228(1), pp. 20–30. Available at: https://doi.org/10.1002/path.4056.

Monteiro (2009) ‘Identification of a 0.4 Kb deletion region in 10q26 associated with endometrial carcinoma’, *Oncology Reports*, 23(2). Available at: https://doi.org/10.3892/or00000664.

Mukerji, B. *et al.* (2018) ‘Racial disparities in young women with endometrial cancer’, *Gynecologic Oncology*, 148(3), pp. 527–534. Available at: https://doi.org/10.1016/j.ygyno.2017.12.032.

Murali, R. *et al.* (2018) ‘Evolving Roles of Histologic Evaluation and Molecular/Genomic Profiling in the Management of Women with Endometrial Cancer’, *Journal of the National Comprehensive Cancer Network: JNCCN*, 16(2), pp. 201–209. Available at: https://doi.org/10.6004/jnccn.2017.7066.

Okuda, T. *et al.* (2010) ‘Genetics of Endometrial Cancers’, *Obstetrics and Gynecology International*, 2010, p. 984013. Available at: https://doi.org/10.1155/2010/984013.

Rios-Doria, E. *et al.* (2023) ‘Integration of clinical sequencing and immunohistochemistry for the molecular classification of endometrial carcinoma’, *Gynecologic Oncology*, 174, pp. 262–272. Available at: https://doi.org/10.1016/j.ygyno.2023.05.059.

Salanti, G., Ades, A.E. and Ioannidis, J.P.A. (2011) ‘Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial’, *Journal of Clinical Epidemiology*, 64(2), pp. 163–171. Available at: https://doi.org/10.1016/j.jclinepi.2010.03.016.

Unwin, A. (2018) *Graphical Data Analysis with R*. Chapman and Hall/CRC. Available at: https://doi.org/10.1201/9781315370088.

Wickham, H. (2011) ‘ggplot2’, *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), pp. 180–185. Available at: https://doi.org/10.1002/wics.147.

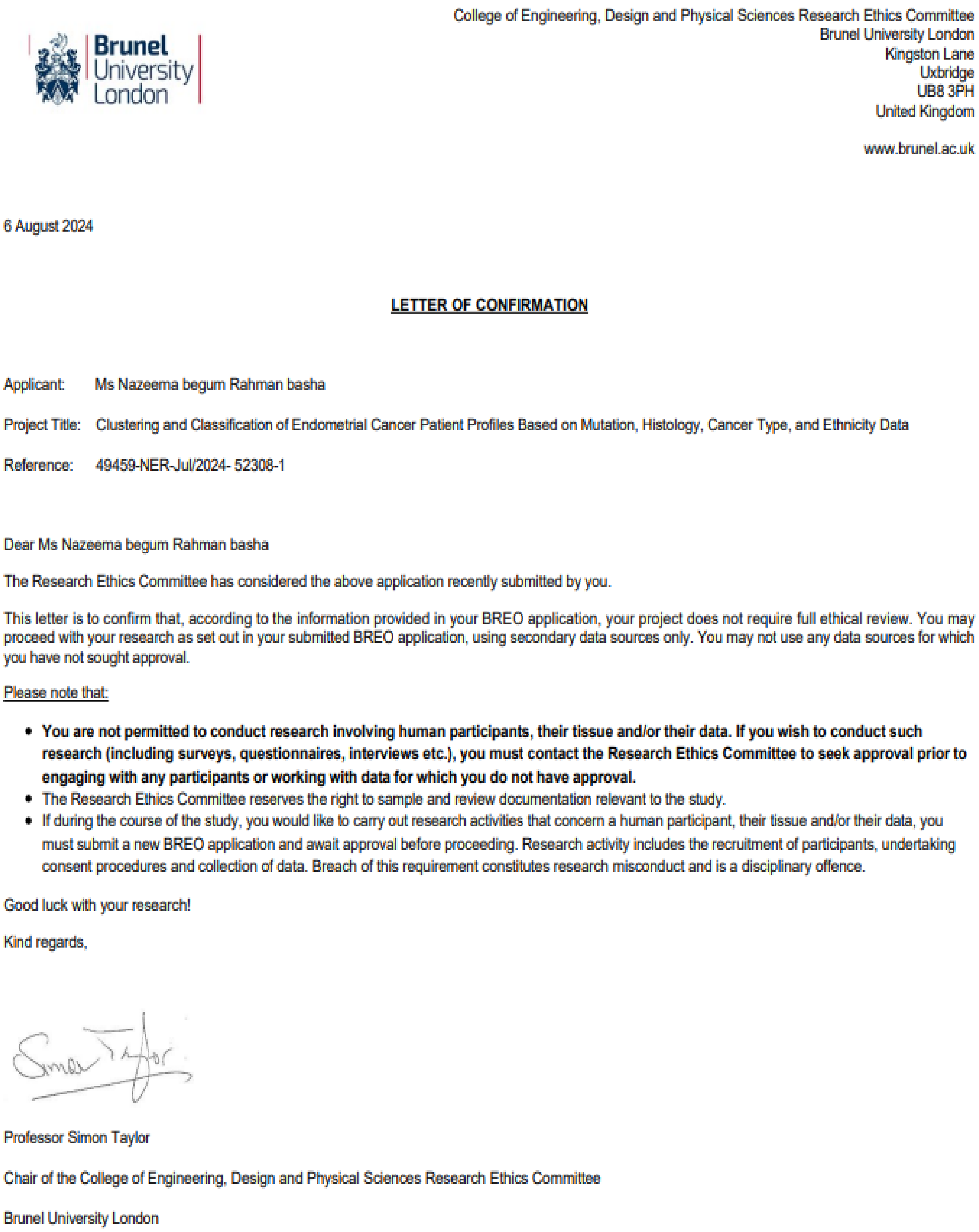
Xu, B., Feng, X. and Burdine, R.D. (2010) ‘Categorical Data Analysis in Experimental Biology’, *Developmental biology*, 348(1), pp. 3–11. Available at: https://doi.org/10.1016/j.ydbio.2010.08.018.

