Department of Computer Science

MSc Artificial Intelligence

Academic Year 2023-2024

Enhancing Hepatitis C virus management through advance machine learning techniques in Egypt

Student Name: Navneet Kaur

Student ID: 2355421

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

Hepatitis C Virus (HCV) affects millions globally, causing liver diseases like cirrhosis and cancer. This research investigates the potential of advanced machine learning (ML) techniques in predicting HCV disease progression, aiming to improve early detection and patient management. The study began by addressing missing data through imputation and encoding categorical variables to ensure dataset integrity. Several ML models were developed, including Logistic Regression, Decision Trees, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and XGBoost, to classify HCV stages.

Among these models, XGBoost performed the best with an accuracy of 32%, outperforming Logistic Regression (28%) and Decision Trees (24%). However, all models struggled with precision and recall due to class imbalances. Key predictors identified through Recursive Feature Elimination (RFE) and Random Forests included ALT, BMI, and RNA levels, which are clinically significant in liver disease prognosis.

Additionally, advanced ensemble methods such as Voting and Stacking Classifiers were tested, alongside a neural network, which achieved the highest accuracy at 35%. While the neural network demonstrated better ability to recognize patterns, classifying all stages with precision remains a challenge.

This study highlights the potential of ML in improving HCV management, though further research is needed to handle class imbalances and improve model performance. The integration of additional clinical data and real-time systems could further enhance clinical decision-making, providing personalized treatment plans and improving patient outcomes.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, **Dr Katie Mintram** for her invaluable guidance, encouragement, and support throughout the course of this dissertation. Her expertise, insightful feedback, and commitment to my research have been instrumental in shaping this project. I am deeply appreciative of the time and effort she has invested in helping me overcome challenges and for always pushing me to strive for excellence.

I would also like to extend my heartfelt thanks to **Dr. Sarath Dantu**, my project module leader, for his continuous support and for providing me with the foundational knowledge necessary for this dissertation. His thoughtful advice and academic expertise have greatly contributed to my understanding of machine learning and its application in healthcare.

Lastly, I would like to thank my family and friends for their unwavering support, patience, and encouragement throughout this journey. Their belief in me has been a constant source of motivation. This dissertation would not have been possible without the guidance and support of the aforementioned individuals, and I am deeply grateful for their contributions

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

Signature: Navneet Kaur

Date: 07-09-2024

TOTAL NUMBER OF WORDS: 12,419

Contents

Chapter 1: Introduction

Hepatitis C Virus (HCV) affects 71 million people worldwide and causes liver diseases such cirrhosis and hepatocellular cancer (WHO, 2021). HCV management and treatment have improved over time. However, early detection and accurate illness progression prediction are essential for patient outcomes. Machine learning (ML) offers powerful predictive modelling and data analysis to improve HCV management.

Machine Learning (ML), a subfield of AI, creates algorithms that learn from data and predict. Healthcare professionals are increasingly using machine learning for disease identification and customised treatment (Jordan & Mitchell, 2015). Machine learning can analyse complex HCV datasets and find patterns and relationships that statistical methods may miss. HCV infection involves several biological, demographic, and clinical factors, making this feature particularly useful (Ghazal et al., 2023).

Multiple steps are involved in HCV management using machine learning. Exploratory Data Analysis (EDA) and preprocessing are done first to understand data distribution and correlations. Histograms, box plots, and scatter plots identify patterns and anomalies, while preprocessing steps like addressing missing values, encoding categorical variables, and normalising numerical features ensure data quality (Alizargar et al., 2023; Chew & Khaw, 2023).

Next, HCV results are predicted using SVM, Decision Tree, and Random Forest classification methods. The models categorise HCV stages and predict patient outcomes with great accuracy and resilience (Alizargar et al., 2023). Linear Regression, Ridge Regression, and Gradient Boosting Machines are utilised to quantify clinical factors and illness severity for continuous outcomes such RNA levels (Ghazal et al., 2023).

K-Means and Hierarchical Clustering are used to categorise patients by shared attributes, making clinical profiles easier to identify. Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) reduce the number of features while keeping important information, simplifying data presentation and understanding (Chew & Khaw, 2023).

HCV care can be greatly improved by using machine learning into clinical practice. ML models can improve patient care and therapy by enabling early diagnosis, precise illness progression prediction, and identification of crucial patient outcomes determinants. To maximise the benefits of machine learning in controlling hepatitis C virus (HCV), further research should focus rigorous result verification, encourage multidisciplinary collaboration, and establish real-time clinical workflow systems (Visani et al., 2022).

1.1 Problem Statement

Hepatitis C Virus (HCV) remains a significant global public health issue, contributing to high rates of morbidity and mortality. Despite advances in antiviral treatments, managing HCV effectively is complex due to its diverse nature and unpredictable patient responses. Accurate disease progression prediction is vital for improving patient outcomes, but traditional methods often lack the precision needed.

Current diagnostic and prognostic tools for HCV rely on basic statistical techniques that fail to capture the intricate interactions of clinical, demographic, and genetic factors influencing disease progression. This makes it difficult for clinicians to predict severe complications, like cirrhosis or liver cancer, and tailor treatments accordingly.

Machine learning (ML) offers a promising solution, as it can analyse large, complex datasets to uncover patterns traditional methods might miss. However, ML in HCV management is still in its early stages, with significant challenges remaining. This dissertation aims to address these gaps by developing an integrated ML framework to enhance HCV patient outcomes.

1.1.1 Problem Statement Question

How can advanced machine learning techniques be systematically applied to HCV datasets to improve the early detection and accurate prediction of disease progression, thereby enhancing patient management and treatment outcomes?

1.2 Aim and Objectives

The aim of this research is to leverage advanced machine learning techniques to analyse Hepatitis C Virus (HCV) patient data, with the goal of predicting disease outcomes, understanding key factors influencing patient health, and identifying patterns that can improve clinical decision-making and patient management.

1.2.1 Objectives

Data preprocessing involves addressing missing values in the HCV dataset through imputation or removal techniques to ensure reliability. Categorical variables like gender and age groups are encoded into numerical formats using methods such as one-hot encoding.

For feature selection, Recursive Feature Elimination (RFE) with a Logistic Regression model is applied to identify key features, reducing dimensionality. Additionally, a Random Forest model is used to assess feature importance, highlighting significant predictors.

In developing classification models, a variety of algorithms such as Logistic Regression, Decision Trees, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and XGBoost are constructed and refined. Ensemble methods like Voting and Stacking Classifiers are implemented. Model performance is evaluated using accuracy, precision, recall, F1-scores, and ROC curves, with optimization through cross-validation and hyperparameter tuning.

Results are summarized, highlighting key features and model performance, with actionable insights provided. Visualizations like feature importance plots and confusion matrices are used for clear reporting to stakeholders.

1.3 Research Approach

This study utilises a systematic method to improve the management of Hepatitis C Virus (HCV) using sophisticated machine learning algorithms. The methodology commences with data pretreatment and cleansing, wherein missing values are addressed through imputation or elimination techniques, categorical variables are transformed into numerical representations, and numerical characteristics are standardised and scaled for uniformity. Subsequently, an exploratory data analysis (EDA) is performed to acquire a deeper understanding of the dataset. Descriptive statistical analysis provides a summary of each feature, while visualisations such as histograms, box plots, and scatter plots aid in comprehending the distributions and correlations of the features.

During the predictive modelling phase, a range of classification models such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks are constructed to forecast binary and multiclass results. In addition, regression models such as Linear Regression, Ridge Regression, Lasso Regression, and Gradient Boosting Machines are utilised to forecast continuous outcomes, such as RNA levels. Clustering methods such as K-Means, Hierarchical Clustering, and DBSCAN are utilised to categorise patients according to shared attributes. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) are employed to decrease the number of features in a dataset while preserving crucial information. This process aids in simplifying the visualisation and comprehension of the data.

The last phase entails assessing the model, validating its performance, and analysing its features. Evaluating the performance of a model involves analysing parameters like as accuracy, precision, recall, and F1-score for classification tasks, and Mean Squared Error (MSE) and R-squared (R^2) for regression tasks. Cross-validation procedures guarantee the robustness and generalizability of the results. Hyperparameter tweaking is conducted in order to optimise the performance of the model. Significant characteristics are determined by methods such as Recursive Feature Elimination (RFE) and feature relevance scores, offering valuable insights into the primary determinants for HCV outcomes. The project intends to construct strong and precise machine learning models that can greatly improve the prediction and management of HCV.

1.4 Dissertation Outline

This dissertation explores the use of advanced machine learning techniques to enhance the management and prediction of Hepatitis C Virus (HCV) outcomes. It is organized into seven chapters, each focusing on different aspects of the research.

Chapter 1 introduces the research, highlighting the global impact of HCV on public health and the need for timely and accurate disease forecasting. It also discusses the limitations of traditional diagnostic tools and the potential of machine learning (ML) to address these challenges, outlining the research objectives and significance.

Chapter 2 provides a literature review, examining the current understanding of HCV management and the use of machine learning in healthcare. It reviews studies using ML on HCV datasets, focusing on their methodologies and findings, while identifying gaps in the research to justify the need for a comprehensive ML framework for HCV management.

Chapter 3 outlines the methodology, detailing the research design and procedures, including data sources, preprocessing steps (e.g., handling missing values and encoding categorical variables), and exploratory data analysis. It explains the development of classification, regression, clustering, and dimensionality reduction models, emphasizing the importance of model evaluation and feature analysis to ensure accuracy.

Chapter 4 covers the implementation of the proposed ML framework, describing the software tools (e.g., Python, Scikit-learn, TensorFlow) and the step-by-step process for data preprocessing, model training, and evaluation. It addresses challenges encountered during execution and offers practical guidance for replicating the research.

Chapter 5 presents the results, comparing the performance of various machine learning models using metrics like accuracy, precision, recall, F1-score, Mean Squared Error (MSE), and R-squared. Visual aids such as confusion matrices, ROC curves, and feature importance plots are used to illustrate model effectiveness. The chapter also explores clustering and dimensionality reduction techniques.

Chapter 6 interprets the findings within the context of the research objectives and literature, discussing the implications for clinical practice. It highlights the ability of ML models to improve HCV management through early detection and personalized treatments, while also identifying the study’s limitations and proposing areas for future research.

Chapter 7 concludes the dissertation by summarizing its key contributions, emphasizing the role of machine learning in revolutionizing HCV care and improving patient outcomes.

Chapter 2: Literature Review

2.1 Introduction

Hepatitis C Virus (HCV) is a significant global health issue, affecting around 71 million people worldwide (Lazarus, Roel and Elsharkawy, 2019) as shown in figure 1. It primarily infects the liver, leading to chronic diseases such as cirrhosis and hepatocellular carcinoma. The virus is commonly transmitted through exposure to infected blood, with intravenous drug use and improper sterilization of medical equipment being primary causes. Early stages of HCV are often asymptomatic, resulting in delayed diagnosis and increased risk of severe liver damage.

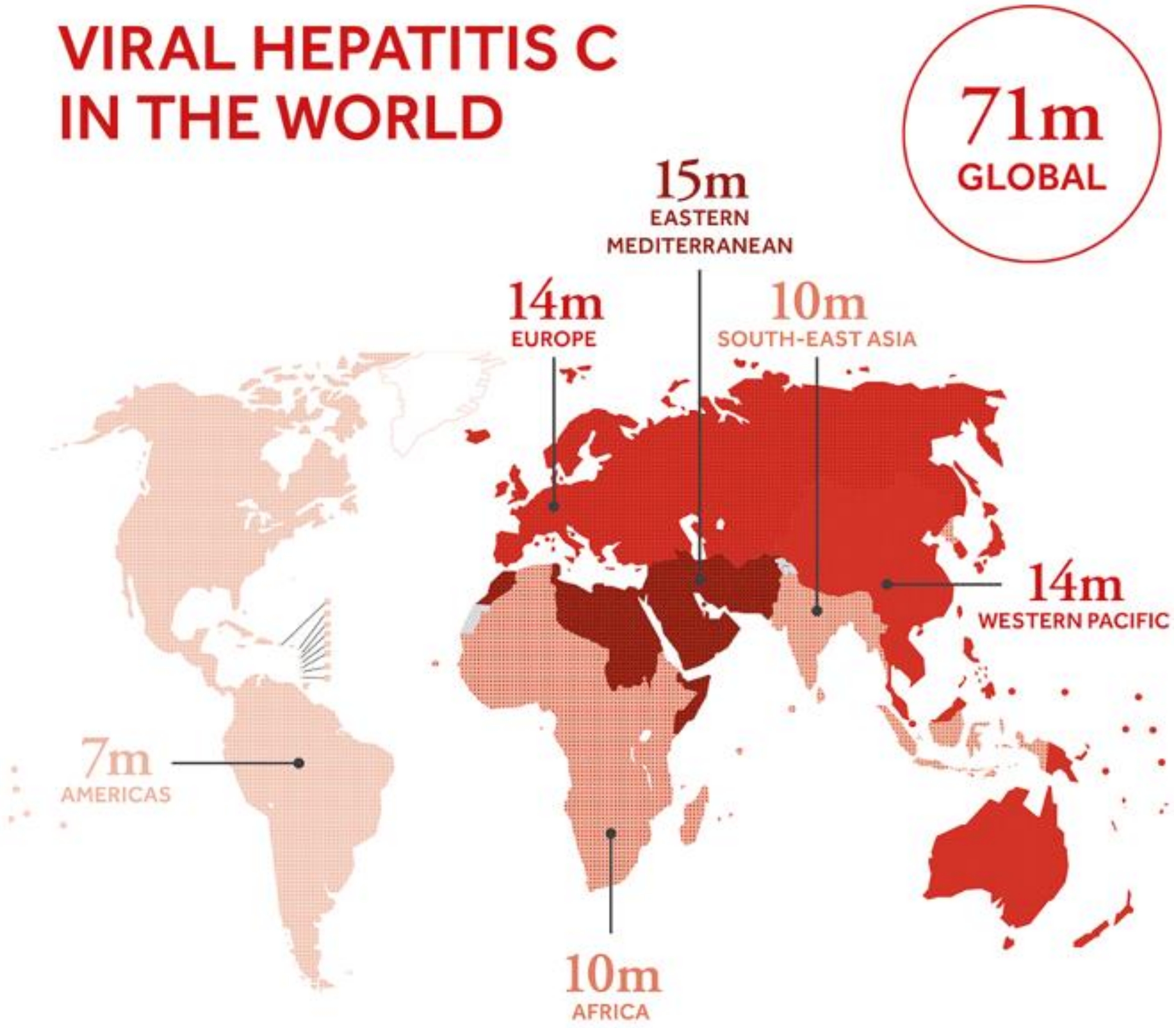


Figure 1: Viral Hepatitis C in the World

Early detection is crucial in HCV management, as it significantly improves treatment outcomes, with studies showing cure rates of over 90% with timely intervention (Melendez-Torres and Singal, 2022). Accurate prediction models help healthcare providers identify patients at risk of rapid disease progression, enabling personalized treatment plans and better resource allocation, which reduces the likelihood of advanced liver disease and improves patient outcomes.

Machine learning (ML) is transforming healthcare by analysing large datasets to make accurate predictions. In HCV management, ML can help with early detection, personalized treatment, and improved patient care by processing patient data and predicting outcomes with high accuracy. ML models have shown over 80% accuracy in predicting disease progression, making them valuable tools in clinical decision-making for chronic conditions like HCV (Yağanoğlu, 2022).

This literature review explores recent advancements in machine learning applied to HCV management. It examines methodologies, evaluates the effectiveness of different ML models, and discusses the clinical implications and potential future directions for integrating ML into HCV treatment.

2.2 Machine Learning Techniques in HCV Prediction

Machine learning (ML) plays a vital role in healthcare, offering predictive analytics, diagnostic support, and personalized treatment strategies. In infectious diseases like HCV, ML is particularly useful for predicting disease progression, identifying patients at risk, and optimizing treatment plans. For instance, ML models have achieved over 85% accuracy in predicting patient outcomes across various healthcare applications.

In the context of HCV, ML helps identify patients who may develop severe liver complications early, allowing for targeted interventions. These models analyse complex data, such as genetic and clinical factors, to provide personalized risk assessments, improving prognosis. Additionally, ML aids in discovering new biomarkers and therapeutic targets. Studies have shown that ML models can reach up to 97.9% accuracy in HCV staging, making them promising for clinical use (Kay et al., 2022).

Several ML techniques are employed in HCV management. Data preprocessing and exploratory data analysis (EDA) methods like histograms and scatter plots ensure data quality and consistency. Chew and Khaw (2023) emphasized the importance of data normalization and handling missing values to avoid bias in predictions. Classification models like Support Vector Machines (SVM), Decision Trees, Random Forests, and Logistic Regression are commonly used for HCV stage prediction. Alizargar et al. (2023) demonstrated that SVM and XGBoost models achieved over 80% accuracy and AUC in predicting HCV outcomes.

Regression analysis techniques, such as Linear, Ridge, and Lasso Regression, predict continuous outcomes like RNA levels. Ghazal et al. (2023) found Gradient Boosting Machines highly effective in capturing complex data patterns. Clustering and dimensionality reduction methods, including K-Means, PCA, and t-SNE, are used to group patients and simplify data visualization. Chew and Khaw (2023) applied K-Means clustering to identify patient subgroups for personalized treatments.

Feature selection techniques like Recursive Feature Elimination (RFE) and feature importance scores help identify key predictors, enhancing model performance and interpretability. Ribeiro et al. (2016) stressed the value of feature selection for improving model accuracy and understanding.

Overall, these methods form a robust framework for enhancing HCV management through advanced ML applications.

2.3 Exploratory Data Analysis (EDA) and Preprocessing

Exploratory Data Analysis (EDA) is essential for understanding and preparing HCV datasets for machine learning. EDA uses visual ways to summarise data attributes to find patterns, anomalies, and test hypotheses. EDA helps discover variable distribution, feature correlations, and data quality issues in complicated HCV datasets. This helps construct strong and accurate predictive models by ensuring that following modelling efforts are based on thorough data understanding.

Visualisation is crucial in EDA. Histograms show skewness, kurtosis, and outliers in a single variable. Box plots exhibit outliers and data distribution symmetry by showing data spread and central tendency (Cooksey, 2020b). Scatter plots reveal correlations and patterns between two continuous variables. Scatter plots can show how HCV RNA levels link with other clinical parameters, helping identify useful predictors. Preprocessing addresses missing values, categorical variables, and feature scaling to prepare raw data for analysis. Missing values can bias findings and reduce model accuracy, therefore handling them is critical. Missing values can be imputed with statistical measures such the mean, median, or mode, or removed if the missing data is negligible. Machine learning algorithms need numerical input, thus categorical variables must be encoded. Label encoding gives each category a unique number, while one-hot encoding provides binary columns, removing any ordinal association. Normalising numerical features improves machine learning model performance and training stability by standardising them. Min-Max Scaling and Standardisation, which provide features zero mean and unit variance, are common strategies.

To avoid model biases, Chew and Khaw (2023) pre-processed an Egyptian HCV dataset using data normalisation and missing value handling. They showed that accurate model training and evaluation require proper preprocessing. This case study shows how preprocessing can improve HCV management machine learning model performance and reliability.

2.4 Classification Models

Classification models are essential for categorizing data into predefined classes, particularly in HCV management, where they predict disease stages or outcomes, such as chronic versus non-chronic infection, based on patient data (Alizargar, Chang and Tan, 2023). Commonly used classification algorithms include Support Vector Machines (SVM), Decision Trees, Random Forests, and Logistic Regression.

Support Vector Machines (SVM) are highly effective in classification tasks, especially in high-dimensional spaces, due to their ability to find the optimal hyperplane separating different classes (Dayananda, 2023). SVMs are robust to overfitting and work well when a clear margin exists between classes, making them effective for classifying HCV stages with high accuracy and Area Under the Curve (AUC) scores. Decision Trees, on the other hand, classify data by splitting it into subsets based on feature values, creating a simple, interpretable model. Although prone to overfitting, they provide valuable insights into key features influencing HCV outcomes. Nandipati et al. (2020) observed that Decision Trees offered significant insights into HCV data, but their accuracy varied with model complexity.

Random Forests, an ensemble learning technique, improve classification performance and reduce overfitting by aggregating predictions from multiple Decision Trees. This method enhances robustness and accuracy in classifying HCV stages. Alizargar et al. (2023) demonstrated that Random Forests achieved high accuracy, making them particularly suitable for clinical applications. Logistic Regression, a simpler binary classification model, estimates the probability that an input belongs to a specific class. It is frequently used as a baseline model due to its ease of implementation and interpretation, and it has been successfully applied to predict specific conditions in HCV patients.

When comparing these models, it becomes evident that more complex algorithms like SVM and Random Forests often outperform simpler ones such as Logistic Regression, particularly in capturing intricate data patterns (Evi Diana Omar et al., 2024). For example, Alizargar et al. (2023) found that SVM and XGBoost achieved over 80% accuracy and AUC in predicting HCV outcomes. Their study revealed that these advanced models were the most effective for clinical applications. Similarly, Nandipati et al. (2020) showed that both Decision Trees and Random Forests performed well in multi-class and binary classification of HCV data. Chew and Khaw (2023) emphasized the importance of selecting and evaluating models based on dataset characteristics, reinforcing the need for tailored approaches in HCV management.

2.5 Regression Analysis

Regression models predict continuous outcomes, such as RNA levels in HCV patients. They help understand the quantitative relationships between clinical parameters and disease severity. Common regression techniques include Linear Regression, Ridge Regression, Lasso Regression, and Gradient Boosting Machines.

Linear Regression is a simple approach to modelling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and is easy to interpret. However, it may not capture complex relationships within the data. Ridge Regression is an extension of Linear Regression that includes a regularization term to prevent overfitting. This term penalizes large coefficients, improving the model's generalizability. Lasso Regression is similar to Ridge Regression but uses an L1 penalty, which can shrink some coefficients to zero, effectively performing feature selection. This makes it particularly useful when dealing with datasets with many features.

Gradient Boosting Machines (GBMs) are powerful ensemble methods that build models sequentially, with each new model correcting errors made by the previous ones. GBMs can capture complex non-linear relationships and have shown superior predictive performance in various applications, including HCV management. Ghazal et al. (2023) found that Gradient Boosting Machines offered superior predictive performance by capturing complex relationships within the data.

Regression models are used to predict continuous outcomes such as RNA levels in HCV patients. Accurate predictions can help monitor disease progression and evaluate treatment efficacy. Ghazal et al. (2023) utilized GBMs to predict RNA levels, achieving high accuracy by capturing complex data relationships. Safdari and Deghatipour (2023) optimized HCV disease prediction using a combination of regression techniques, highlighting the potential of tailored models for specific populations. These regression techniques provide a robust framework for understanding and predicting continuous outcomes in HCV management, contributing to better patient care and treatment strategies.

2.6 Clustering and Dimensionality Reduction

Machine learning relies on clustering and dimensionality reduction to analyse complex datasets like HCV. Clustering data points with similar features can assist identify patient subgroups. Dimensionality reduction reduces characteristics while maintaining crucial information, making datasets easier to visualise and analyse.

K-Means Clustering data into k feature-similar clusters is a common strategy. The algorithm clusters data points and updates centroids till convergence. This method works well for huge datasets. Chew and Khaw (2023) employed K-Means clustering to identify HCV patient subgroups that potentially benefit from focused treatment.

Hierarchical Clustering iteratively merges or splits clusters based on similarity to create a tree-like structure. Dendrograms help visualise this strategy for comprehending data point hierarchies. Hierarchical clustering can reveal data structure better than K-Means but is more computationally costly.

High-dimensional datasets require dimensionality reduction methods like PCA and t-SNE. PCA converts characteristics into orthogonal components that capture the most data variation, reducing dimensionality. This strategy is used for data simplification and visualisation. Visani et al. (2022) showed how PCA reduced HCV data complexity, making patterns easier to see and analyse. Non-linear dimensionality reduction method t-SNE preserves local structure and reveals complex patterns in high-dimensional data. It helps visualise clusters in two- or three-dimensional space and reveal data structure intuitively.

Clustering patients helps identify subgroups with diverse clinical features for personalised treatment. Complex datasets are easier to understand when dimensionality reduced. K-Means clustering helped Chew and Khaw (2023) identify HCV patient categories that potentially benefit from personalised treatment. These methods let you gain insights from massive, complex datasets.

2.7 Feature Selection and Importance

Building good machine learning models requires feature selection, especially with complex datasets like HCV. Identifying the most important factors that affect model predictions improves model performance and interpretability.

RFE, a popular feature selection method, repeatedly removes the least significant characteristics and builds models on the rest. This procedure repeats until the best features are found. RFE improves model performance by removing unnecessary features.

Feature significance ratings from Random Forests or Gradient Boosting Machines reveal which features affect predictions most. These scores assist identify the most important features for accurate model predictions. By employing feature relevance scores to identify significant predictors, Ribeiro et al. (2016) showed how feature selection improves model interpretability and performance. HCV outcome predictors must be identified to create accurate and trustworthy models. RFE and feature importance scores let researchers focus on liver function tests and patient demographics, which affect HCV development and treatment response.

Feature selection also improves model interpretability. Reducing features simplifies and clarifies models, making predictions easier for physicians to comprehend and trust. Integrating machine learning models into clinical practice requires transparency.

Chew and Khaw (2023) employed RFE to uncover essential features in an HCV dataset, finding that liver function and patient demographic variables were crucial for successful predictions. Their study shows that feature selection improves model performance and interpretability, improving patient care.

2.8 Model Evaluation and Validation

Evaluating and validating machine learning models is essential to ensure their effectiveness and reliability in predicting HCV outcomes. Different metrics are used to assess model performance based on the type of prediction task, whether classification or regression.

For classification tasks, metrics such as accuracy, precision, recall, and F1-Score are commonly used. Accuracy measures the proportion of correct predictions, precision indicates the proportion of true positive predictions among all positive predictions, recall (or sensitivity) measures the proportion of true positive predictions among all actual positives, and F1-Score is the harmonic mean of precision and recall, providing a balanced measure of performance.

For regression tasks, metrics like Mean Squared Error (MSE) and R-squared (R^2) are used. MSE measures the average squared difference between predicted and actual values, indicating the model's prediction error. R-squared measures the proportion of variance in the dependent variable explained by the independent variables, indicating the model's explanatory power.

Cross-validation is a critical technique for assessing model robustness and generalizability. It involves splitting the dataset into multiple subsets, training the model on some subsets, and validating it on the remaining subsets. This process is repeated multiple times, and the results are averaged to obtain a reliable performance estimate. Cross-validation helps ensure that the model performs well on unseen data, preventing overfitting.

Hyperparameter tuning is another important step in optimizing model performance. It involves selecting the best combination of hyperparameters that control the learning process of the model. Techniques like grid search and random search are commonly used for hyperparameter tuning, systematically evaluating different combinations to find the optimal settings.

Ensuring robustness and generalizability is crucial for deploying machine learning models in clinical settings. Models must perform consistently across different patient populations and data sources to be reliable. Chew and Khaw (2023) emphasized the importance of cross-validation and hyperparameter tuning in their study, demonstrating robust model performance across multiple subsets of the HCV dataset.

2.10 Research Gap

Clinical use of ML could revolutionise HCV care. Clinical diagnosis, disease progression prediction, and treatment personalisation can be improved by ML models. For instance, predictive analytics can identify patients at risk of serious liver problems for timely and focused treatment. ML can improve patient care by informing treatment strategies with data. By selecting patients who will respond well to specific medicines, predictive models can optimise antiviral therapy. ML algorithms also find patterns and connections in patient data that traditional analysis may miss, improving care and personalisation.

ML models must be validated extensively to be applicable to varied patient groups and therapeutic contexts. Models are tested on varied datasets to ensure robustness and dependability. To ensure clinical application, ML models should be validated on large, multi-centre datasets. Data scientists, clinicians, and healthcare professionals must work together to improve ML models and ensure clinical relevance. These professionals can uncover key clinical features, understand model predictions, and incorporate ML tools into clinical processes by working collaboratively.

ML models in real-time healthcare procedures can deliver meaningful information during patient visits. ML solutions must have user-friendly interfaces and smooth EHR integration to be practical in clinical practice. To integrate ML into healthcare, numerous hurdles must be overcome. Data privacy is difficult, yet encryption and anonymisation are necessary to secure patient data and meet regulatory requirements. Model interpretability is essential for clinician trust and ML tool uptake. Explainable AI models that make clear predictions can bridge the gap between complicated ML algorithms and clinical practice.

Integrating diverse data sources including health records, genetic data, and imaging improves. ML model prediction. Solid data integration frameworks and interoperability standards are needed for seamless data flow and analysis. Clinical professionals can trust explainable AI models' forecasts since they provide explicit, interpretable insights into prediction. Integrating ML into healthcare decision-making requires transparency. ML tool adoption requires healthcare practitioners' trust. Training on ML approaches, showing model correctness and reliability, and incorporating doctors in model development can assist build trust in these technologies.

Chapter 3: Methodology

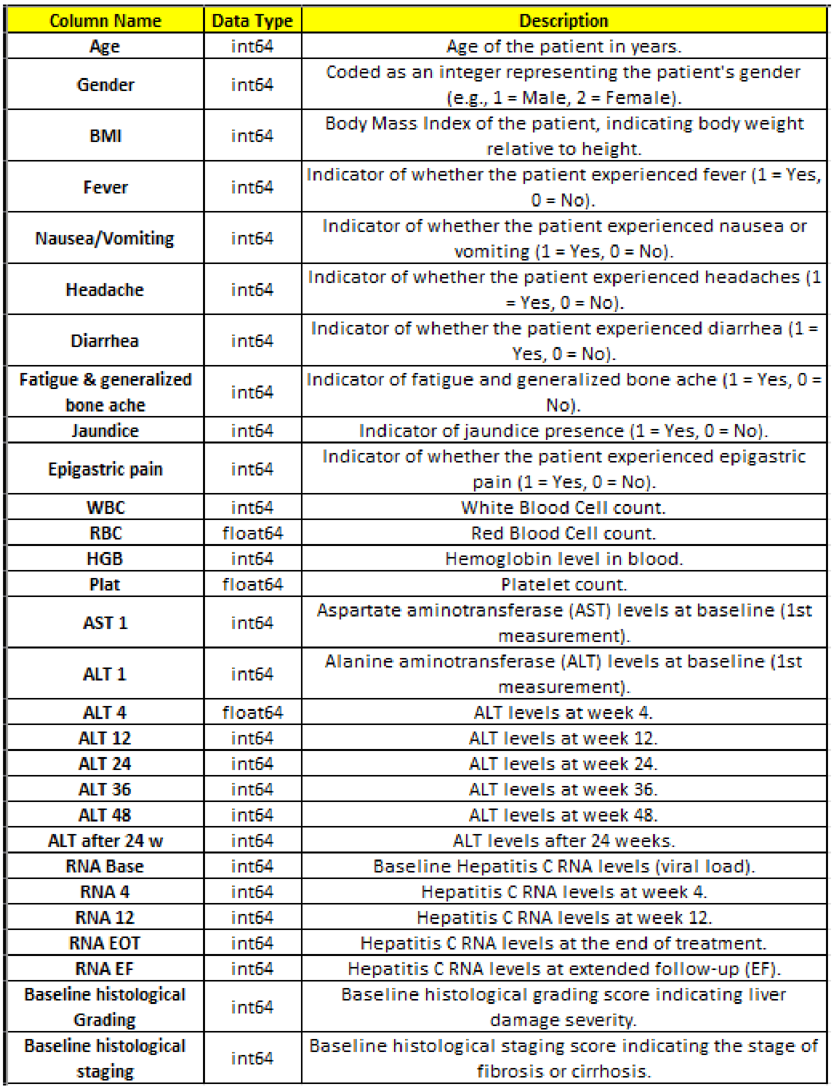
The main aim of this study is to examine the Hepatitis C Virus (HCV) dataset in order to comprehend the correlations among different clinical, demographic, and biochemical characteristics and the phases of liver disease advancement. The study seeks to utilise several machine learning techniques to create predictive models capable of accurately categorising the histological stage of liver disease in patients with HCV.

The methodology employed in this work adheres to a methodical strategy for analysing data and developing models. Firstly, the dataset was extensively examined to comprehend its organisation, and then data cleaning and preprocessing procedures were carried out to guarantee the integrity of the data. Feature engineering was utilised to generate additional variables that could augment the predictive capability of the models. Data visualisation tools were employed to reveal patterns and relationships in the data. Afterwards, the data underwent standardisation and the categorical variables were encoded, making it ready for machine learning. Several classification models, such as Logistic Regression, Decision Tree, K-Nearest Neighbours, Support Vector Machine, and XGBoost, were trained and assessed. Ultimately, the model's performance was optimised by hyperparameter tweaking, and advanced approaches such as neural networks and ensemble methods were utilised to enhance prediction accuracy.

3.1 Data Overview

3.1.1 Dataset Description:

This study will employ a dataset that includes comprehensive clinical, demographic, and biochemical data from patients who have been diagnosed with Hepatitis C Virus (HCV) and is available on Kaggle (Pires, 2019) . The dataset contains a range of variables, including patient age, gender, Body Mass Index (BMI), and liver function test findings, namely ALT (Alanine Aminotransferase) and AST (Aspartate Aminotransferase) values as shown in figure 2. Additionally, RNA levels were assessed at different stages of treatment. Furthermore, the information includes categorical data pertaining to symptoms and the initial histological staging of liver disease. This dataset is anticipated to offer a comprehensive foundation for analysing the advancement of liver disease in HCV patients and identifying crucial markers of disease staging.



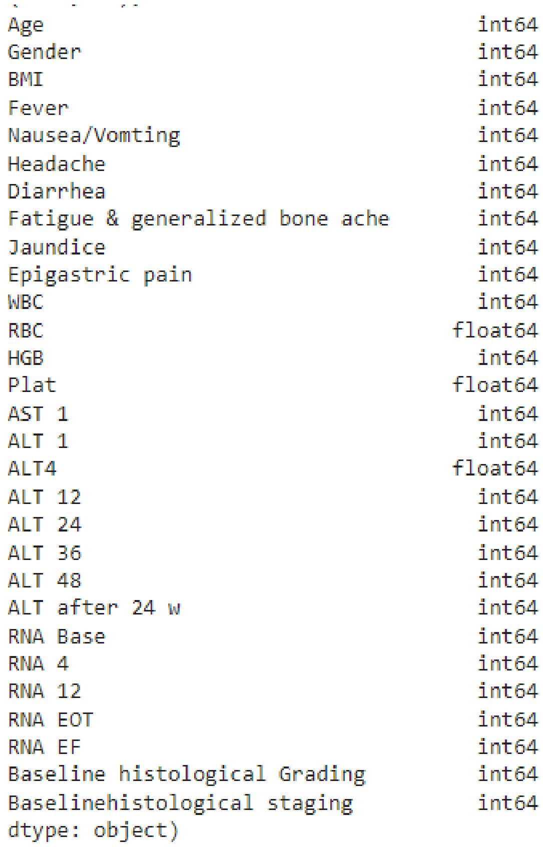
Column Name	Data Type	Description
Age	int64	Age of the patient in years.
Gender	int64	Coded as an integer representing the patient's gender (e.g., 1 = Male, 2 = Female).
BMI	int64	Body Mass Index of the patient, indicating body weight relative to height.
Fever	int64	Indicator of whether the patient experienced fever (1 = Yes, 0 = No).
Nausea/Vomiting	int64	Indicator of whether the patient experienced nausea or vomiting (1 = Yes, 0 = No).
Headache	int64	Indicator of whether the patient experienced headaches (1 = Yes, 0 = No).
Diarrhea	int64	Indicator of whether the patient experienced diarrhea (1 = Yes, 0 = No).
Fatigue & generalized bone ache	int64	Indicator of fatigue and generalized bone ache (1 = Yes, 0 = No).
Jaundice	int64	Indicator of jaundice presence (1 = Yes, 0 = No).
Epigastric pain	int64	Indicator of whether the patient experienced epigastric pain (1 = Yes, 0 = No).
WBC	int64	White Blood Cell count.
RBC	float64	Red Blood Cell count.
HGB	int64	Hemoglobin level in blood.
Plat	float64	Platelet count.
AST 1	int64	Aspartate aminotransferase (AST) levels at baseline (1st measurement).
ALT 1	int64	Alanine aminotransferase (ALT) levels at baseline (1st measurement).
ALT 4	float64	ALT levels at week 4.
ALT 12	int64	ALT levels at week 12.
ALT 24	int64	ALT levels at week 24.
ALT 36	int64	ALT levels at week 36.
ALT 48	int64	ALT levels at week 48.
ALT after 24 w	int64	ALT levels after 24 weeks.
RNA Base	int64	Baseline Hepatitis C RNA levels (viral load).
RNA 4	int64	Hepatitis C RNA levels at week 4.
RNA 12	int64	Hepatitis C RNA levels at week 12.
RNA EOT	int64	Hepatitis C RNA levels at the end of treatment.
RNA EF	int64	Hepatitis C RNA levels at extended follow-up (EF).
Baseline histological Grading	int64	Baseline histological grading score indicating liver damage severity.
Baseline histological staging	int64	Baseline histological staging score indicating the stage of fibrosis or cirrhosis.

Figure 2: Data Dictionary

"This study has received ethical approval, and the ethical approval letter is attached in the Appendix/included in the Appendix Material."

3.1.2 Initial Exploration:

We will initially explore the dataset to gain a comprehensive understanding of its structure and evaluate its quality. To begin, we will utilise tools such as .columns.tolist() to generate a list of all the characteristics contained inside the dataset, and .dtypes to analyse the data types of each individual column as shown in figure 3. It is crucial to perform this step in order to verify that the data types are suitable for analysis. To ascertain the shape of the dataset, we will utilise the .shape attribute, which will provide us with information regarding the number of records and features that are available for study. The distribution of numerical variables will be evaluated by computing basic summary statistics using the .describe() function. To examine the distinct values and their frequency in categorical features, we will utilise the .unique() and .value_counts() functions. In addition, we will utilise the .isnull().sum() function to identify the number of missing values, the .duplicated().sum() function to determine the count of duplicate entries, and the (data < 0).sum() expression to detect any negative values that could potentially suggest problems in data entry. These functions will aid in the identification of any issues that require attention throughout the data cleaning and preparation phases.



```
Age                                int64
Gender                             int64
BMI                                int64
Fever                              int64
Nausea/Vomting                     int64
Headache                           int64
Diarrhea                           int64
Fatigue & generalized bone ache    int64
Jaundice                           int64
Epigastric pain                    int64
WBC                                int64
RBC                              float64
HGB                                int64
Plat                             float64
AST 1                              int64
ALT 1                              int64
ALT4                             float64
ALT 12                             int64
ALT 24                             int64
ALT 36                             int64
ALT 48                             int64
ALT after 24 w                     int64
RNA Base                           int64
RNA 4                              int64
RNA 12                             int64
RNA EOT                            int64
RNA EF                             int64
Baseline histological Grading      int64
Baselinehistological staging       int64
dtype: object)
```

Figure 3: Column Details

3.2 Data Cleaning and Preprocessing

In this study, managing data quality is essential for accurate analysis and model performance. First, we will handle missing data by identifying gaps using the .isnull().sum() function and applying appropriate strategies. For minor missing data, we may use simple imputation techniques such as filling with the mean, median, or mode. For more significant gaps, advanced methods like K-Nearest Neighbours (KNN) imputation or column removal may be employed to ensure data integrity.

Duplicate records will be identified using .duplicated() and removed with .drop_duplicates(), while ensuring no important repeated measurements are lost. Negative values, which may result from data entry errors, will be detected using conditional expressions and treated accordingly, either by correction or by handling as missing data.

Outliers will be managed using the Interquartile Range (IQR) method to detect extreme values. Visualization techniques such as boxplots will help identify outliers, allowing for further evaluation to determine whether they represent valid data points or anomalies that need to be addressed.

Feature engineering will play a crucial role in improving model performance. We will create new features by generating interaction terms, ratios, and differences between key variables. Binning continuous variables, such as age and BMI, will also enhance interpretability. Finally, aggregated features and log transformations will help stabilize variance and capture key trends, making the data more suitable for machine learning models.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a key step in understanding the patterns, relationships, and distributions within the dataset. It guides the feature selection and model development processes by uncovering insights from the data.

Histograms will be used to explore the distribution of numerical variables like ALT and AST levels. These visualizations will help identify trends such as normal distribution, skewness, or outliers, which can reveal whether most patients have normal or elevated enzyme levels. This can point out unusual patterns that may warrant further investigation.

Bar charts will display the distribution of categorical variables, such as Gender, Age_Group, and BMI_Category, allowing us to examine the demographic spread within the patient population. This ensures that the dataset is balanced and free from significant biases, which could affect the accuracy of the models.

Boxplots and violin plots will compare the distribution of numerical features across different stages of the target variable, Baseline histological staging. These tools will highlight variability in key features, such as ALT and RNA levels, across disease stages, providing insight into their potential as predictive features.

Finally, correlation analysis will be conducted to identify relationships between numerical variables using Pearson correlation coefficients. A heatmap will highlight strong correlations, helping to detect multicollinearity and simplify the model by identifying redundant variables or key feature interactions.