Yin, F. *et al.* (2019) ‘Predicting prognosis of endometrioid endometrial adenocarcinoma on the basis of gene expression and clinical features using Random Forest’, *Oncology Letters*, 18(2), pp. 1597–1606. Available at: https://doi.org/10.3892/ol.2019.10504.

Zhang, Z. & Wang, L. (2017). Advanced statistics using R. https://advstats.psychstat.org.

.

APPENDIX A: ETHICAL APPROVAL



College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

6 August 2024

LETTER OF CONFIRMATION

Applicant: Ms Nazeema begum Rahman basha

Project Title: Clustering and Classification of Endometrial Cancer Patient Profiles Based on Mutation, Histology, Cancer Type, and Ethnicity Data

Reference: 49459-NER-Jul/2024- 52308-1

Dear Ms Nazeema begum Rahman basha

The Research Ethics Committee has considered the above application recently submitted by you.

This letter is to confirm that, according to the information provided in your BREO application, your project does not require full ethical review. You may proceed with your research as set out in your submitted BREO application, using secondary data sources only. You may not use any data sources for which you have not sought approval.

Please note that:

- **You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research (including surveys, questionnaires, interviews etc.), you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.**
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must submit a new BREO application and await approval before proceeding. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,

Simon Taylor

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

APPENDIX B

~~~
---
title: "endometrial cancer"
author: "Nazeema"
date: "2024-08-21"
output: html_document
---


```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```


## R Markdown


This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.


When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:


```{r}
##install.packages("dplyr")
##install.packages("ggplot2")
##install.packages("factoextra")
##install.packages("cluster")
##install.packages("tidyr")
##install.packages("caret")