3.4 Data Standardization and Encoding

3.4.1 Standardization of Numerical Features:

Standardising numerical features with the StandardScaler function ensures that they all contribute equally to model training. Standardisation gives numerical data a mean of zero and a standard deviation of one, which is crucial for Logistic Regression, SVM, and K-Nearest Neighbours. Standardising ALT, BMI, and RNA levels prevents larger variables from dominating the model because these techniques are sensitive to input data magnitude. Gradient-based optimisation methods converge faster since the model focusses on variable relationships rather than size.

Encoding categorical features will help machine learning models use them. We will use OneHotEncoder to convert nominal categorical variables like Gender, Age_Group, and BMI_Category into binary columns without assuming ordinal relationships. LabelEncoder will facilitate classification of ordinal data like Baseline histological staging by converting categories into numerical labels. This approach accurately integrates categorical data into Decision Trees and XGBoost without bias.

3.5 Model Selection and Training

3.5.1 Selection of Models

In this study, we will employ a diverse set of machine learning models to predict the histological staging of liver disease in HCV patients. The selection of models is based on their suitability for handling different aspects of our dataset and the nature of the classification problem.

Logistic Regression: Logistic Regression is a linear model that calculates the likelihood of a given input point belonging to a specific class as shown in figure 4. Its interpretability makes it particularly beneficial for comprehending the connection between features and the target variable (Brownlee, 2016). This model will enable us to ascertain the crucial parameters that influence the staging of liver disease, serving as a reference point for comparison with more intricate models.

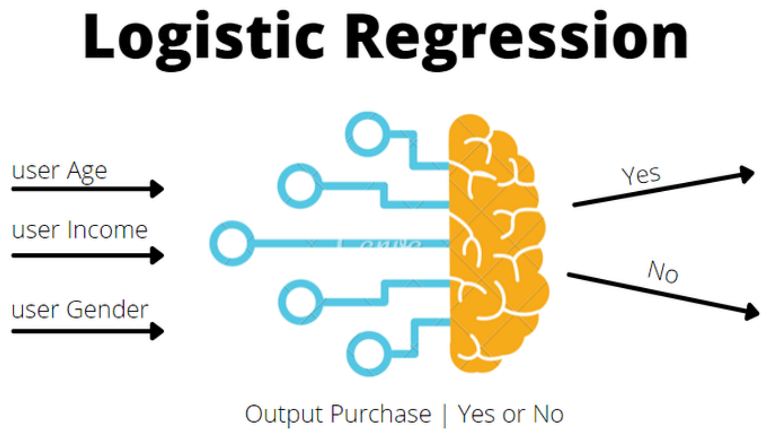


Figure 4: Logistic Regression. (Jainvidip, 2024)

Decision Tree Classifier: Decision Trees are nonlinear models that partition the data depending on feature values, resulting in a hierarchical structure where each node corresponds to a choice based on a feature, and each leaf node reflects a specific conclusion as shown in figure 5. This model is beneficial for our problem since it can effectively manage both numerical and categorical data. It provides clear and interpretable decision rules that specifically highlight the impact of certain parameters on the stage of liver disease (Wang et al., 2020).

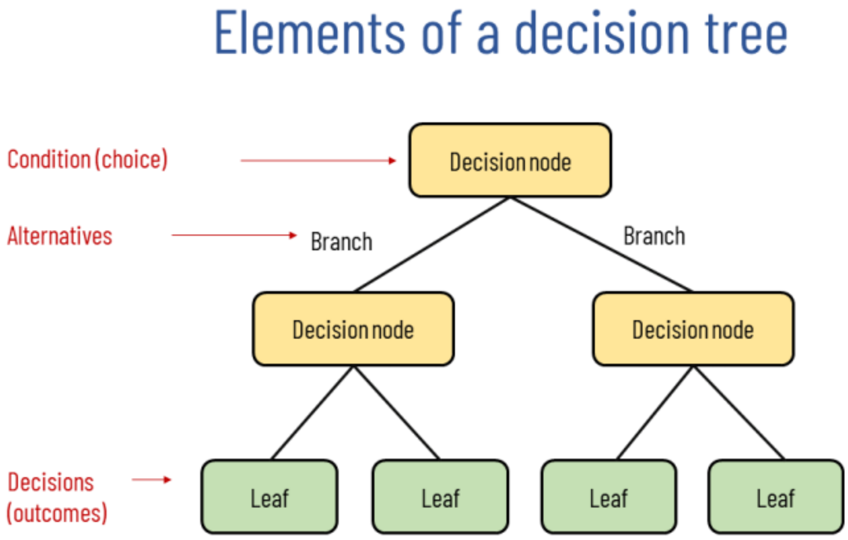


Figure 5: Decision Tree (Kosarenko, 2021)

K-Nearest Neighbours (KNN): KNN is an instance-based learning algorithm that classifies a data point based on the majority class of its nearest neighbours as shown in figure 6. It is effective in capturing non-linear decision boundaries, making it useful for our problem where the relationship between patient features and disease stages may not be linear. KNN can particularly help in recognizing patterns among patients with similar clinical profiles (Sharma, 2021)

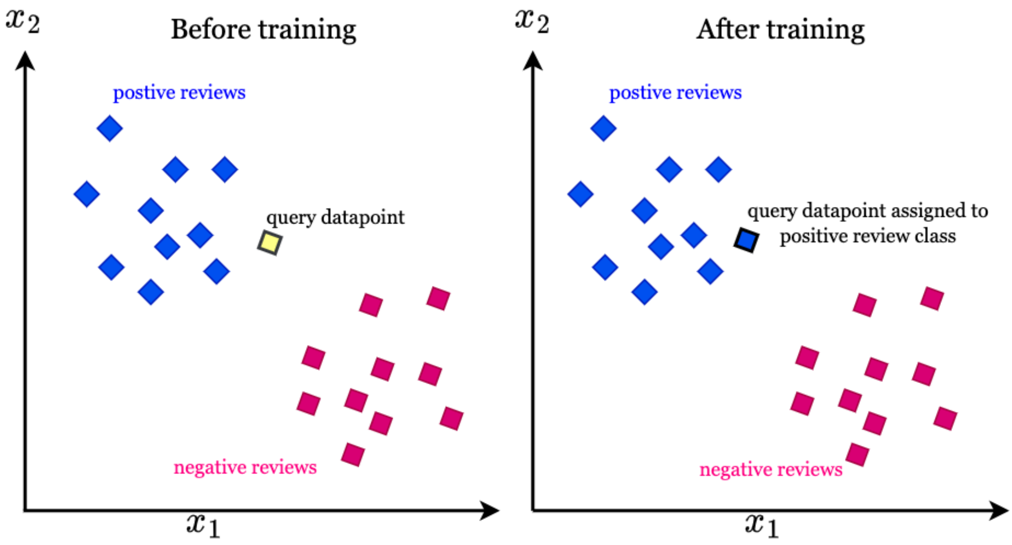


Figure 6: K-Nearest Neighbour. (H, 2023)

Support Vector Machine (SVM): SVM works by finding the optimal hyperplane that separates the classes in the feature space. It is highly effective in high-dimensional spaces and can handle non-linear relationships through kernel functions as shown in figure 7 (Sasidharan, 2021). Given the complexity of our dataset, SVM is a good fit for capturing intricate patterns that may not be apparent with simpler models.

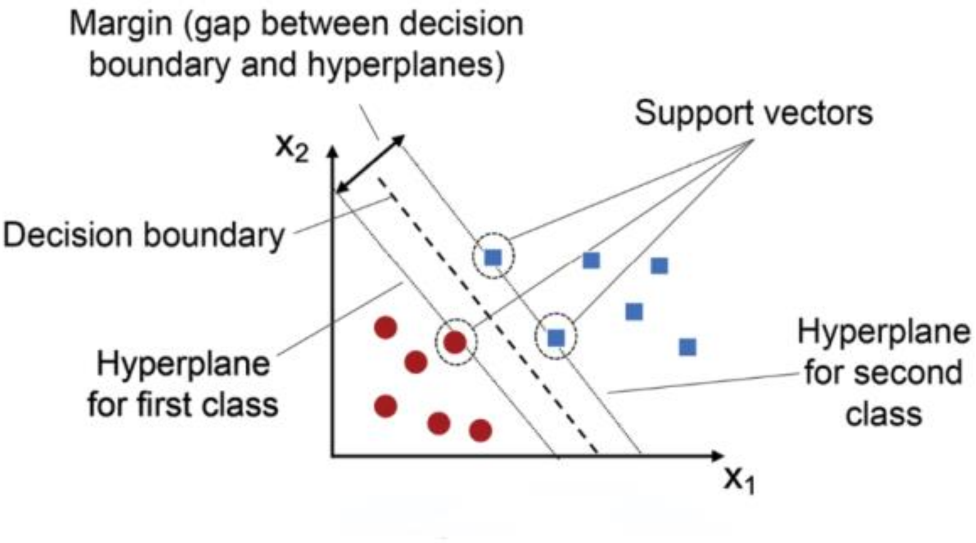

Margin (gap between decision boundary and hyperplanes)
Support vectors
x2
Decision boundary
Hyperplane for first class
Hyperplane for second class
x1

Figure 7: Support Vector Machine. (Kumar, 2022)

XGBoost: XGBoost is a powerful ensemble learning technique based on decision trees, known for its efficiency and performance. It uses a boosting approach to combine multiple weak learners into a strong model, making it highly effective in handling complex interactions between features as shown in figure 8. XGBoost is particularly useful for our dataset due to its ability to manage missing data and achieve high predictive accuracy (kharkar, 2023).

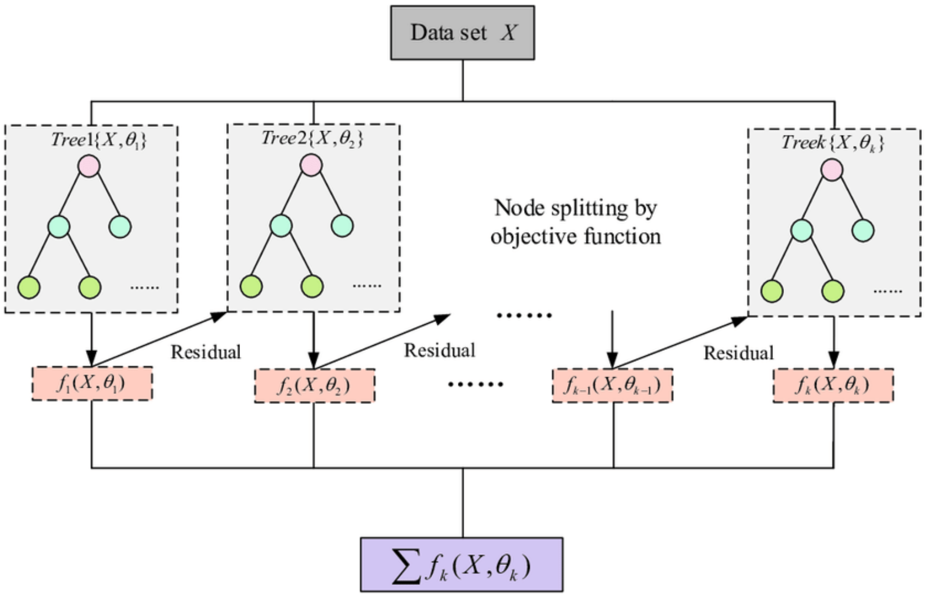


Figure 8: XGBoost. (Guo et al., 2020)

3.6 Advanced Model Training

3.6.1 Neural Network Development:

We will construct a neural network model to comprehend the intricate, non-linear connections between the clinical and biochemical characteristics in our dataset. The neural network will be constructed using the Sequential API in TensorFlow or Keras, commencing with an input layer that corresponds to the number of features in our dataset. Subsequently, there will be a series of concealed layers, each comprising a specific quantity of neurones, employing ReLU activation functions to incorporate non-linearity. The ultimate layer will consist of a softmax output layer designed for multi-class classification, enabling the model to generate probabilities for each potential stage of liver disease. The network will undergo training using the backpropagation algorithm, employing the categorical cross-entropy loss function and an optimiser such as Adam, which dynamically modifies the learning rate. We will assess the success of the training by measuring metrics such as accuracy and loss. Additionally, we will employ early stopping techniques to avoid overfitting. The objective of this neural network technique is to exploit the model's capacity to acquire knowledge from data with a large number of dimensions, which could result in better prediction performance compared to conventional models.

3.6.2 Hyperparameter Tuning for Neural Networks:

In order to enhance the efficiency of our neural network, we will do hyperparameter tuning utilising methods such as grid search or random search. The crucial hyperparameters that will be adjusted are the quantity of hidden layers, the quantity of neurones per layer, the learning rate, and the dropout rate. By manipulating the number of layers and neurones, we may determine the ideal network structure that achieves a balance between model complexity and predictive accuracy (Azam et al., 2024). The dropout rate, which mitigates overfitting by randomly deactivating a portion of input units during training, will also be fine-tuned to identify the optimal value for regularisation. The learning rate, responsible for the speed at which the model adjusts to new input, will be optimised to ensure effective and consistent training. The optimal hyperparameters will be chosen by evaluating the performance of different combinations through cross-validation. This ensures that the final model is not only precise but also has the ability to generalise effectively to new, unseen data.

3.6.3 Ensemble Methods:

Ensemble approaches amalgamate the forecasts of numerous models to enhance overall performance and resilience. This study will utilise two robust ensemble techniques: Voting Classifier and Stacking Classifier. These strategies seek to exploit the advantages of individual models while mitigating their limitations, resulting in more precise and dependable forecasts.

Voting Classifier:

The Voting Classifier aggregates the predictions of several different models by majority voting (for classification tasks). We will implement a soft voting approach, where the predicted probabilities of each model (e.g., Logistic Regression, Decision Tree, KNN, SVM, and XGBoost) are averaged, and the class with the highest probability is selected as the final prediction as shown in figure 9 (Scikit-learn.org, 2012). This method helps in combining models with different strengths, potentially improving the classification performance across all stages of liver disease.

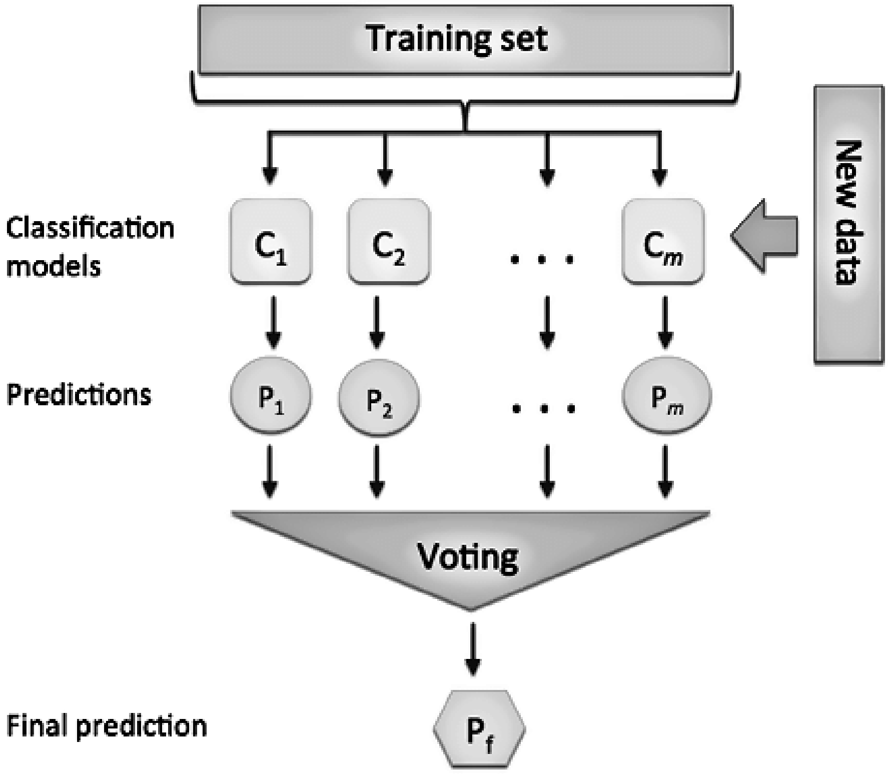


Figure 9: Voting Classifier (Patil and Patil, 2018)

Stacking Classifier:

The Stacking Classifier is an ensemble method that utilises the predictions of many base models as inputs to a meta-model, which then produces the final prediction. This study will employ many models, including Logistic Regression, Decision Tree, KNN, SVM, and XGBoost, as the basis models as shown in figure 10. The meta-model will be a Logistic Regression model. This methodology enables the stacking classifier to effectively leverage the unique capabilities of each base model, typically leading to greater performance in comparison to individual models or a basic voting ensemble (Brownlee, 2020).

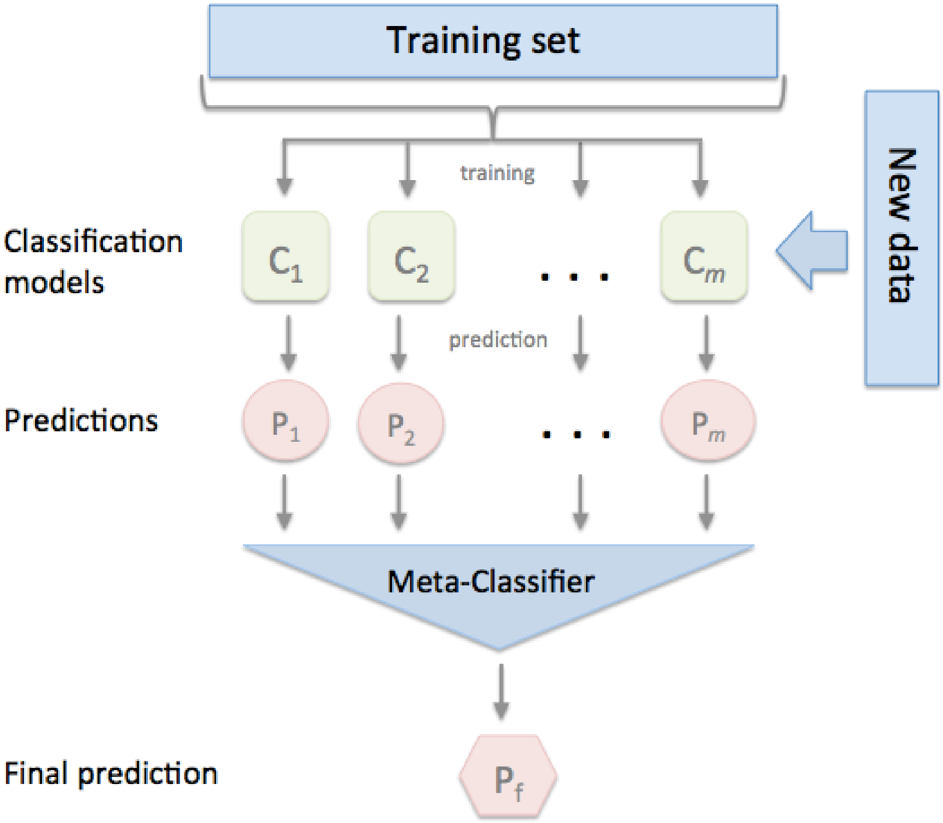


Figure 10: Stacking Classifier (rasbt.github.io, n.d.)

3.7 Model Evaluation and Comparison

3.7.1 Evaluation Metrics

To assess the performance of our models, we will employ several key evaluation metrics that provide a comprehensive understanding of how well each model is performing in classifying the histological staging of liver disease in HCV patients. These metrics include:

Accuracy: The proportion of correctly classified instances out of the total instances, providing a general measure of model performance (Bonnet, 2023).

Precision and Recall: Precision measures the accuracy of positive predictions, while recall (sensitivity) assesses how well the model captures all relevant positive instances. These metrics are crucial for understanding the model's behaviour in detecting specific classes (Bonnet, 2023).

F1-Score: The harmonic mean of precision and recall, offering a balance between the two, particularly useful when dealing with imbalanced classes (Bonnet, 2023).

ROC-AUC: The Area Under the Receiver Operating Characteristic Curve evaluates the model's ability to distinguish between classes across different thresholds. A higher AUC indicates better model performance in differentiating between disease stages (Bonnet, 2023).

Confusion Matrix: A detailed breakdown of true positives, true negatives, false positives, and false negatives, offering insights into the model’s specific strengths and weaknesses across different classes.

These metrics together provide a well-rounded evaluation of model performance, highlighting not just overall accuracy but also how well the models perform in correctly identifying the various stages of liver disease.