```
~~~

~~~
```{r}
library(dplyr)
library(ggplot2)
library(cluster)    # For clustering
library(factoextra) # For visualizing clusters
library(tidyr)
library(caret)
```
##1.data preparation and cleaning
##1.1 load data
```{r}
## 1.1 load the data from the endometrial.csv file and inspect it
endometrial_dataset <- read.csv("endometrial.csv")

str(endometrial_dataset)

##having a look at the summary statistics of the dataset before cleaning

summary(endometrial_dataset)

##looking at the top 10 entries of the dataset
head(endometrial_dataset,10)
```
## 1.2. handling missing values
```{r}

# Calculate the sum of mutated gene counts for each patient
endometrial_dataset$calculated_mutation_sum <- rowSums(mutation_columns, na.rm = TRUE)

# Compare the calculated mutation sum with the original Mutation.Count column
~~~

~~~
endometrial_dataset$matches <- endometrial_dataset$Mutation.Count ==
endometrial_dataset$calculated_mutation_sum


# Identify rows with discrepancies

discrepancies <- endometrial_dataset %>% filter(matches == FALSE)

discrepancy_row_numbers <- which(endometrial_dataset$matches == FALSE)

print(discrepancy_row_numbers)

```




##1.3. checking for duplicate entries in the column name


```{r}

# Check for duplicate column names

duplicate_columns <- names(endometrial_cleaned)[duplicated(names(endometrial_cleaned))]


# Display the duplicate columns (if any)

if (length(duplicate_columns) > 0) {

  print(paste("Duplicate column names found:", paste(duplicate_columns, collapse = ", ")))

} else {

  print("No duplicate column names found.")

}

```

##1.4 cleaned data

```{r}

endometrial_cleaned <- endometrial_dataset[-1832,]

View(endometrial_cleaned)

```



#2. EDA

```{r}

# Descriptive statistics for categorical variables
~~~

~~~
table(endometrial_cleaned$Cancer.Type.Detailed)
table(endometrial_cleaned$Ethnicity.Category)
table(endometrial_cleaned$Histology)

# Summary statistics for numerical variables like mutation counts
summary(endometrial_cleaned$Mutation.Count)
```
##2.1Visualizing the distribution of key categorical variables using bar plots.
```{r}
# Bar plot for Cancer Type Detailed
ggplot(endometrial_cleaned, aes(x = Cancer.Type.Detailed)) +
 geom_bar(fill = "skyblue") +
 theme_minimal() +
 labs(title = "Distribution of Cancer Types", x = "Cancer Type Detailed", y = "Count") +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Bar plot for Ethnicity Category
ggplot(endometrial_cleaned, aes(x = Ethnicity.Category)) +
 geom_bar(fill = "lightgreen") +
 theme_minimal() +
 labs(title = "Distribution of Ethnicity", x = "Ethnicity Category", y = "Count") +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Bar plot for Histology
ggplot(endometrial_cleaned, aes(x = Histology)) +
 geom_bar(fill = "lightcoral") +
 theme_minimal() +
 labs(title = "Distribution of Histology", x = "Histology", y = "Count") +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```
~~~

~~~~
##2.2 Investigate relationships between categorical variables


##2.2.1 cancer type detailed vs Ethinicity


```{r}


# Group by Cancer Type and Ethnicity to calculate the count of patients within each Ethnicity
group_ethnicity <- endometrial_cleaned %>%
  group_by(`Ethnicity.Category`, `Cancer.Type.Detailed`) %>%
  summarise(Count = n(), .groups = "drop")

# Calculate the total number of patients within each ethnicity category
total_patients_by_ethnicity <- group_ethnicity %>%
  group_by(`Ethnicity.Category`) %>%
  summarise(Total = sum(Count))

# Join the total patients by ethnicity back to the main dataset
group_ethnicity <- group_ethnicity %>%
  left_join(total_patients_by_ethnicity, by = "Ethnicity.Category") %>%
  mutate(Percentage = (Count / Total) * 100)  # Calculate percentage within each ethnicity
category

# Print the percentage data to the console
print(group_ethnicity)

# Create a bar plot to show the percentage of tumor types within each ethnicity category
ggplot(group_ethnicity, aes(x = `Cancer.Type.Detailed`, y = Percentage, fill =
`Ethnicity.Category`)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(round(Percentage, 1), "%")),
        position = position_dodge(width = 0.9), vjust = -0.5) +  # Adding percentage labels
~~~~

~~~
  labs(title = "Percentage of Tumor Types by Ethnicity (100% Within Each Ethnicity)",
       x = "Tumor Type",
       y = "Percentage Within Each Ethnicity Category") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("Hispanic" = "#FF9999", "Non-Hispanic" = "#9999FF", "Unknown" = "#CCCCCC"))


```


###2.2.2 Distribution of Mutation Counts
##Analyze the distribution of mutation counts. Use histograms or density plots to visualize this distribution.
```{r}
# Histogram of Mutation Count with proper x-axis scale
ggplot(endometrial_cleaned, aes(x = Mutation.Count)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Mutation Counts", x = "Number of different genes mutated", y = "Frequency") +
  scale_x_continuous(breaks = seq(0, max(endometrial_cleaned$Mutation.Count, na.rm = TRUE), by = 5),
                     limits = c(0, max(endometrial_cleaned$Mutation.Count, na.rm = TRUE)))
```



###2.3. Correlations and Pairwise Relationships


###2.3.1 Explore pairwise relationships between numeric variables, such as mutation counts, using box plots
```{r}
# Scatter plot: Mutation Count vs. Cancer Type Detailed
ggplot(endometrial_cleaned, aes(x = Cancer.Type.Detailed, y = Mutation.Count)) +
  geom_boxplot(fill = "lightblue") +
~~~

~~~
 theme_minimal() +
 labs(title = "Mutation Count by Cancer Type", x = "Cancer Type Detailed", y = "Mutation Count") +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Correlation matrix for numeric variables (if applicable)
numeric_data <- endometrial_cleaned %>%
 select(where(is.numeric))

cor_matrix <- cor(numeric_data, use = "complete.obs")
print(cor_matrix)
```



## 3 Cluster Analysis
#3.1 cluster analysis of patients profile based on histology, ethinicity and cancer type detailed, and mutation count

```{r}

# Select relevant columns for clustering
endometrial_select <- endometrial_cleaned %>%
 select(Patient.ID, Cancer.Type.Detailed, Ethnicity.Category, Histology, Mutation.Count)

# Convert categorical variables to factors
endometrial_select$Cancer.Type.Detailed <- as.factor(endometrial_select$Cancer.Type.Detailed)
endometrial_select$Ethnicity.Category <- as.factor(endometrial_select$Ethnicity.Category)
endometrial_select$Histology <- as.factor(endometrial_select$Histology)

# One-hot encode the categorical variables
endometrial_encod <- endometrial_select %>%
~~~

```
  mutate(across(where(is.factor), as.numeric))

# Scale the numeric columns (mutation count and encoded categorical variables)
endometrial_scale <- endometrial_encod %>%
  mutate(across(c(Mutation.Count, Cancer.Type.Detailed, Ethnicity.Category, Histology), scale))

# Determine the optimal number of clusters using the elbow method
fviz_nbclust(endometrial_scale %>% select(-Patient.ID), kmeans, method = "wss")

# Set the number of clusters based on the elbow plot
set.seed(123)
k <- 4  # the number of clusters based on elbow plot
kmeans_result2 <- kmeans(endometrial_scale %>% select(-Patient.ID), centers = k, nstart = 25)

# Add cluster results to the original data
endometrial_select$Cluster <- as.factor(kmeans_result2$cluster)

# Visualize the clusters
fviz_cluster(kmeans_result2, data = endometrial_scale %>% select(-Patient.ID),
        geom = "point",
        ellipse.type = "norm",
        main = "Cluster Plot of Patients Based on Histology, Ethnicity, Cancer Type, and Mutation Count")

# Check the distribution of Cancer Types, Ethnicity, Histology, and Mutation Count within each cluster
cancer_type_distribution <- table(endometrial_select$Cluster, endometrial_select$Cancer.Type.Detailed)
ethnicity_distribution <- table(endometrial_select$Cluster, endometrial_select$Ethnicity.Category)
histology_distribution <- table(endometrial_select$Cluster, endometrial_select$Histology)

# Display the distributions
```

~~~
print("Distribution of Cancer Type Detailed within clusters:")
print(cancer_type_distribution)

print("Distribution of Ethnicity Category within clusters:")
print(ethnicity_distribution)

print("Distribution of Histology within clusters:")
print(histology_distribution)


```




## 3.2 clustering including patient ID, categorical columns, and all gene columns
```{r}