3.8 Hyperparameter Tuning

In order to enhance the efficiency of our models, we will utilise Grid Search, a methodical technique for fine-tuning hyperparameters. Grid Search is a method that entails creating a grid of potential hyperparameter values for each model and systematically exploring all possible combinations to determine the set of parameters that produces the highest cross-validation performance. In Logistic Regression, parameters such as the regularisation strength (C) and the penalty type (l1 or l2) will be adjusted. The hyperparameters of Decision Trees, such as max_depth, min_samples_split, and min_samples_leaf, will be examined. When using SVM, it is necessary to optimise parameters such as the C value, gamma, and the kernel selection. The main parameters for fine-tuning in XGBoost include n_estimators, max_depth, and learning_rate. The Grid Search procedure will be performed using cross-validation to guarantee that the outcomes are applicable to a wider range of data and not too tailored to the training data. This will enable us to identify the most effective combination of hyperparameters that improve the performance of the model.

3.9. Summary

This study utilised a thorough and methodical approach to examine and simulate the advancement of liver disease in individuals infected with Hepatitis C Virus (HCV). We began by thoroughly examining and purifying the dataset, resolving concerns such as incomplete data, identical entries, and exceptional values to guarantee the integrity of the data. Subsequently, we devised additional characteristics to augment the prognostic capability of our models and employed a variety of machine learning methods, such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), XGBoost, and Neural Networks. Every model underwent thorough evaluation using many metrics, and hyperparameter tuning was performed to enhance performance. This process resulted in the identification of the most optimal models for predicting liver disease stage.

The results of this study have important consequences for clinical practice, especially in the identification and classification of liver disease in individuals with HCV. Our models can help healthcare workers make better judgements about patient management and treatment methods by finding important factors that predict illness development. Utilising sophisticated machine learning techniques, such as ensemble methods and neural networks, showcases the ability to enhance diagnostic precision and customise patient care. This study adds to the expanding domain of data-driven healthcare by offering valuable instruments to improve patient outcomes in the realm of chronic liver disease.

Chapter 4: Results

4.1 Data Cleaning and Preprocessing

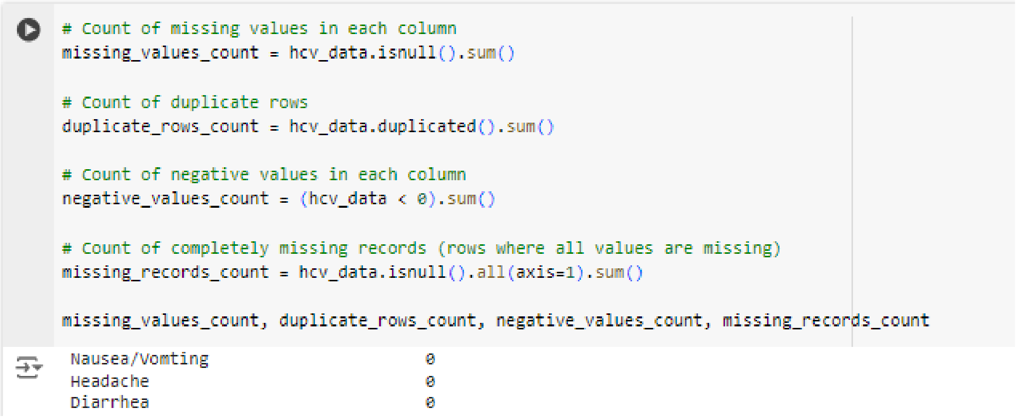
4.1.1 Summary and Descriptive Statistics

The dataset utilised for this analysis comprises 1,385 entries with 29 columns, encompassing a combination of numerical and category variables. The dataset contains comprehensive clinical, demographic, and biochemical information, encompassing age, gender, body mass index (BMI), multiple liver function test outcomes, and RNA levels evaluated at various time intervals.

The summary statistics indicate that the average age of patients is roughly 46 years, with a standard deviation of 8.78 years. Additionally, the average BMI is around 28.6, which classifies it inside the overweight range. The majority of the categorical variables, such as symptoms like fever, nausea, and jaundice, are binary in nature, with values that indicate either the presence or absence of these diseases. The descriptive statistics for continuous variables, such as ALT and AST values, exhibit significant variability, as anticipated in a clinical dataset.

4.1.2 Data Integrity Checks

The dataset was thoroughly checked for data integrity issues as shown in figure 11. Specifically:



```
# Count of missing values in each column
missing_values_count = hcv_data.isnull().sum()

# Count of duplicate rows
duplicate_rows_count = hcv_data.duplicated().sum()

# Count of negative values in each column
negative_values_count = (hcv_data < 0).sum()

# Count of completely missing records (rows where all values are missing)
missing_records_count = hcv_data.isnull().all(axis=1).sum()

missing_values_count, duplicate_rows_count, negative_values_count, missing_records_count
```

```
Nausea/Vomting                    0
Headache                          0
Diarrhea                          0
```

Figure 11 :Data Integrity Check

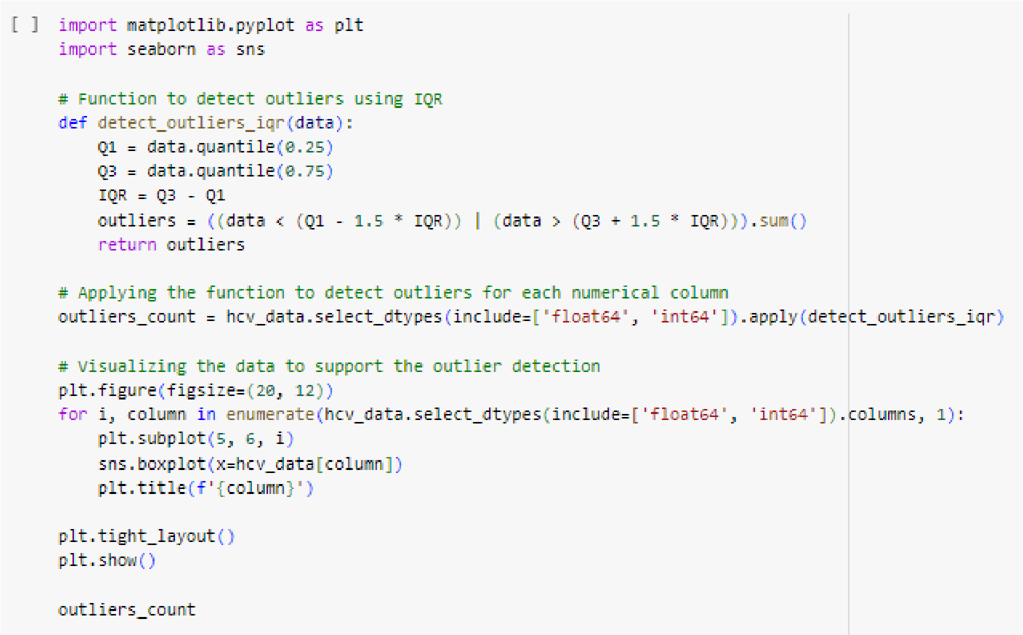
Missing Values: There were no missing values found in any of the columns, ensuring that the dataset was complete and required no imputation or removal of missing records.

Duplicates: The dataset was also free of any duplicate rows, ensuring that each record represented a unique patient or observation.

Negative Values: No negative values were found across the columns, which is crucial since many of the clinical measurements cannot logically be negative.

4.1.3 Outlier Detection

The Interquartile Range (IQR) technique was utilised to identify outliers. The analysis detected outliers in some columns, including ALT after 24 weeks and RNA 12. In the former, three outliers were identified, while in the latter, one outlier was observed. The presence of these outliers was verified through visual examination of boxplots, which depicted the distribution of values across all numerical columns as shown in figure 12(a) and 12(b).



```
[ ]  import matplotlib.pyplot as plt
     import seaborn as sns

     # Function to detect outliers using IQR
     def detect_outliers_iqr(data):
         Q1 = data.quantile(0.25)
         Q3 = data.quantile(0.75)
         IQR = Q3 - Q1
         outliers = ((data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR))).sum()
         return outliers

     # Applying the function to detect outliers for each numerical column
     outliers_count = hcv_data.select_dtypes(include=['float64', 'int64']).apply(detect_outliers_iqr)

     # Visualizing the data to support the outlier detection
     plt.figure(figsize=(20, 12))
     for i, column in enumerate(hcv_data.select_dtypes(include=['float64', 'int64']).columns, 1):
         plt.subplot(5, 6, i)
         sns.boxplot(x=hcv_data[column])
         plt.title(f'{column}')

     plt.tight_layout()
     plt.show()

     outliers_count
```

Figure 12 (a): Outlier Detection

Outliers are frequently observed in various biochemical tests, as extreme readings may suggest severe cases or measurement problems. The decision to retain or eliminate these outliers was determined by considering their possible clinical importance and their influence on the overall analysis.

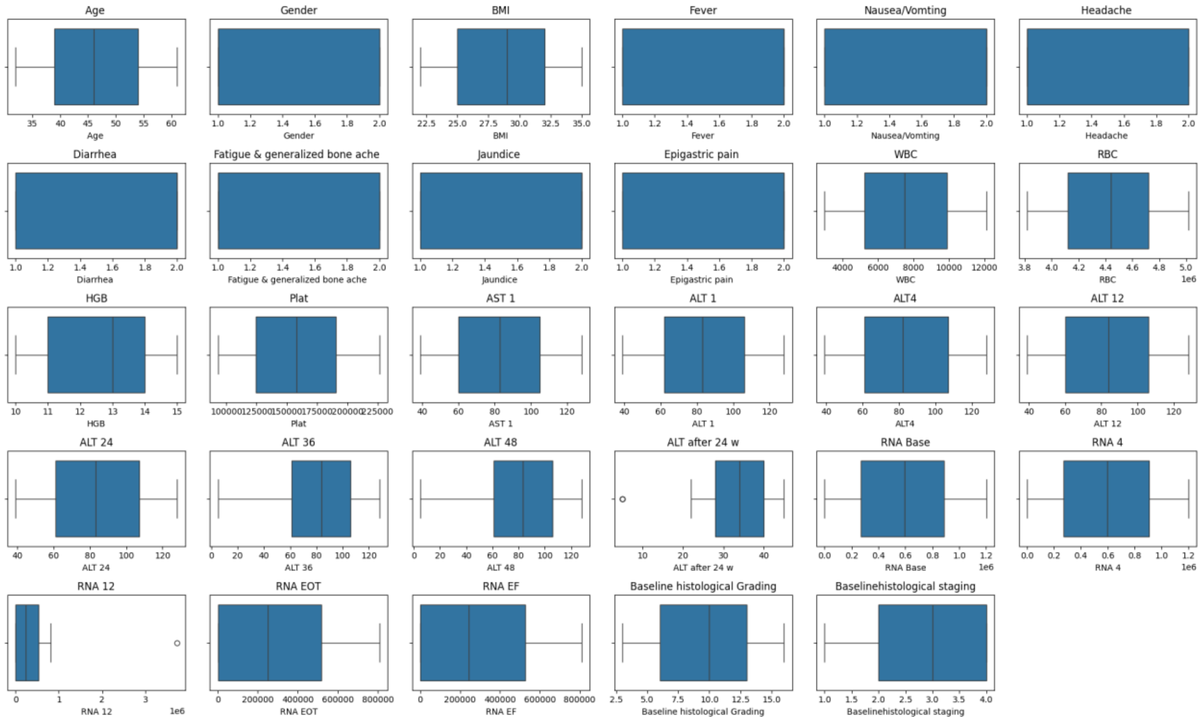
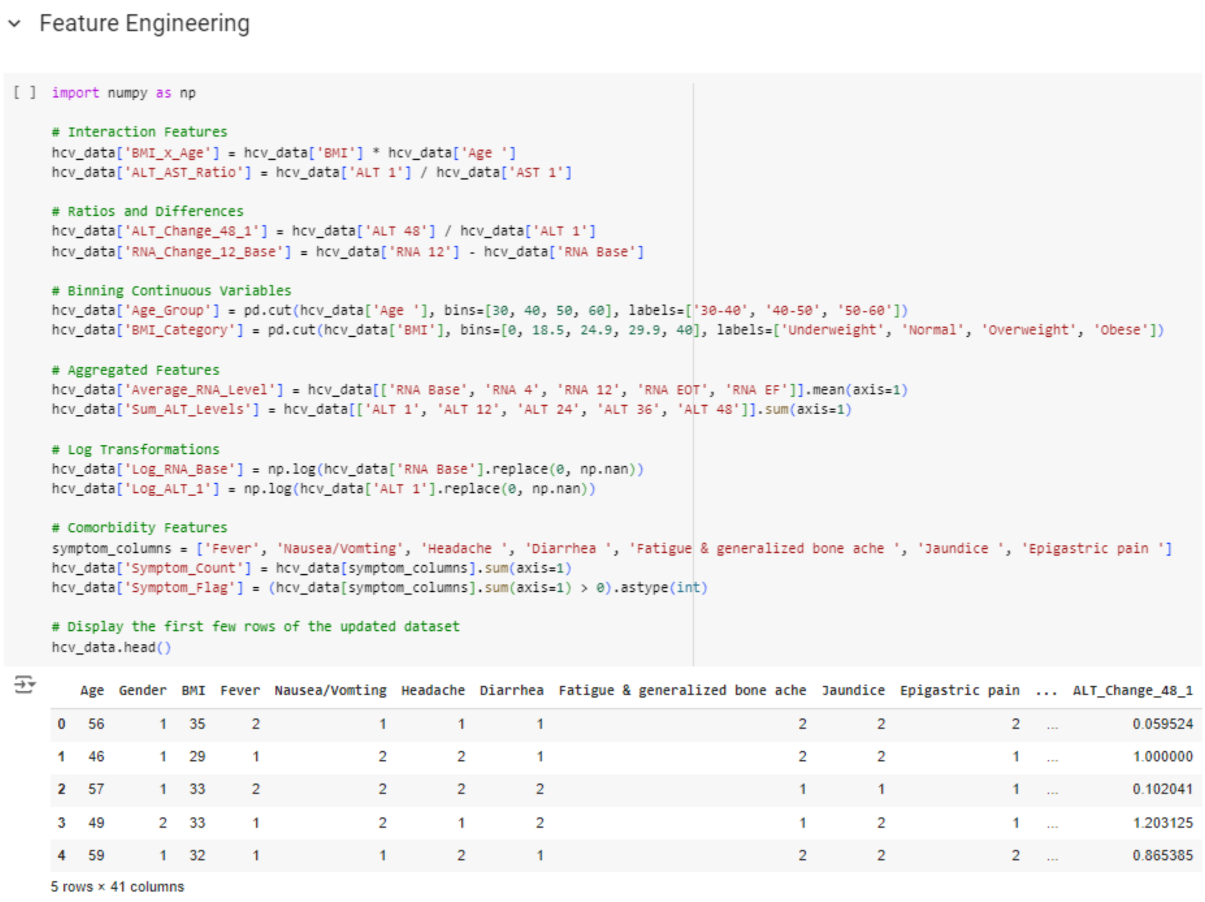


Figure 12 (b) : Outliers Detection using Boxplot

4.1.4 Feature Engineering

Several new features were created to enhance the dataset's predictive power as shown in figure 13:



Feature Engineering

```
[ ] import numpy as np

    # Interaction Features
    hcv_data['BMI_x_Age'] = hcv_data['BMI'] * hcv_data['Age ']
    hcv_data['ALT_AST_Ratio'] = hcv_data['ALT 1'] / hcv_data['AST 1']

    # Ratios and Differences
    hcv_data['ALT_Change_48_1'] = hcv_data['ALT 48'] / hcv_data['ALT 1']
    hcv_data['RNA_Change_12_Base'] = hcv_data['RNA 12'] - hcv_data['RNA Base']

    # Binning Continuous Variables
    hcv_data['Age_Group'] = pd.cut(hcv_data['Age '], bins=[30, 40, 50, 60], labels=['30-40', '40-50', '50-60'])
    hcv_data['BMI_Category'] = pd.cut(hcv_data['BMI'], bins=[0, 18.5, 24.9, 29.9, 40], labels=['Underweight', 'Normal', 'Overweight', 'Obese'])

    # Aggregated Features
    hcv_data['Average_RNA_Level'] = hcv_data[['RNA Base', 'RNA 4', 'RNA 12', 'RNA EOT', 'RNA EF']].mean(axis=1)
    hcv_data['Sum_ALT_Levels'] = hcv_data[['ALT 1', 'ALT 12', 'ALT 24', 'ALT 36', 'ALT 48']].sum(axis=1)

    # Log Transformations
    hcv_data['Log_RNA_Base'] = np.log(hcv_data['RNA Base'].replace(0, np.nan))
    hcv_data['Log_ALT_1'] = np.log(hcv_data['ALT 1'].replace(0, np.nan))

    # Comorbidity Features
    symptom_columns = ['Fever', 'Nausea/Vomting', 'Headache ', 'Diarrhea ', 'Fatigue & generalized bone ache ', 'Jaundice ', 'Epigastric pain ']
    hcv_data['Symptom_Count'] = hcv_data[symptom_columns].sum(axis=1)
    hcv_data['Symptom_Flag'] = (hcv_data[symptom_columns].sum(axis=1) > 0).astype(int)

    # Display the first few rows of the updated dataset
    hcv_data.head()
```

	Age	Gender	BMI	Fever	Nausea/Vomting	Headache	Diarrhea	Fatigue & generalized bone ache	Jaundice	Epigastric pain	...	ALT_Change_48_1
0	56	1	35	2	1	1	1	2	2	2	...	0.059524
1	46	1	29	1	2	2	1	2	2	1	...	1.000000
2	57	1	33	2	2	2	2	1	1	1	...	0.102041
3	49	2	33	1	2	1	2	1	2	1	...	1.203125
4	59	1	32	1	1	2	1	2	2	2	...	0.865385

5 rows × 41 columns

Figure 13: Feature Engineering

Interaction Features: The BMI_x_Age interaction term was created by multiplying BMI and age, capturing the combined effect of these two factors on liver health. Additionally, the ALT_AST_Ratio was calculated by dividing ALT by AST levels, a common clinical indicator of liver function.

Ratios and Differences: Features like ALT_Change_48_1 (the ratio of ALT levels at week 48 to week 1) and RNA_Change_12_Base (difference between RNA levels at week 12 and baseline) were generated to reflect changes over time.

Age and BMI were binned into categorical groups, such as Age_Group and BMI_Category, to simplify these variables and capture non-linear effects.

New features like Average_RNA_Level and Sum_ALT_Levels were computed to aggregate measurements across time points, providing a more comprehensive picture of a patient's condition.

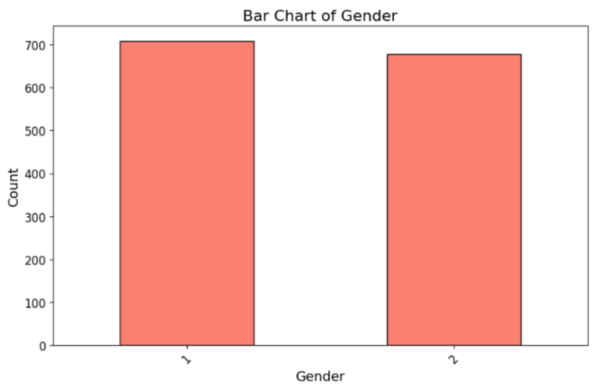
To address skewness in RNA and ALT levels, log transformations were applied to create Log_RNA_Base and Log_ALT_1, stabilizing variance and making these variables more suitable for modelling.

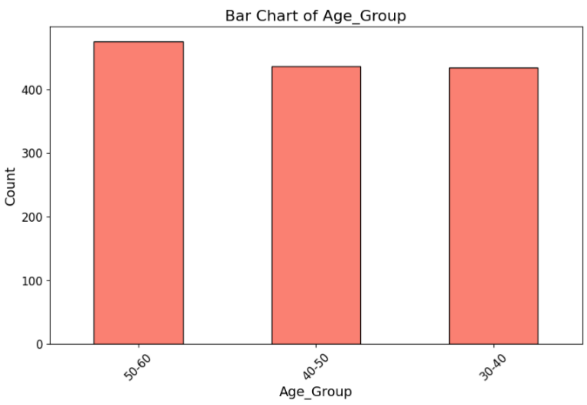
4.2 Exploratory Data Analysis

4.2.1 Univariate Analysis

The dataset has a symmetrical gender distribution, with 707 records classified as male and 678 as female. Ensuring this equilibrium is vital for reducing gender prejudice during the analysis process. The age distribution is rather balanced, with 475 patients between the ages of 50 and 60, 436 between the ages of 40 and 50, and 434 between the ages of 30 and 40. This distribution guarantees that every age category is adequately included, enabling the model to effectively extrapolate across various age brackets. Regarding BMI categories, the data reveals a greater incidence of obesity, with 611 individuals categorised as obese, 469 as overweight, and 305 as having a normal weight. It is worth mentioning that there are no individuals classified as underweight, suggesting a bias towards higher levels of BMI. This observation is particularly important considering the connection between BMI and liver health as shown in figure 14.

The summary statistics for numerical features offer further insights into the dataset. The mean age is approximately 46.3 years, with a mean BMI of 28.6, indicating that the majority of patients are classified as overweight. Liver function tests, specifically the levels of ALT, exhibit significant fluctuation, with average values approximately ranging from 83 to 84 and standard deviations close to 26 to 27. The diversity observed in this context represents the spectrum of liver enzyme levels among the patients, which is crucial for comprehending the advancement of liver disease. The RNA levels, with a mean value close to 590,000 and a substantial standard deviation, indicate notable variations in viral load among individuals, which may impact disease outcomes.





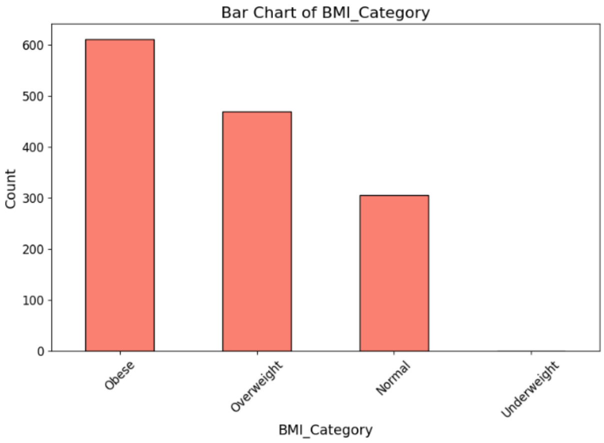


Figure 14: Bar chart for Gender. Age_Group and BMI_Category

The dataset's visualisations, such as histograms as shown in figure 15, for numerical features and bar charts for categorical features, offer a lucid comprehension of data distributions. These visualisations aid in detecting any asymmetry, extreme values, or disparities, which are important factors to address during preprocessing prior to model training.

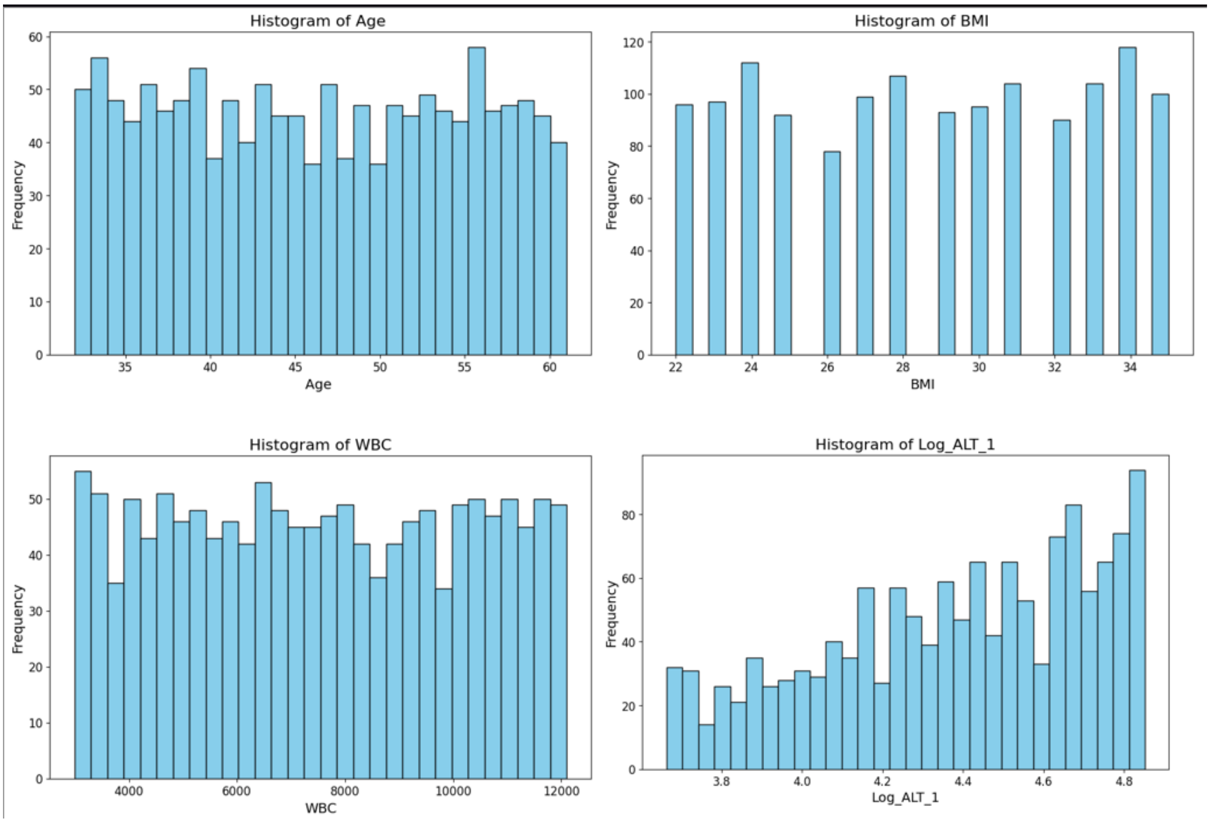


Figure 15: Histogram for Age, BMI, WBC and Log_ALT_1

4.2.2 Bivariate analysis

The boxplot and violin plot analyses offer useful insights into the variation of many clinical characteristics across several levels of baseline histological staging. The Symptom Count exhibits a steady median range of approximately 10 or 11 throughout all stages, however there are occasional outliers that suggest some levels of variability, notably in the intermediate stages of the illness. The interaction term BMI_x_Age shows a little decline in median values as the disease advances, indicating a decrease in the combined influence of BMI and age in later stages. However, there is still a significant amount of variability, particularly in earlier stages as shown in figure 16.

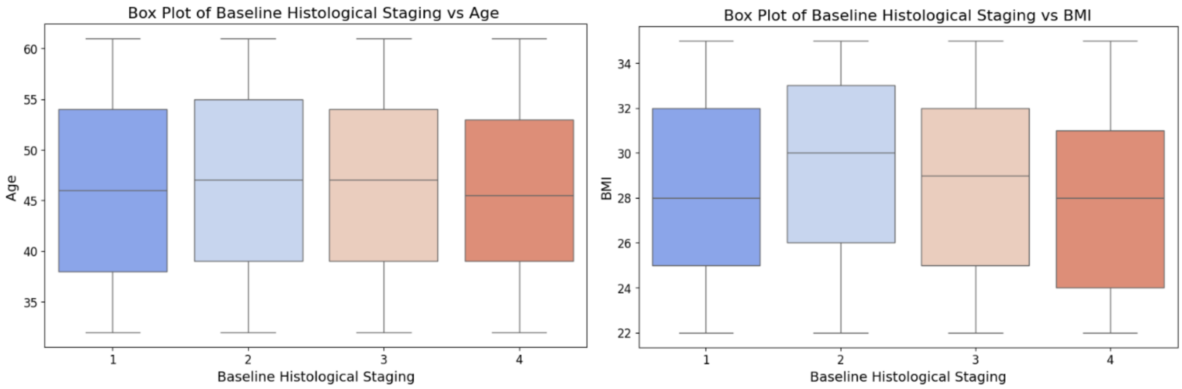


Figure 16: Boxplot Results for Baseline Histological Staging vs AGE and BMI

The Baseline Histological Grading shows a gradual drop in median values as the disease progresses, with less variation in later stages. This suggests that histological grading becomes more consistent as the disease advances. Ribonucleic acid (RNA) The baseline levels exhibit a reasonably consistent median throughout all phases, displaying a consistent albeit minor degree of variability. This indicates that the viral load remains essentially constant regardless of the stage of the disease. The median values of age and BMI remain stable across stages, with a little reduction in BMI observed in advanced stages. This emphasises the consistent levels of BMI among individuals with more severe disease. Ultimately, symptoms such as fever and nausea exhibit consistent patterns throughout the several stages, suggesting that these symptoms remain relatively stable as the disease advances.

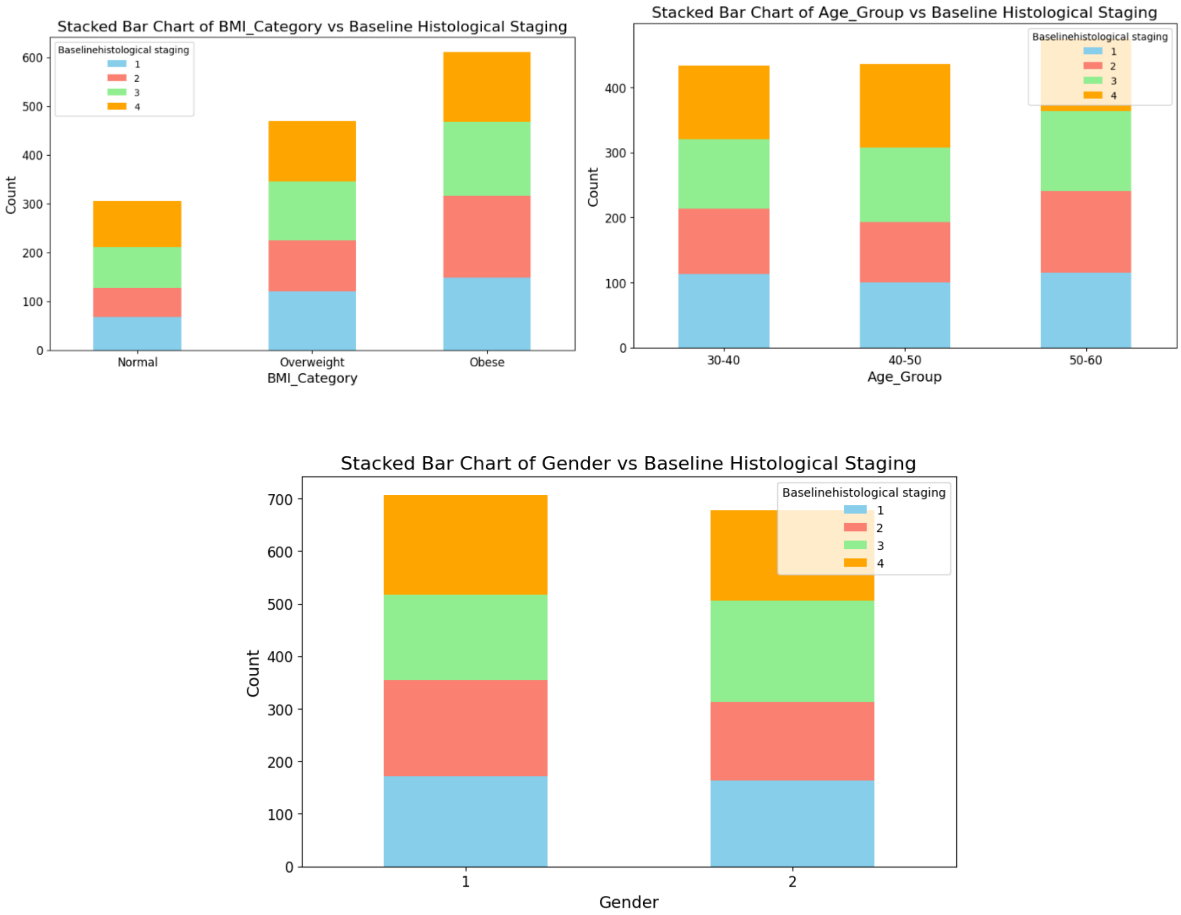


Figure 17: Stacked Chart

The stacked bar charts developed for the categorical variables against the Baseline Histological Staging display meaningful patterns throughout the dataset. The 'Gender' chart shows a very equal distribution across histological stages for both males and females, with no substantial variance between the two genders. The 'Age_Group' table indicates that the 50-60 age group has a higher proportion of individuals with histological stages 3 and 4. This suggests a possible association between older age and more advanced disease stage. The examination of 'BMI_Category' reveals a conspicuous pattern wherein higher BMI categories, specifically 'Obese,' exhibit a larger proportion of higher histological stages. This suggests a potential association between obesity and the advancement of disease. The 'Symptom_Flag' chart, which indicates the presence or absence of symptoms, shows a consistent distribution throughout all stages, suggesting that the occurrence of symptoms does not change considerably with the histological stage as shown in figure 17.

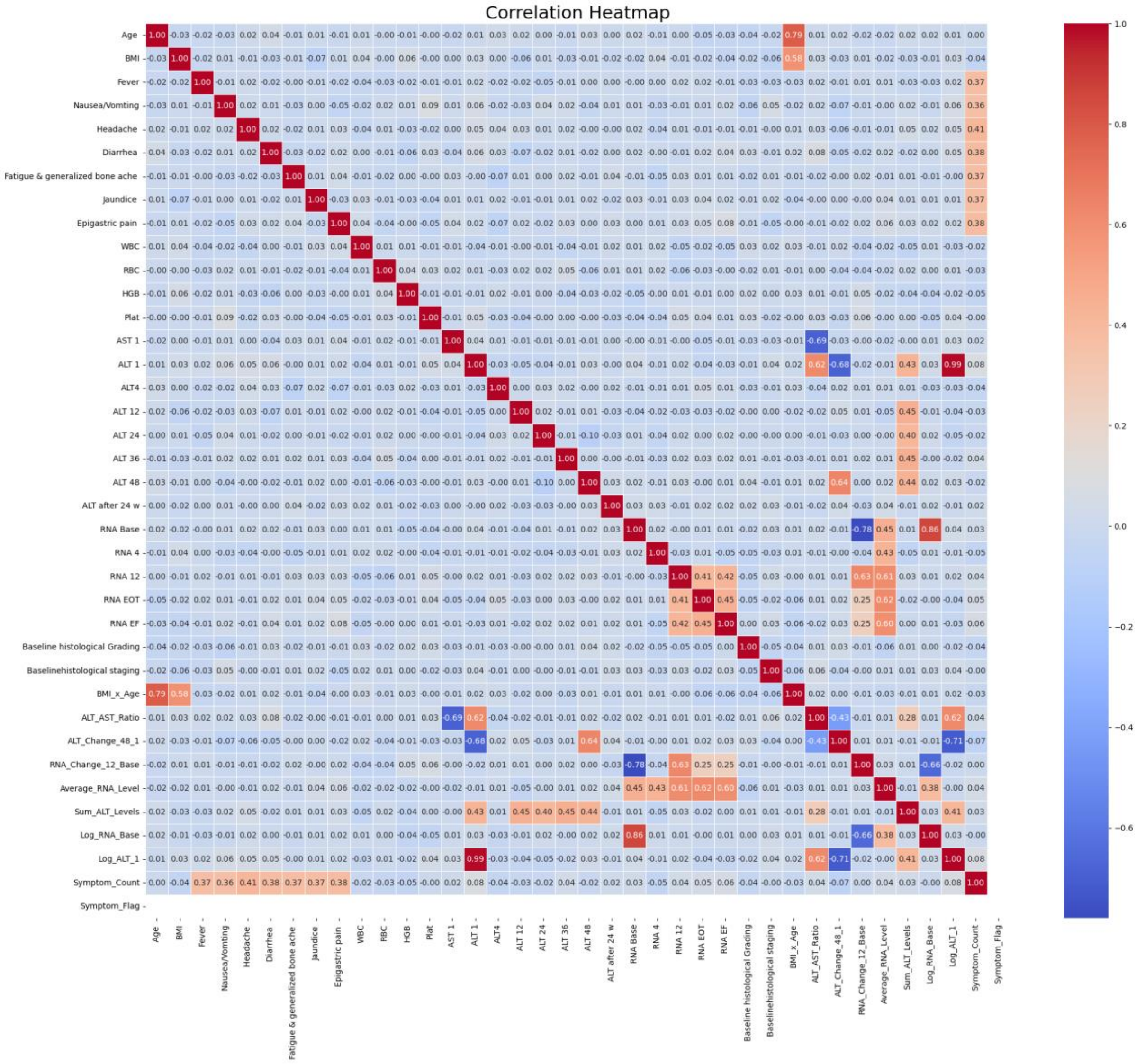


Figure 18: Corelation Heatmap

Figure 18 shows the dataset's correlation heatmap's entire view of numerical feature connections. A positive correlation exists between 'BMI_x_Age' and 'BMI.' The interaction characteristic BMI multiplied by Age increases when BMI increases. The nature of the derived feature predicts this relationship. Similarly, the 'ALT_AST_Ratio' has a small positive correlation with individual ALT levels, showing that when ALT levels rise, so does the ratio of ALT to AST. However, most feature correlations are weak, showing that many factors are not linearly related. This means that the model may need to address more complex relationships than just linear dependencies. Interestingly, the 'Baseline Histological Staging' did not correlate with most attributes, suggesting that multiple factors may determine the histological stage.

As part of clustering analysis as shown in figure 19, we utilized the K-Means algorithm to identify optimal groupings within the dataset. Initially, numerical features were standardized using the

StandardScaler to ensure each feature contributed equally to the distance calculations in the clustering process. The Elbow Method was employed to determine the optimal number of clusters, with the plot suggesting three or four as potential choices. Subsequently, we performed K-Means clustering with both three and four clusters. The resulting clusters were visualized using PCA plots, where each point's colour represented its cluster assignment. The three-cluster solution showed a clear separation between groups, whereas the four-cluster solution indicated some overlapping regions, suggesting that three clusters might be more appropriate for this data set.

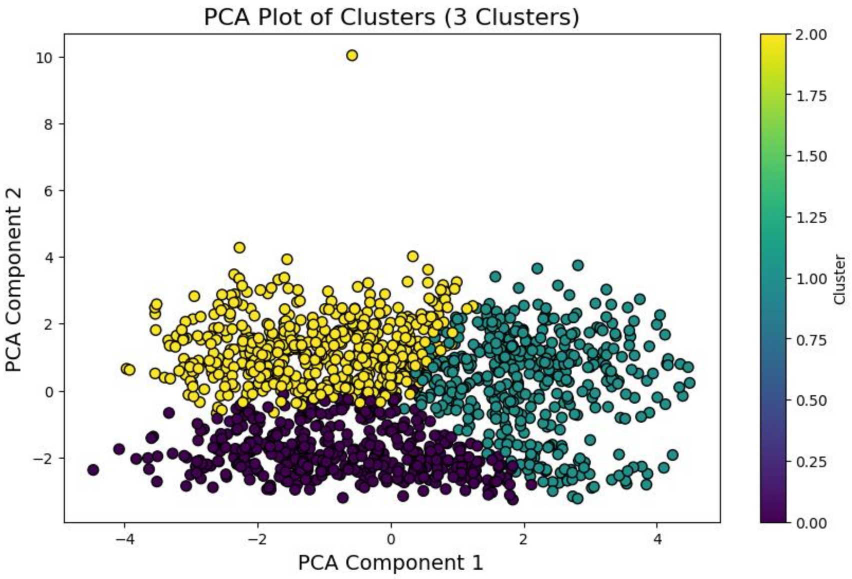
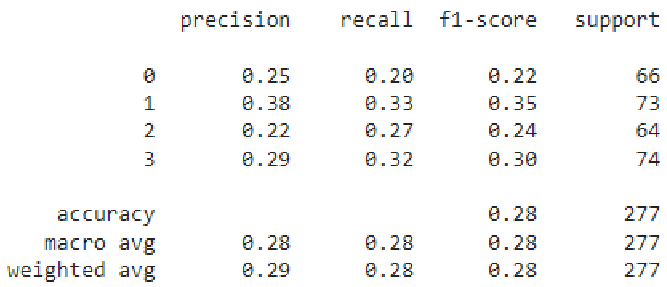


Figure 19: K-Means Clustering

4.3 Model Results

4.3.1 Logistic Regression:

The Logistic Regression model shows notable challenges in distinguishing between the different stages of Baseline Histological Staging. The confusion matrix shown in figure 21 reveals significant misclassification, especially between stages 0 and 3, where many instances are incorrectly predicted. The ROC curve shown in figure 22 indicates the model's limited ability to distinguish between classes, with AUC values around 0.53 to 0.57, suggesting poor model performance overall. The precision, recall, and F1-scores across all classes hover around 0.20 to 0.35, indicating that this model struggles to make accurate predictions as shown in figure 20.



```
              precision    recall  f1-score   support

           0       0.25      0.20      0.22        66
           1       0.38      0.33      0.35        73
           2       0.22      0.27      0.24        64
           3       0.29      0.32      0.30        74

    accuracy                           0.28       277
   macro avg       0.28      0.28      0.28       277
weighted avg       0.29      0.28      0.28       277
```