# Step 2: Select the relevant columns for clustering, including patient ID, categorical columns,
and all gene columns
# Identify the gene columns using a pattern match
gene_columns <- grep("^Gene:", names(endometrial_cleaned), value = TRUE)  # Gene columns

# Select relevant columns including patient data, clinical information, and gene mutation data
endometrial_select <- endometrial_cleaned %>%
  select(Patient.ID, Cancer.Type.Detailed, Ethnicity.Category, Histology, Mutation.Count,
all_of(gene_columns))

# Step 3: Ensure that the clinical categorical variables are treated as factors
endometrial_select <- endometrial_select %>%
  mutate(across(c(Cancer.Type.Detailed, Ethnicity.Category, Histology), as.factor))
~~~

```
# Step 4: Convert categorical variables (like Cancer Type, Ethnicity, Histology) to numeric for clustering
# Use one-hot encoding if necessary for categorical data
endometrial_encoded <- endometrial_select %>%
  mutate(across(where(is.factor), as.numeric))

# Step 5: Scale the numeric columns, including mutation count, the encoded categorical variables, and all gene columns
endometrial_scaled <- endometrial_encoded %>%
  mutate(across(c(Mutation.Count, starts_with("Gene:")), scale))

# Step 6: Check that the gene columns are correctly included
print("Gene columns included for clustering:")
print(gene_columns)

# Step 7: Determine the optimal number of clusters using the elbow method
# Exclude Patient.ID from clustering
fviz_nbclust(endometrial_scaled %>% select(-Patient.ID), kmeans, method = "wss")

# Step 8: Perform k-means clustering
set.seed(123)
k <- 4  # Set the number of clusters (this can be adjusted based on elbow plot)
kmeans_result <- kmeans(endometrial_scaled %>% select(-Patient.ID), centers = k, nstart = 25)

# Step 9: Add the cluster results to the original data
endometrial_select$Cluster <- as.factor(kmeans_result$cluster)

# Step 10: Visualize the clusters using PCA to reduce dimensionality for visualization
fviz_cluster(kmeans_result, data = endometrial_scaled %>% select(-Patient.ID),
        geom = "point", ellipse.type = "norm",
        main = "Cluster Plot of Patients Based on mutated Gene, Ethnicity, histology, cancer type detailed")
```

~~~
# Step 11: Analyze the distribution of genes and clinical data across clusters
# Summarize the number of patients with mutations in each gene across clusters
gene_distribution <- endometrial_select %>%
  group_by(Cluster) %>%
  summarise(across(starts_with("Gene:"), ~ sum(. > 0, na.rm = TRUE)))  # Count patients with mutations in each gene

# Print gene mutation distribution across clusters
print("Gene mutation distribution within clusters:")
print(as.data.frame(gene_distribution))

# Step 12: Analyze the distribution of clinical features within each cluster
# Distribution of Cancer Type, Ethnicity, and Histology within each cluster
cancer_type_distribution <- table(endometrial_select$Cluster, endometrial_select$Cancer.Type.Detailed)
ethnicity_distribution <- table(endometrial_select$Cluster, endometrial_select$Ethnicity.Category)
histology_distribution <- table(endometrial_select$Cluster, endometrial_select$Histology)

# Print clinical feature distributions
print("Distribution of Cancer Type within clusters:")
print(cancer_type_distribution)

print("Distribution of Ethnicity Category within clusters:")
print(ethnicity_distribution)

print("Distribution of Histology within clusters:")
print(histology_distribution)

```
#4. Modelling and Evaluation
~~~

~~~~
##4.1 classifying  the patient's cancer type based on the gene mutation

```{r}
# Load the required libraries
library(dplyr)
library(readr)

# Extract gene mutation columns (from column 17 to 319) and the cancer type detailed column (column 5)
mutation_data <- endometrial_cleaned[, 17:319]  # Extract gene mutation columns
cancer_types <- endometrial_cleaned[, 5]  # Extract cancer type detailed column

# Identify and exclude non-numeric columns
non_numeric_columns <- sapply(mutation_data, function(col) !is.numeric(as.numeric(col)))
mutation_data_clean <- mutation_data[, !non_numeric_columns]  # Keep only numeric columns

# Convert remaining columns to numeric (this should now avoid introducing NA values)
mutation_data_clean <- mutation_data_clean %>%
  mutate(across(everything(), ~as.numeric(.)))