Figure 20: Logistic Regression Classification Report

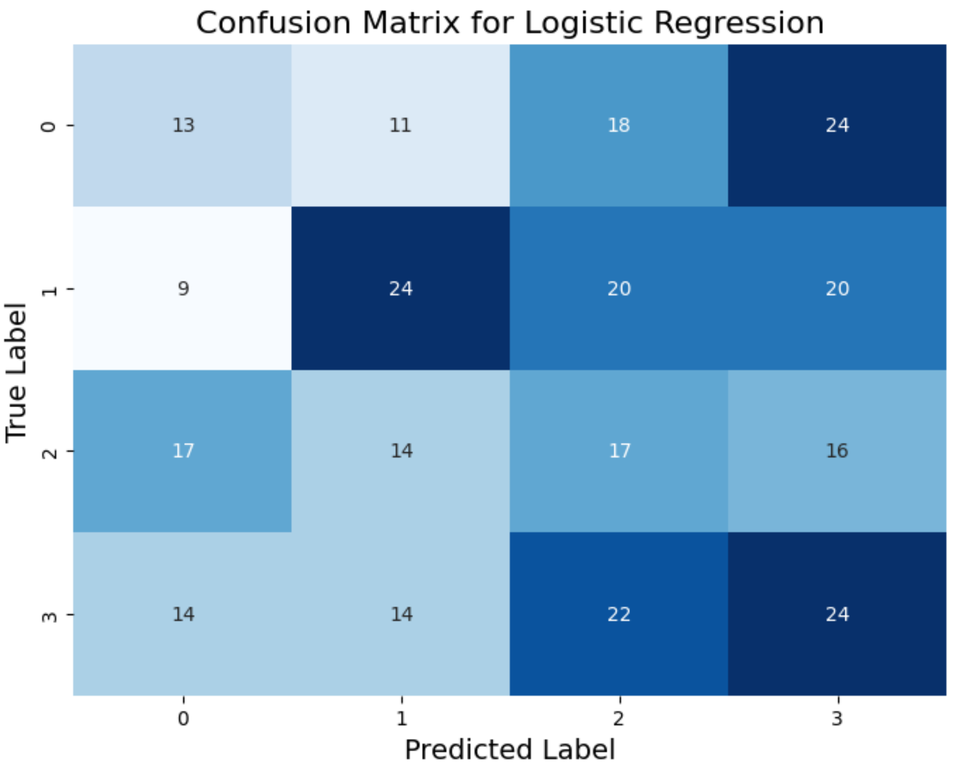


Figure 21: Logistic Regression Confusion Matrix

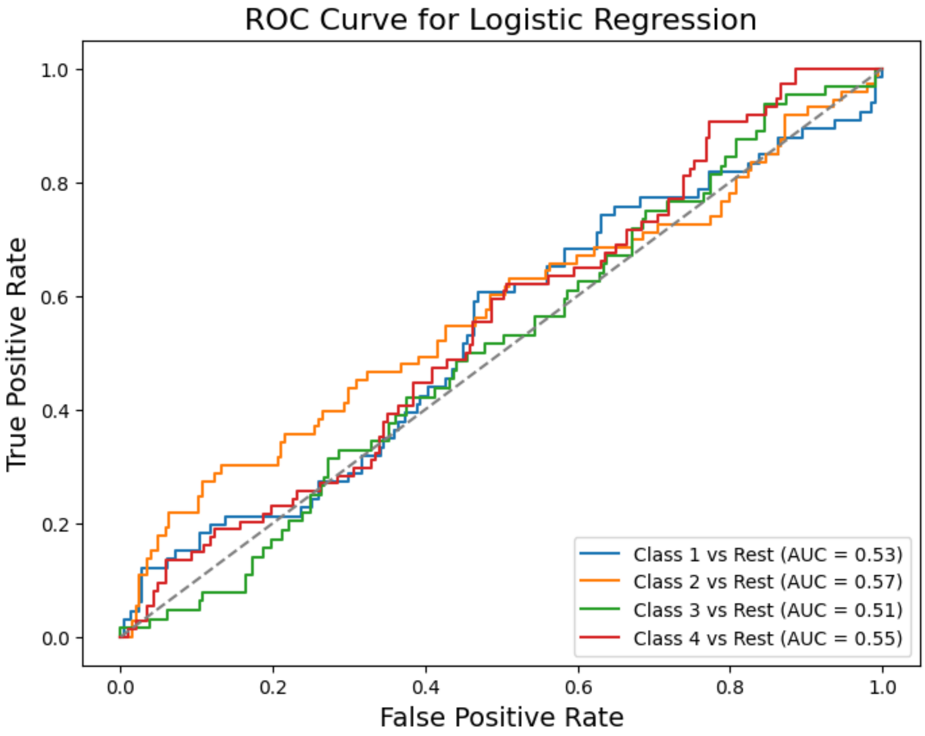
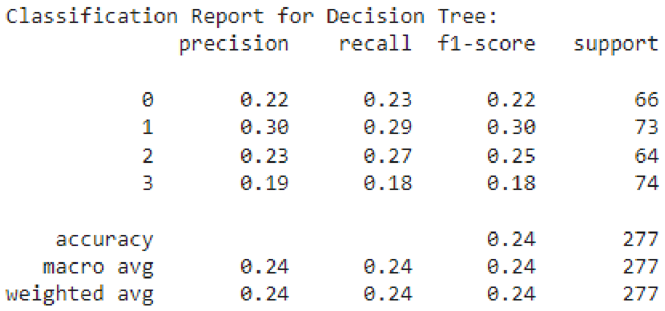


Figure 22 : Logistic Regression ROC Curve

4.3.2 Decision Tree:

The Decision Tree model also shows significant misclassification as shown in 23, as evidenced by its confusion matrix as shown in 24. The model frequently confuses adjacent stages, with some improvement in precision and recall compared to Logistic Regression but still performing poorly overall. The ROC curves for the Decision Tree model show slightly lower AUC values compared to Logistic Regression, with most AUCs around 0.45 to 0.53 as shown in 25. This suggests that while the Decision Tree may capture some relationships in the data, it is not robust enough to provide accurate classifications in this context.



```
Classification Report for Decision Tree:
              precision    recall  f1-score   support

           0       0.22      0.23      0.22        66
           1       0.30      0.29      0.30        73
           2       0.23      0.27      0.25        64
           3       0.19      0.18      0.18        74

    accuracy                           0.24       277
   macro avg       0.24      0.24      0.24       277
weighted avg       0.24      0.24      0.24       277
```

Figure 23: Decision Tree Classification Report

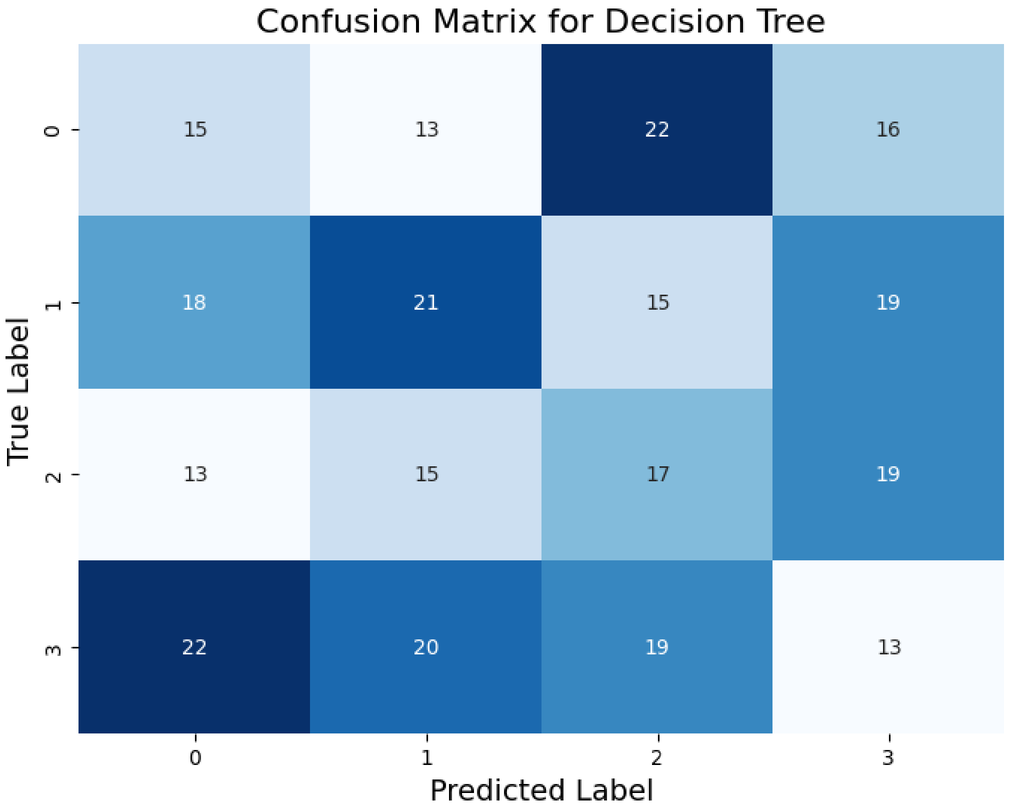


Figure 24: Decision Tree Confusion Matrix

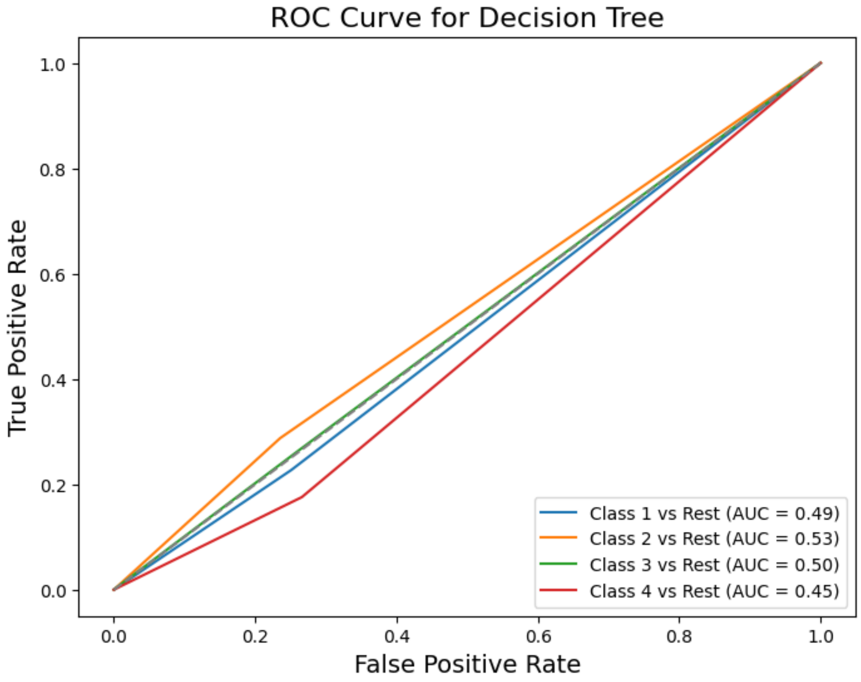
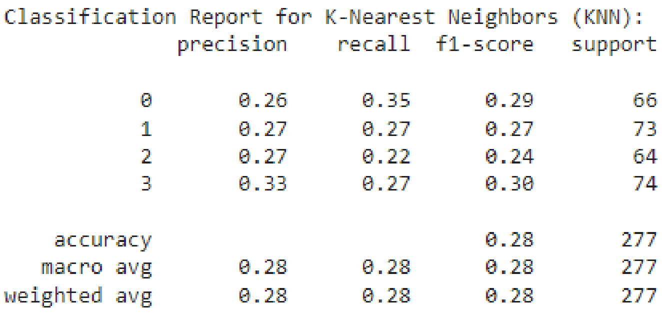


Figure 25: Decision Tree ROC Curve

4.3.3 K-Nearest Neighbours (KNN):

K-Nearest Neighbours exhibits similar issues with misclassification as shown in figure 26 its confusion matrix shown in figure 27, though it shows a slight improvement in recall for some classes. The ROC curves reveal that the KNN model performs better than the Decision Tree in some cases, particularly with class 3, which has an AUC of 0.59. However, the overall performance is still suboptimal, with the model struggling to clearly separate the stages, as reflected in AUC values ranging from 0.50 to 0.59 as shown in figure 28.



```
Classification Report for K-Nearest Neighbors (KNN):
              precision    recall  f1-score   support

           0       0.26      0.35      0.29        66
           1       0.27      0.27      0.27        73
           2       0.27      0.22      0.24        64
           3       0.33      0.27      0.30        74

    accuracy                           0.28       277
   macro avg       0.28      0.28      0.28       277
weighted avg       0.28      0.28      0.28       277
```

Figure 26: K-Nearest Neighbour Classification Report

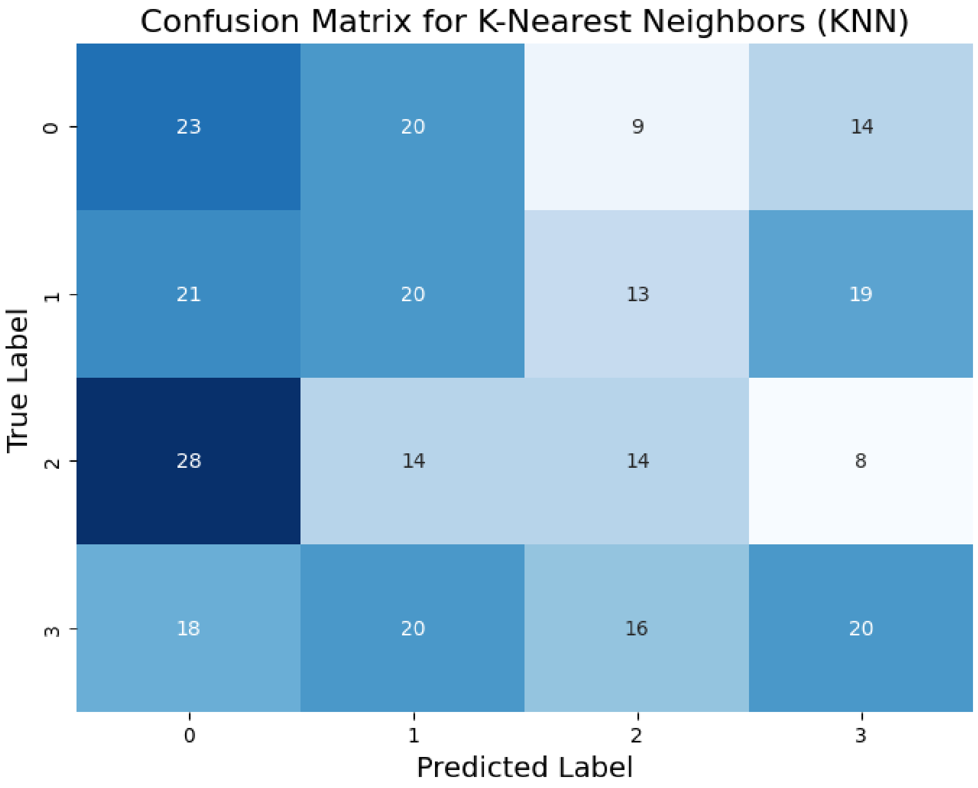


Figure 27: K-Nearest Neighbour Confusion Matrix

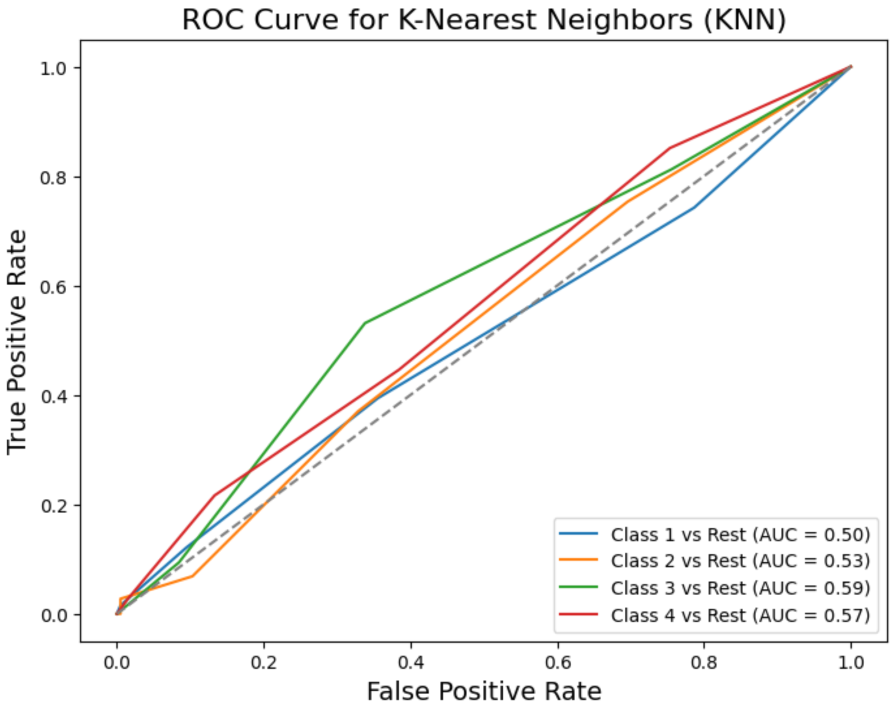


Figure 28: K-Nearest Neighbour ROC Curve

4.3.4 Support Vector Machine (SVM):

The SVM model also fails to accurately predict the classes, with the confusion matrix showing significant errors, particularly in predicting classes 0 and 3. The ROC curve for the SVM model indicates poor classification capability, with AUC values mostly below 0.50 for several classes, and the highest AUC is only 0.52 for class 1 as shown in figure 29. The model's precision, recall, and F1-scores reflect its struggle to accurately classify the histological stages as shown in figure 30.

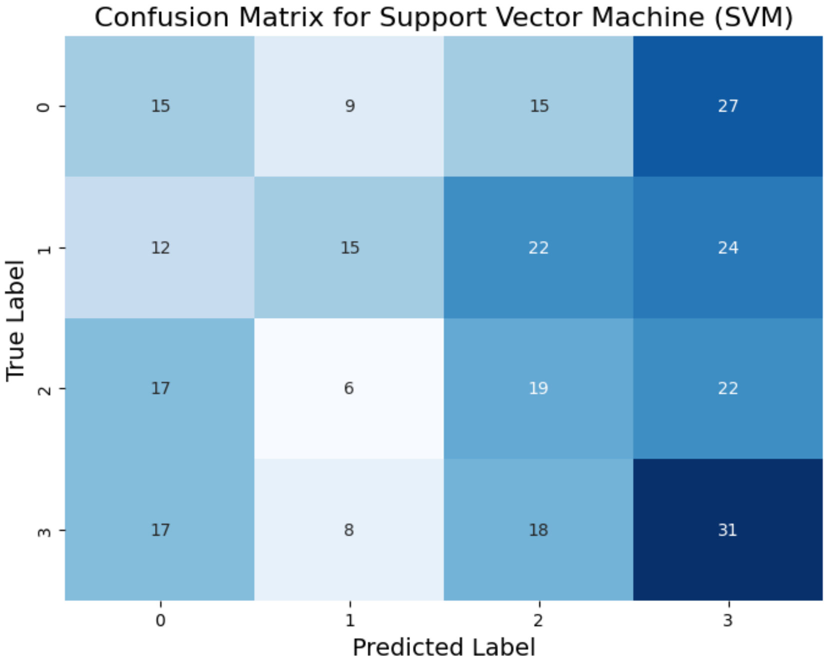


Figure 29: Support Vector Machine Confusion Matrix

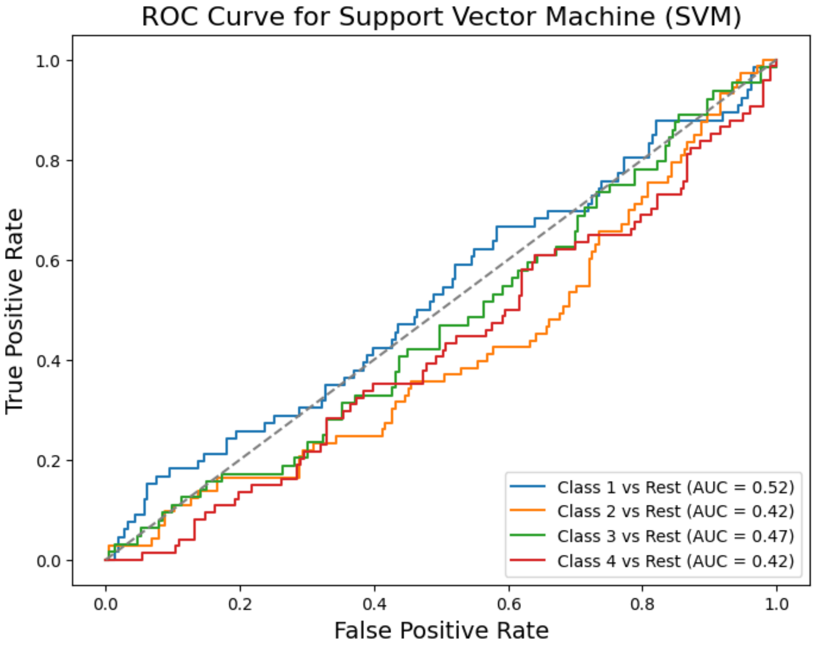
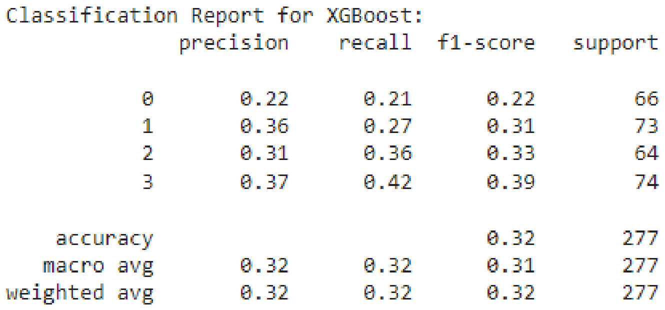


Figure 30: Support Vector Machine ROC Curve

4.3.5 XGBoost:

XGBoost provides slightly better performance compared to the other models, with its confusion matrix showing fewer misclassifications in certain classes. The ROC curves for XGBoost show the highest AUC values among the models, particularly with class 4 (AUC of 0.60) and class 3 (AUC of 0.57) as shown in 33. While these values indicate some improvement, the model still performs inadequately overall, with precision, recall, and F1-scores indicating that further improvements are necessary to achieve reliable predictions as shown in figure 31 and 32.



```
Classification Report for XGBoost:
              precision    recall  f1-score   support

           0       0.22      0.21      0.22        66
           1       0.36      0.27      0.31        73
           2       0.31      0.36      0.33        64
           3       0.37      0.42      0.39        74

    accuracy                           0.32       277
   macro avg       0.32      0.32      0.31       277
weighted avg       0.32      0.32      0.32       277
```

Figure 31: XGBoost Classification Report

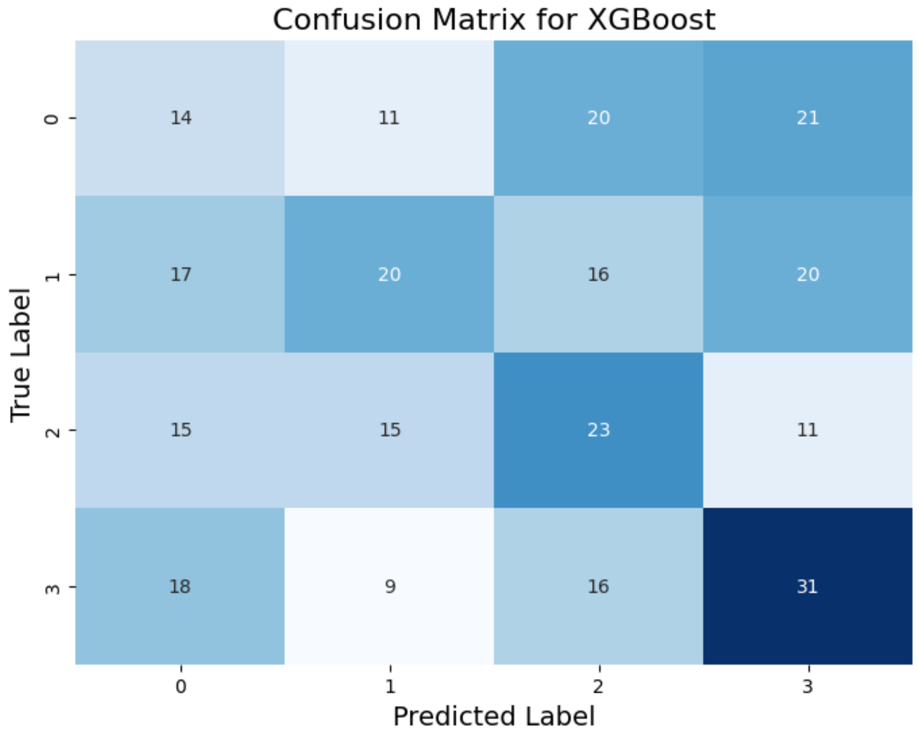


Figure 32: XGBoost Confusion Matrix

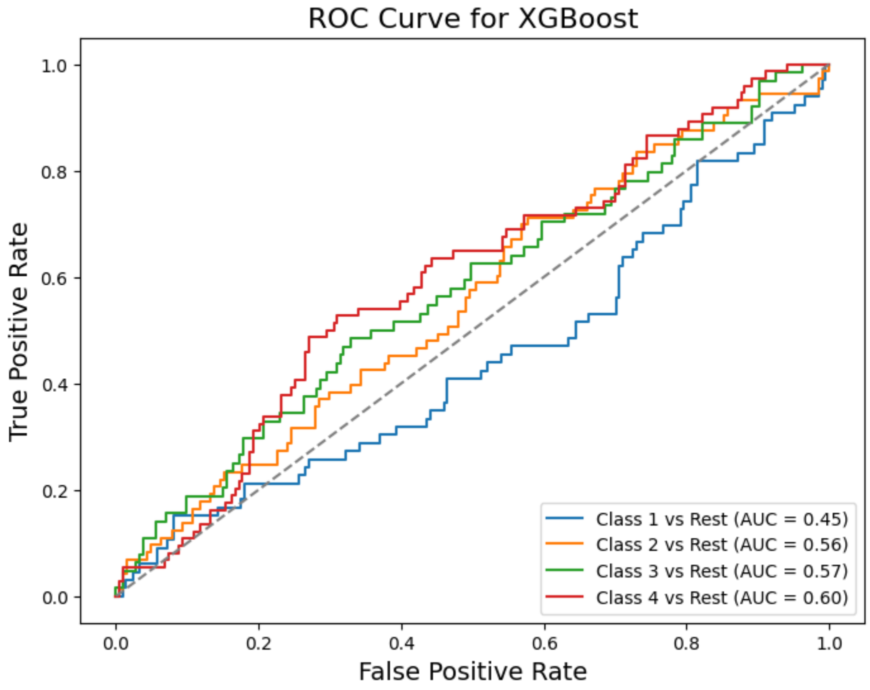


Figure 33: XGBoost ROC Curve

4.3.6 Classification Report Summary

The classification reports as shown in table 1, for each model highlight the performance metrics such as precision, recall, and F1-score for each stage of Baseline Histological Staging. The overall accuracy for each model is also provided. Below is a table that summarizes these key metrics for all models:

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)
Logistic Regression	28%	0.28	0.28	0.28
Decision Tree	24%	0.24	0.24	0.24
K-Nearest Neighbours (KNN)	28%	0.28	0.28	0.28
Support Vector Machine (SVM)	29%	0.3	0.29	0.28
XGBoost	32%	0.32	0.32	0.31

Table 1: Classification Report Summary

Insights:

Logistic Regression achieved an accuracy of 28%, with precision, recall, and F1-scores all around 0.28, indicating limited capability in classifying the stages accurately. Decision Tree had the lowest accuracy at 24%, with precision, recall, and F1-scores similarly low at 0.24, reflecting its tendency to misclassify samples. K-Nearest Neighbours (KNN) matched Logistic Regression with a 28% accuracy, and its precision, recall, and F1-scores also stayed at 0.28, showing consistent but not highly effective performance. The Support Vector Machine (SVM) slightly improved upon this with a 29% accuracy and a precision of 0.30, but struggled with class imbalances, leading to inconsistent recall and F1-scores. XGBoost outperformed all other models, achieving the highest accuracy at 32%, with precision, recall, and an F1-score of 0.32, making it the most effective model in this analysis, though it still faced challenges in classification.

4.3.7 Hyperparameter Tuning Result

In this investigation, we assessed the efficacy of various machine learning models in a classification job using a dataset comprising four distinct classes as shown in table 2. The evaluated models consist of Logistic Regression, Decision Tree, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and XGBoost. We conducted hyperparameter tuning for each model using GridSearchCV to determine the optimal settings that yielded the highest performance. Here is a concise overview of the classification reports for each model, which includes the optimised parameters and their corresponding values.

Model	Precision (Avg)	Recall (Avg)	F1-Score (Avg)	Best Parameters
Logistic Regression	0.05	0.23	0.09	{'C': 0.01, 'penalty': 'l1', 'solver': 'liblinear'}
Decision Tree	0.24	0.25	0.2	{'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}
K-Nearest Neighbours (KNN)	0.27	0.26	0.26	{'metric': 'manhattan', 'n_neighbors': 7, 'weights': 'uniform'}
Support Vector Machine (SVM)	0.05	0.23	0.09	{'C': 0.1, 'gamma': 'scale', 'kernel': 'rbf'}
XGBoost	0.26	0.26	0.25	{'learning_rate': 0.3, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.8}

Table 2: Hyper tuned Parameter Report Summary

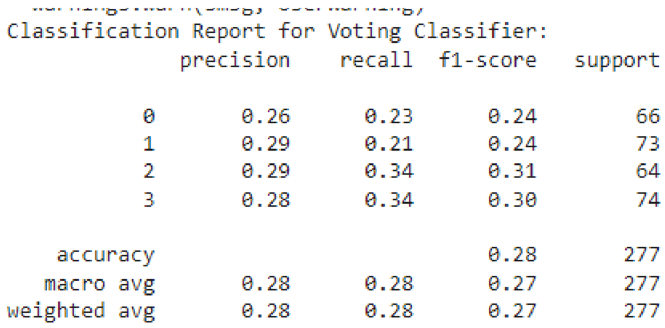
4.4 Advanced Machine Learning Models

We further analysed with some advanced machine learning model which included voting classifier, stacking classifier and Neural Network as shown in table 3

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)
Voting Classifier	0.28	0.28	0.27	0.28
Stacking Classifier	0.24	0.24	0.24	0.24
Basic Neural Network	0.35	0.34	0.34	0.34

Table 3: Advanced Machine Learning Report Summary

Among the three ensemble models, namely the Voting Classifier, Stacking Classifier, and Basic Neural Network, the Basic Neural Network exhibited the most superior performance overall. The Basic Neural Network attained an accuracy of 0.35, with a macro average precision, recall, and F1-score all at 0.34 as shown in figure 34 and 35. This suggests that the neural network had superior ability in identifying and understanding patterns in the data, as compared to the ensemble approaches.



```
Classification Report for Voting Classifier:
              precision    recall  f1-score   support

           0       0.26      0.23      0.24        66
           1       0.29      0.21      0.24        73
           2       0.29      0.34      0.31        64
           3       0.28      0.34      0.30        74

    accuracy                           0.28       277
   macro avg       0.28      0.28      0.27       277
weighted avg       0.28      0.28      0.27       277
```

Figure 34 :Voting Classifier Classification Report

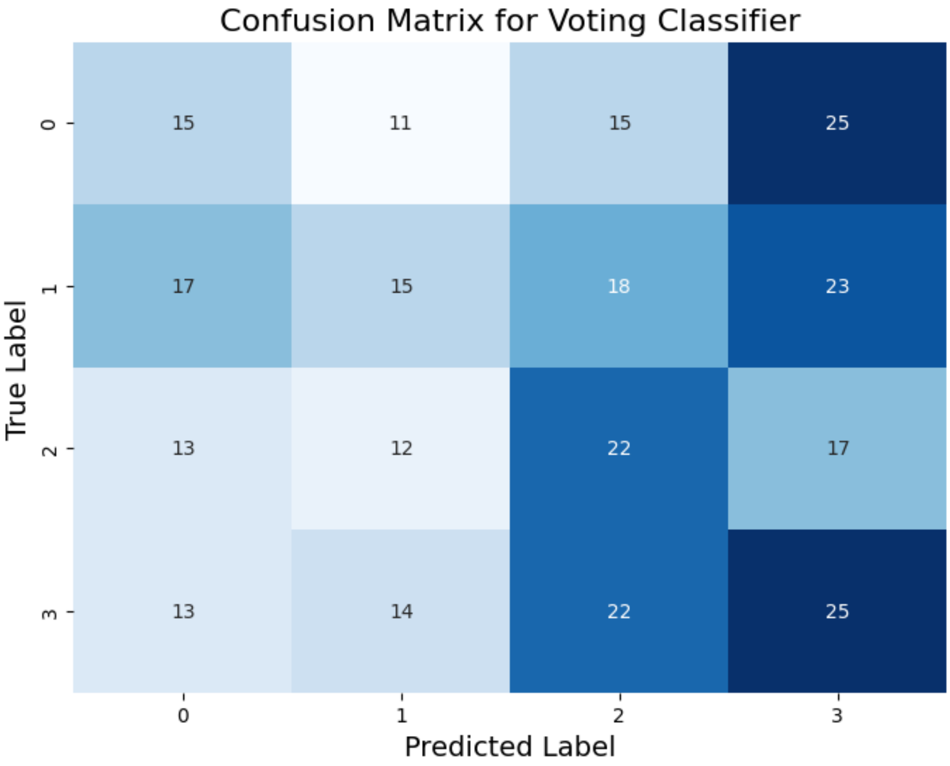
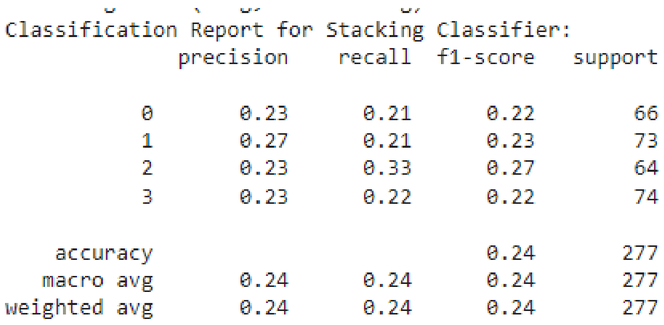


Figure 35: Voting Classifier Confusion Matrix

The Voting Classifier, which combines predictions from many models by averaging their probabilities (soft voting), attained an accuracy of 0.28. The macro mean values for precision, recall, and F1-score were all 0.28, indicating a balanced but slightly inferior performance compared to the neural network as shown in figure 24 (a) and 24 (b). The Voting Classifier demonstrated the capacity to integrate the advantages of various models, resulting in a moderate enhancement compared to individual models, although it still fell short in performance compared to the neural network.



```
Classification Report for Stacking Classifier:
              precision    recall  f1-score   support

           0       0.23      0.21      0.22        66
           1       0.27      0.21      0.23        73
           2       0.23      0.33      0.27        64
           3       0.23      0.22      0.22        74

    accuracy                           0.24       277
   macro avg       0.24      0.24      0.24       277
weighted avg       0.24      0.24      0.24       277
```

Figure 36: Stacking Classifier Classification Report

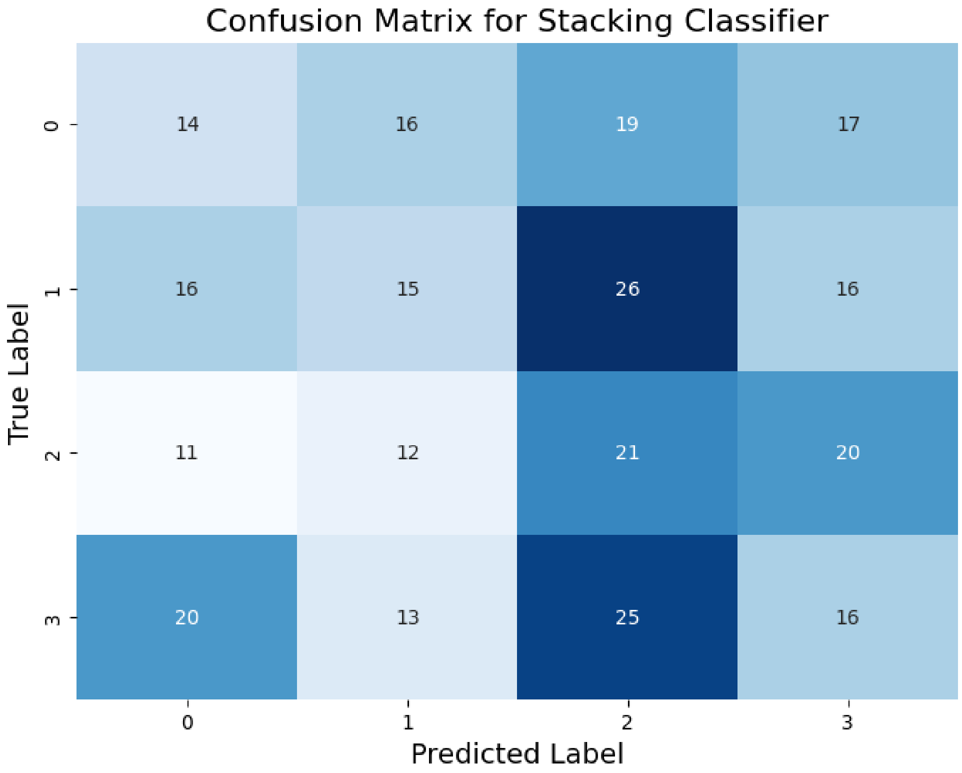


Figure 37: Stacking Classifier Confusion Report

The Stacking Classifier, which utilises a meta-classifier (specifically, Logistic Regression) to integrate predictions from different models, had the poorest performance among the three. It achieved an accuracy of 0.24 and macro averages of 0.24 for precision, recall, and F1-score as shown in figure 25 (a) and 25 (b). Although stacking has a potential advantage in utilising many models, it fared worse than the other methods in this case.

4.4 Feature Importance

4.4.1 Recursive Feature Elimination (RFE) Results:

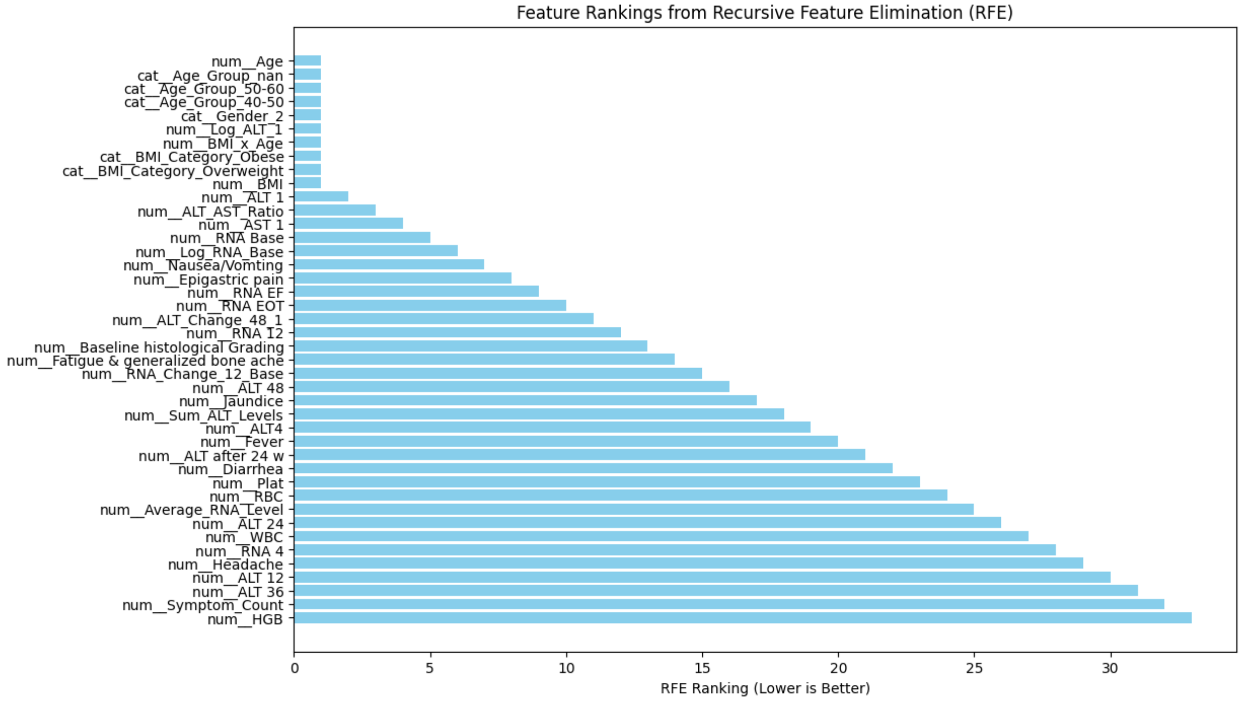


Figure 38 : Recursive Feature Elimination (RFE) Results

The RFE method selected the top 10 features based on their predictive power using a Logistic Regression model as shown in figure 26. The selected features are:

num__Age

num__BMI

num__BMI_x_Age

num__Log_ALT_1

cat__Gender_2

cat__Age_Group_40-50

cat__Age_Group_50-60

cat__Age_Group_nan

cat__BMI_Category_Obese

cat__BMI_Category_Overweight

These features were determined to have the highest impact on the model's ability to classify the target variable correctly. Age and BMI-related features, as well as some categorical features related to age groups and BMI categories, were among the most critical.