# Combine cancer type detailed with cleaned gene mutation data
combined_data <- bind_cols(cancer_types, mutation_data_clean)

# Rename the first column to something easier to work with
colnames(combined_data)[1] <- "Cancer_Type_Detailed"

# Calculate the sum of mutations for each gene by cancer type
gene_mutation_summary <- combined_data %>%
  group_by(Cancer_Type_Detailed) %>%
  summarise(across(everything(), sum, na.rm = TRUE))
~~~~

````
# View the result: which genes are mutated in which cancer type
print(gene_mutation_summary)

` ` `



##4.2 Random Forest Model

` ` `{r}

# Load the required libraries
 # For machine learning functions
library(randomForest)  # For Random Forest model
library(readr)

 #Extract gene mutation columns (from column 17 to 319) and the cancer type detailed column (column 5)
mutation_data <- endometrial_cleaned[, 17:319]  # Extract gene mutation columns
cancer_types <- endometrial_cleaned[, 5]  # Extract cancer type detailed column

# Combine cancer type detailed with gene mutation data
combined_data <- bind_cols(cancer_types, mutation_data)

# Rename the first column to something easier to work with
colnames(combined_data)[1] <- "Cancer_Type_Detailed"

# Convert cancer type to a factor (for classification purposes)
combined_data$Cancer_Type_Detailed <- as.factor(combined_data$Cancer_Type_Detailed)

# Ensure all mutation data is numeric
combined_data <- combined_data %>%
````

```
  mutate(across(starts_with("Gene"), ~as.numeric(.)))

# Split the data into training and testing sets (80% training, 20% testing)
set.seed(123)  # For reproducibility
train_index <- createDataPartition(combined_data$Cancer_Type_Detailed, p = 0.8, list = FALSE)
train_data <- combined_data[train_index, ]
test_data <- combined_data[-train_index, ]

# Fit a Random Forest model to classify cancer types based on gene mutations
rf_model <- randomForest(Cancer_Type_Detailed ~ ., data = train_data, importance = TRUE)

# Predict the cancer types on the test set
predictions <- predict(rf_model, test_data)

# Evaluate the model's accuracy
conf_matrix <- confusionMatrix(predictions, test_data$Cancer_Type_Detailed)

# Print the confusion matrix to check model performance
print(conf_matrix)

# Optional: View the importance of each gene in predicting cancer types
importance(rf_model)

# Save the model results to a file if needed
write.csv(data.frame(test_data$Cancer_Type_Detailed, predictions),
"cancer_type_predictions.csv")

```

##4.3 Classify the cancer type based on Top 10 most important genes

~~~
```{r}
# Load necessary libraries
  # For machine learning functions
library(randomForest)  # For Random Forest model
library(readr)



# Extract gene mutation columns (from column 17 to 319) and the cancer type detailed column (column 5)
mutation_data <- endometrial_cleaned[, 17:319]  # Extract gene mutation columns
cancer_types <- endometrial_cleaned[, 5]  # Extract cancer type detailed column

# Combine cancer type detailed with gene mutation data
combined_data <- bind_cols(cancer_types, mutation_data)

# Rename the first column to something easier to work with
colnames(combined_data)[1] <- "Cancer_Type_Detailed"

# Convert cancer type to a factor (for classification purposes)
combined_data$Cancer_Type_Detailed <- as.factor(combined_data$Cancer_Type_Detailed)

# Check for non-numeric values in gene columns
non_numeric_values <- combined_data %>%
  select(starts_with("Gene")) %>%
  summarise_all(~ sum(!is.numeric(.)))

cat("Non-numeric values in gene columns:\n")
print(non_numeric_values)

#replace with NA
combined_data <- combined_data %>%
~~~

```
  mutate(across(starts_with("Gene"), ~ as.numeric(as.character(.)), .names = "numeric_{col}"))

# Replace NA values with 0
combined_data[is.na(combined_data)] <- 0

# Get the column names of the gene mutation data
gene_column_names <- colnames(mutation_data)

# Function to get top 10 important genes for each cancer type
get_top_genes <- function(cancer_type, data, gene_column_names) {
  # Create a binary target variable for the specific cancer type (1 = cancer_type, 0 = other types)
  data$Cancer_Type_Binary <- ifelse(data$Cancer_Type_Detailed == cancer_type, 1, 0)

  # Select the gene mutation columns by their names
  gene_data <- data[, gene_column_names]
  target <- as.factor(data$Cancer_Type_Binary)  # Convert to factor for binary classification

  # Split data into training and testing sets
  set.seed(123)
  train_index <- createDataPartition(target, p = 0.8, list = FALSE)
  train_data <- gene_data[train_index, ]
  train_target <- target[train_index]
  test_data <- gene_data[-train_index, ]
  test_target <- target[-train_index]

  # Check for and remove rows with missing values in train and test data
  train_data <- train_data[complete.cases(train_data), ]
  test_data <- test_data[complete.cases(test_data), ]

  # Train Random Forest model
  rf_model <- randomForest(x = train_data, y = train_target, importance = TRUE)
```

```
 # Predict on test data
 predictions <- predict(rf_model, test_data)

 # Evaluate model performance
 confusion_mat <- confusionMatrix(predictions, test_target, positive = "1")

 # Get importance of genes
 importance_df <- as.data.frame(importance(rf_model))

 # Rank genes by MeanDecreaseAccuracy and MeanDecreaseGini
 importance_df$Gene <- rownames(importance_df)

 # Sort by MeanDecreaseAccuracy and take top 10 genes
 top_genes <- importance_df %>%
  arrange(desc(MeanDecreaseAccuracy)) %>%
  slice(1:10)  # Get top 10 genes

 # Add the cancer type label to the top_genes table
 top_genes <- top_genes %>%
  mutate(Cancer_Type = cancer_type)

 # Return top genes and model performance
 return(list(top_genes = top_genes, confusion_matrix = confusion_mat))
}

# List of unique cancer types in endometrial dataset
cancer_types <- unique(combined_data$Cancer_Type_Detailed)

# Store results for each cancer type
results_by_cancer_type <- list()
```

~~~~
# Loop through each cancer type and get top 10 genes and model performance
for (cancer_type in cancer_types) {
  cat("\nProcessing cancer type:", cancer_type, "\n")  # Show progress
  result <- get_top_genes(cancer_type, combined_data, gene_column_names)
  results_by_cancer_type[[cancer_type]] <- result
}

# Print top 10 genes and confusion matrix for each cancer type
for (result in results_by_cancer_type) {
  cat("\n======== Top 10 genes for:", result$cancer_type, "========\n")
  print(result$top_genes)

  cat("\n======== Confusion Matrix for:", result$cancer_type, "========\n")
  print(result$confusion_matrix)
}

```
##4.4 Visualisation for the top 10 gene for various cancer type
```{r}
all_top_genes <- data.frame()

# Loop through each cancer type and get top 10 genes and model performance
for (cancer_type in cancer_types) {
  cat("\nProcessing cancer type:", cancer_type, "\n")  # Show progress
  result <- get_top_genes(cancer_type, combined_data, gene_column_names)
  results_by_cancer_type[[cancer_type]] <- result

  # Combine top genes for each cancer type into the all_top_genes dataframe
  all_top_genes <- rbind(all_top_genes, result$top_genes)
}
~~~~

~~~~
` ` `


` ` `{r}
# Plot only for "Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma" with pink bars
ggplot(filter(all_top_genes, Cancer_Type == "Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma"),
       aes(x = reorder(Gene, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy, fill = Cancer_Type)) +
  geom_bar(stat = "identity", color = "black") +
  coord_flip() +
  labs(
    title = "Top Genes by Mean Decrease Accuracy for Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma",
    x = "Genes",
    y = "Mean Decrease Accuracy"
  ) +
  scale_fill_manual(values = c("Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma" = "pink")) +  # Set the cancer type color to pink
  theme_minimal()


` ` `



` ` `{r}
# Plot only for "Uterine Mixed Endometrial Carcinoma" with yellow bars
ggplot(filter(all_top_genes, Cancer_Type == "Uterine Mixed Endometrial Carcinoma"),
       aes(x = reorder(Gene, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy, fill = Cancer_Type)) +
  geom_bar(stat = "identity", color = "black") +
  coord_flip() +
  labs(
    title = "Top Genes by Mean Decrease Accuracy for Uterine Mixed Endometrial Carcinoma",
    x = "Genes",
~~~~

~~~
    y = "Mean Decrease Accuracy"
  ) +
  scale_fill_manual(values = c("Uterine Mixed Endometrial Carcinoma" = "yellow")) +  # Set the cancer type color to yellow
  theme_minimal()

```



```{r}
# Plot only for "Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor" with orange bars
ggplot(filter(all_top_genes, Cancer_Type == "Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor"),
       aes(x = reorder(Gene, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy, fill = Cancer_Type)) +
  geom_bar(stat = "identity", color = "black") +
  coord_flip() +
  labs(
    title = "Top Genes by Mean Decrease Accuracy for Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor",
    x = "Genes",
    y = "Mean Decrease Accuracy"
  ) +
  scale_fill_manual(values = c("Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor" = "orange")) +  # Set the cancer type color to orange
  theme_minimal()

```


```{r}
# Plot only for "Uterine Undifferentiated Carcinoma" with purple bars
ggplot(filter(all_top_genes, Cancer_Type == "Uterine Undifferentiated Carcinoma"),
~~~

~~~
      aes(x = reorder(Gene, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy, fill = Cancer_Type)) +
  geom_bar(stat = "identity", color = "black") +
  coord_flip() +
  labs(
    title = "Top Genes by Mean Decrease Accuracy for Uterine Undifferentiated Carcinoma",
    x = "Genes",
    y = "Mean Decrease Accuracy"
  ) +
  scale_fill_manual(values = c("Uterine Undifferentiated Carcinoma" = "purple")) +  # Set the cancer type color to purple
  theme_minimal()

```



```{r}
# Plot only for "Uterine Clear Cell Carcinoma" with green bars
ggplot(filter(all_top_genes, Cancer_Type == "Uterine Clear Cell Carcinoma"),
      aes(x = reorder(Gene, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy, fill = Cancer_Type)) +
  geom_bar(stat = "identity", color = "black") +
  coord_flip() +
  labs(
    title = "Top Genes by Mean Decrease Accuracy for Uterine Clear Cell Carcinoma",
    x = "Genes",
    y = "Mean Decrease Accuracy"
  ) +
  scale_fill_manual(values = c("Uterine Clear Cell Carcinoma" = "green")) +  # Set the cancer type color to green
  theme_minimal()
```
~~~

~~~
```{r}
# Plot only for "Uterine Endometrioid Carcinoma"
ggplot(filter(all_top_genes, Cancer_Type == "Uterine Endometrioid Carcinoma"),
       aes(x = reorder(Gene, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy, fill = Cancer_Type)) +
  geom_bar(stat = "identity", color = "black") +
  coord_flip() +
  labs(
    title = "Top Genes by Mean Decrease Accuracy for Uterine Endometrioid Carcinoma",
    x = "Genes",
    y = "Mean Decrease Accuracy"
  ) +
  theme_minimal()

```

```{r}
# Plot only for "Endometrial Carcinoma" with blue bars
ggplot(filter(all_top_genes, Cancer_Type == "Endometrial Carcinoma"),
       aes(x = reorder(Gene, MeanDecreaseAccuracy), y = MeanDecreaseAccuracy, fill = Cancer_Type)) +
  geom_bar(stat = "identity", color = "black") +
  coord_flip() +
  labs(
    title = "Top Genes by Mean Decrease Accuracy for Endometrial Carcinoma",
    x = "Genes",
    y = "Mean Decrease Accuracy"
  ) +
  scale_fill_manual(values = c("Endometrial Carcinoma" = "blue")) +  # Set the correct cancer type color to blue
~~~

~~~
 theme_minimal()
` ` `
~~~