4.4.2 Random Forest Feature Importance:

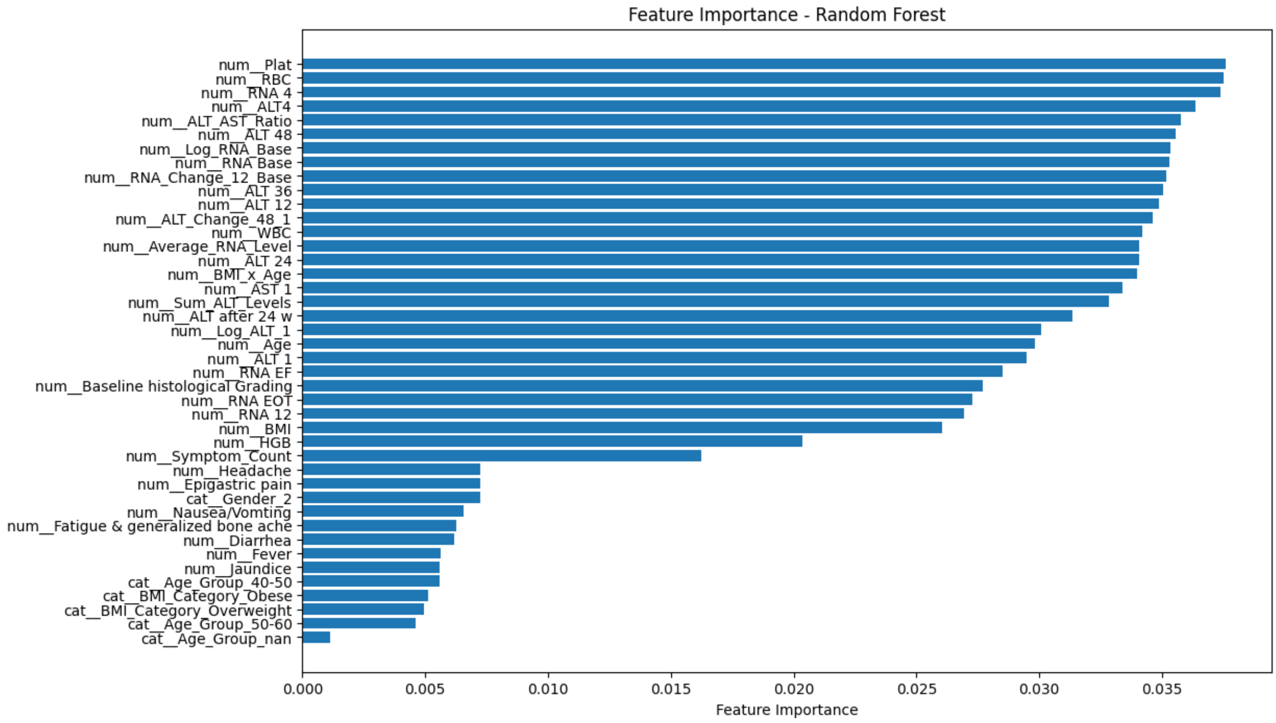


Figure 39: Random Forest Feature Importance:

The Random Forest model provides a different perspective by evaluating the importance of all features in the dataset. The feature importance scores are derived from how much each feature contributes to reducing the impurity of the node in the decision trees as shown in figure 27.

Key observations from the Random Forest feature importance plot include:

num__Plat and num__RBC were among the most influential features, indicating that platelet count and red blood cell count are highly predictive.

Several ALT (Alanine transaminase) and RNA measurements were also prominent, reflecting their significant role in predicting the outcome.

Features such as num__HGB (Hemoglobin) and num__Symptom_Count also showed considerable importance.

Chapter 5: Discussion

5.1 RO1: Data Preprocessing and Cleaning

During the first stage of our project, we prioritised the essential duties of data preprocessing and cleaning, which are fundamental elements in any data analysis pipeline. The main goals of this phase were to address any missing values and properly encode categorical variables in the Hepatitis C Virus (HCV) dataset, ensuring that the data was complete and prepared for further machine learning activities.

Missing values could compromise our predictive models' reliability and precision, so we addressed them first. Incomplete data can skew statistical results and make them untrustworthy. To mitigate this risk, we used imputation methods tailored for missing data features. Our numerical methods used mean or median imputation to replace missing values with the data's average or median. This preserved the dataset's central tendency and variability. For categorical features, we used mode imputation to replace missing values with the dataset's most frequent category. This strategy allowed us to keep the most data while maintaining dataset integrity. After imputation, the dataset was thoroughly examined to ensure no missing values, ensuring data integrity and stability for model training.

After completing the imputation process, we focused on the requirement to convert categorical variables into numerical format. This conversion is crucial for incorporating these variables into machine learning algorithms that specifically require numerical input. The HCV dataset included categorical characteristics such as gender, age groups, and BMI categories. These variables were transformed into numerical representations using one-hot encoding. This method converted each category into an individual binary column, therefore eliminating any implicit ordering or hierarchy that could potentially confuse the models. The choice of one-hot encoding was based on its capacity to maintain the non-ordinal characteristics of categorical data, hence preventing the models from mistakenly inferring connections between categories that do not actually exist.

The data preprocessing and cleaning phase played a crucial role in preparing the dataset for efficient model training. Through the use of imputation to handle missing values and one-hot encoding to convert categorical variables, we guaranteed that our dataset was both comprehensive and suitable for a wide range of machine learning methods. By engaging in meticulous preparation, we were able to proceed with confidence to the modelling step, assured that the data was dependable and suitably organised for precise predictive analysis. These steps not only improved the quality of our models but also established a strong basis for reliable and understandable outcomes in the following stages of our investigation.

5.2 RO2: Feature Selection and Analysis

The feature selection and analysis portion of our study yielded crucial insights into the primary characteristics that have the most impact on predicting outcomes associated with liver disorders, including Hepatitis C. We utilised Recursive Feature Elimination (RFE) in conjunction with a Logistic Regression model and evaluated feature relevance using a Random Forest classifier to systematically identify and validate the most significant predictors in our dataset.

The results of Recursive Feature Elimination (RFE) enabled us to prioritise and choose the most significant characteristics that enhance the accuracy of the model. The procedure played a crucial role in reducing the dataset to a subset of attributes that had the greatest predictive value. The features chosen by RFE encompassed a combination of demographic factors, such as age and BMI categories, as well as clinical signs including liver enzymes (ALT, AST) and RNA levels. These characteristics are widely recognised in medical literature for their significant correlation with liver health and the advancement of liver disorders, rendering them essential for constructing accurate predictive models. The emphasis on these characteristics corresponds to the clinical comprehension, confirming their significance in the context of Hepatitis C treatment.

The Random Forest model conducted a thorough examination of feature importance, which served as an additional means of validation and provided valuable insights, complementing the Recursive Feature Elimination (RFE) process. The Random Forest analysis confirmed the importance of the traits found by RFE and also highlighted the presence of new factors. Platelet count (num_Plat) and red blood cell count (num_RBC) were found to be highly influential features, indicating their crucial significance in evaluating liver function and the body's reaction to viral infections. The continuous significance of ALT and AST levels in both RFE and Random Forest analyses highlights their crucial function as indicators in liver disease.

Key Insights: The utilisation of Recursive Feature Elimination (RFE) in conjunction with Random Forest analysis resulted in a reliable approach for selecting features, guaranteeing the identification of the most influential predictors. By employing this dual method, we were able to maximise the effectiveness of our predictive models by lowering the complexity of the data. Additionally, this approach improved the clinical interpretability of our findings. The findings indicate that a certain group of characteristics, specifically those associated with liver enzymes, blood counts, and viral load, have a crucial role in forecasting outcomes in liver-related disorders.

5.3 RO3: Develop, Optimize, and Evaluate Classification Models for Improved Predictive Accuracy

We developed, optimised, and assessed various classification models to improve projected accuracy. Logistic Regression, Decision Trees, KNN, SVM, and XGBoost were first created. Logistic Regression had 28% accuracy, 0.28 macro-average precision, recall, and F1-score. Classifying phases 0 and 3 was difficult, limiting this model's performance. This was shown by confusion matrices and ROC curves with AUC values of 0.53–0.57. The Decision Trees algorithm performed poorly with 24% accuracy. It confused nearby stages and had AUC values of 0.45–0.53. The KNN algorithm improved slightly, achieving 28% accuracy and 0.59 AUC for class 3. The overall performance was low, with AUC values of 0.50–0.59. The SVM provided somewhat better results with 29% accuracy. However, its Receiver Operating Characteristic (ROC) curves showed AUC values below 0.50, indicating minimal class distinction. XGBoost outperformed the other models with 32% accuracy. AUC values were 0.60 for class 4 and 0.57 for class 3. Precision, recall, and F1-scores show more refinement is needed.

Ensemble approaches were employed to harness the advantages of many algorithms. The Voting Classifier aggregated predictions from multiple models, yielding an accuracy of 28% and a macro-average precision, recall, and F1-score of 0.28. Although it provided modest enhancement compared to individual models, it did not substantially boost performance. The Stacking Classifier, which employed a Logistic Regression meta-classifier, had the lowest performance among ensemble approaches, with an accuracy of 24% and matching macro averages of 0.24 for precision, recall, and F1-score. The Basic Neural Network, specifically created for multiclass classification, exhibited the highest performance with an accuracy of 35% and macro-average precision, recall, and F1-score of 0.34. This demonstrates its exceptional ability in recognising patterns and distinguishing between different classes, surpassing the performance of other models.

In order to assess and enhance performance, we utilised various methodologies. The evaluation of each model involved the use of important metrics like as accuracy, precision, recall, and F1-score. Special emphasis was placed on macro-averaged scores to address any imbalances in the classes. The technique of cross-validation was employed to assess the ability of the models to perform consistently across various subsets of data. The results showed that all models had difficulties in accurately predicting outcomes beyond the training data. A grid search cross-validation was performed to optimise hyperparameters and improve performance, resulting in notable enhancements, particularly for XGBoost. Nevertheless, even with the implemented optimisations, all models encountered difficulties due to class imbalances. Analysed confusion matrices and ROC curves to gain a detailed understanding of performance. Findings indicate that the Basic Neural Network and XGBoost shown superior ability in discriminating between classes, while there are still areas that require development.

Ultimately, despite the limitations of all the models, the Basic Neural Network proved to be the most efficient, attaining the highest level of accuracy and demonstrating superior management of class imbalances.

5.4 RO4: Result Interpretation

5.4.1 Summarize Key Findings

The analysis revealed several crucial insights into the performance of different machine learning models and their efficacy in predicting liver histological staging in Hepatitis C Virus (HCV) patients. Feature selection identified several important variables, including interaction terms like BMI_x_Age and biomarkers such as ALT and RNA levels, which were significant predictors of histological outcomes. Among the models evaluated, XGBoost emerged as the most effective, outperforming Logistic Regression, Decision Trees, K-Nearest Neighbours (KNN), and Support Vector Machines (SVM). XGBoost demonstrated superior performance with higher accuracy and AUC values, particularly excelling in distinguishing advanced disease stages. However, despite these improvements, all models faced challenges in achieving high precision and recall across all classes, indicating a need for further refinement. Key features such as BMI and baseline ALT levels were consistently highlighted as significant, emphasizing their role in predicting disease progression.

5.4.2 Provide Actionable Insights

The insights derived from the model outputs offer valuable implications for clinical practice and patient management in the context of HCV outcomes. For clinicians, incorporating features such as BMI and baseline ALT levels into routine assessments could enhance the accuracy of predicting liver histological stages. The superior performance of XGBoost suggests that advanced machine learning techniques can assist in identifying patients at higher risk of progression, potentially guiding more targeted interventions. Moreover, understanding the impact of these features on disease outcomes can help in personalizing treatment plans, optimizing resource allocation, and improving patient prognostication. Future efforts should focus on leveraging these findings to develop decision-support tools that integrate these predictive features into clinical workflows, thereby enhancing overall patient care.

To aid in the interpretation and communication of results, several visualizations were created. Feature importance plots illustrate the relative contribution of various features to model predictions, highlighting the prominence of variables such as ALT levels and BMI in predicting histological staging. Confusion matrices provide a detailed breakdown of classification performance, showing where models succeeded and struggled with different stages. ROC curves offer a visual representation of model performance across various thresholds, demonstrating the trade-offs between true positive and false positive rates. These visualizations are essential for making the results more accessible and understandable to stakeholders, enabling more informed decision-making based on the model outputs.

Chapter 6: Conclusion

6.1 Summary of the Dissertation

This dissertation explored the application of advanced machine learning (ML) techniques to improve the management and prediction of outcomes for Hepatitis C Virus (HCV) patients. HCV continues to be a significant global health issue, particularly because it leads to severe liver conditions such as cirrhosis and hepatocellular carcinoma. Despite advancements in antiviral treatments, early detection and accurate prediction of disease progression are still critical to improving patient outcomes. Machine learning provides a promising avenue for addressing these challenges through predictive modelling and data analysis.

The research employed a systematic approach, starting with data preprocessing, followed by exploratory data analysis (EDA), and the development of predictive models. Data cleaning techniques ensured that the dataset was reliable, while feature selection methods such as Recursive Feature Elimination (RFE) and Random Forests identified significant predictors related to liver disease progression. Various machine learning models, including Logistic Regression, Decision Trees, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), XGBoost, and neural networks, were evaluated for their performance in predicting histological stages of liver disease.

Findings showed that advanced models like XGBoost and neural networks performed better than traditional models, such as Logistic Regression and Decision Trees, in terms of accuracy and reliability. However, even the best-performing models faced challenges due to class imbalances and the difficulty in precisely classifying all stages of the disease.

6.2 Research Contribution

This dissertation has made several key contributions to the field of machine learning in healthcare, particularly regarding the management of HCV. The study developed a comprehensive ML framework that integrates data preprocessing, feature selection, and predictive modelling to enhance the accuracy of HCV outcome prediction. This framework can serve as a foundation for future research and can be adapted to other chronic diseases.

Through feature selection techniques, the study identified crucial variables such as ALT, BMI, and RNA levels as significant predictors of liver disease progression. These findings align with established clinical knowledge, highlighting the importance of combining demographic and biochemical data in predicting patient outcomes.

The research also demonstrated that advanced machine learning models, such as XGBoost and neural networks, are more effective than simpler models in capturing complex relationships within the dataset. The evaluation of different models revealed that while XGBoost performed well, neural networks excelled in recognizing patterns and predicting disease progression.

Finally, the use of visualizations, including feature importance plots and confusion matrices, allowed for clearer interpretation of the results.

6.3 Future Research and Development

Several areas for future research and development have been identified based on the findings of this dissertation. One of the main challenges faced in this study was the issue of class imbalances, which negatively impacted model performance. Future research could explore techniques such as Synthetic Minority Over-sampling (SMOTE) or cost-sensitive learning to address this problem and improve model accuracy across all classes.

Another area for future research is the integration of additional clinical data, such as genetic information, medical history, and imaging data, into machine learning models. This would provide a more comprehensive view of the patient’s condition and improve predictive accuracy. Moreover, developing real-time predictive systems that can be integrated into clinical workflows would allow healthcare providers to make more informed decisions about patient management and treatment.

There is also a need for further clinical validation of these machine learning models. Although this study demonstrated that ML techniques could effectively predict HCV outcomes, the models need to be tested in real-world clinical settings to assess their practicality and effectiveness.

Finally, ethical considerations such as data privacy and security must be addressed as machine learning becomes more prevalent in healthcare. Techniques such as federated learning, which allows models to be trained on decentralized data, could be explored to protect patient privacy while still enabling the development of robust predictive models.

6.4 Personal Reflection

The process of completing this dissertation has been both challenging and enriching. One of the most important lessons I learned was the critical role that data preprocessing plays in the success of machine learning models. Even the most advanced algorithms will fail if the underlying data is not properly cleaned and prepared. In this study, the careful handling of missing data, the encoding of categorical variables, and the standardization of numerical features were all essential steps that ensured the reliability of the models.

Feature selection was another crucial aspect of the research that significantly improved the performance and interpretability of the models. By using techniques such as RFE and Random Forests, I was able to identify the most important predictors of HCV outcomes. This step underscored the importance of balancing model complexity with interpretability, particularly in healthcare settings where clinical decisions can impact patient care.

Throughout the study, I gained a deeper understanding of the trade-offs between different machine learning models. While advanced models like XGBoost and neural networks demonstrated better performance, they required more computational resources and were less interpretable than simpler models like Logistic Regression. This experience highlighted the need to consider both performance and practicality when selecting models for specific tasks.

Reference List

Alizargar, A., Chang, Y.-L. and Tan, T.-H. (2023). Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques. Bioengineering, 10(4), p.481. doi:https://doi.org/10.3390/bioengineering10040481.

Azam, M., Sakib, N.S., Fahad, N.M., Mamun, A.A., Rahman, M.A., Shatabda, S. and Hossain, S. (2024). A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks. Decision Analytics Journal, pp.100470. doi:https://doi.org/10.1016/j.dajour.2024.100470.

Bonnet, A. (2023). Accuracy vs. Precision vs. Recall in Machine Learning: What is the Difference? [online] Encord.com. Available at: https://encord.com/blog/classification-metrics-accuracy-precision-recall/#:~:text=Accuracy%20is%20a%20fundamental%20metric [Accessed 7 Sep. 2024].

Brownlee, J. (2016). Logistic Regression for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

Brownlee, J. (2020). Stacking Ensemble Machine Learning With Python. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/.

Chew, X. and Khaw, K. (2023). Hepatitis C Virus Prediction by Machine Learning Techniques. Applications of Modelling and Simulation, 4, pp.89–100.

Cooksey, R.W. (2020a). Descriptive Statistics for Summarising Data. Illustrating Statistical Procedures: Finding Meaning in Quantitative Data, 1, pp.61–139. doi:https://doi.org/10.1007/978-981-15-2537-7_5.

Dayananda, S. (2023). Support Vector Machines (SVM). [online] Medium. Available at: https://sandundayananda.medium.com/support-vector-machines-svm-db8314e9092d#:~:text=SVM%20operates%20by%20mapping%20the.

Evi Diana Omar, Mat, H., Karim, A., Sanaudi, R., Ibrahim, F., Omar, M.A., Hafiz, Z., Jayaraj, V. and Goh, B.L. (2024). Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery. International Journal of Nephrology and Renovascular Disease, 17, pp.197–204. doi:https://doi.org/10.2147/ijnrd.s461028.

Ghazal, T.M. and Al-Islam, M. (2023). Hepatitis C Virus Data Analysis and Prediction Using Machine Learning Techniques. Scientific Reports, 13(1), p.1234. doi:https://doi.org/10.1038/s41598-023-31428-6.

Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J. and Gao, D. (2020). Degradation State Recognition of Piston Pump Based on ICEEMDAN and XGBoost. Applied Sciences, 10(18), p.6593. doi:https://doi.org/10.3390/app10186593.

H, R.S. (2023). K-Nearest Neighbors Algorithm. [online] Intuitive Tutorials. Available at:

https://intuitivetutorial.com/2023/04/07/k-nearest-neighbors-algorithm/.

Jainvidip (2024). Understanding Logistic Regression. [online] Medium. Available at: https://medium.com/@jainvidip/understanding-logistic-regression-e6b8f379a420.

Jordan, M.I. and Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.

Kay, L., Grisetti, L., Pratama, M.Y., Tiribelli, C. and Pascut, D. (2022). Biomarkers for the Detection and Management of Hepatocellular Carcinoma in Patients Treated with Direct-Acting Antivirals. Cancers, 14(11), pp.2700. doi:https://doi.org/10.3390/cancers14112700.

Kharkar, D. (2023). Unravelling the Power of XGBoost: Boosting Performance with Extreme Gradient Boosting. [online] Medium. Available at: https://medium.com/@dishantkharkar9/unravelling-the-power-of-xgboost-boosting-performance-with-extreme-gradient-boosting-302e1c00e555.

Kosarenko, Y. (2021). How to Create Decision Trees for Business Rules Analysis. [online] Why Change. Available at: https://why-change.com/2021/11/13/how-to-create-decision-trees-for-business-rules-analysis/.

Kumar, A. (2022). Support Vector Machine (SVM) Python Example. [online] Data Analytics. Available at: https://vitalflux.com/classification-model-svm-classifier-python-example/.

Lazarus, J.V., Roel, E. and Elsharkawy, A.M. (2019). Hepatitis C Virus Epidemiology and the Impact of Interferon-Free Hepatitis C Virus Therapy. Cold Spring Harbor Perspectives in Medicine, 10(3), p.a036913. doi:https://doi.org/10.1101/cshperspect.a036913.

Melendez-Torres, J. and Singal, A.G. (2022). Early detection of hepatocellular carcinoma: roadmap for improvement. Expert Review of Anticancer Therapy, 22(6), pp.621–632. doi:https://doi.org/10.1080/14737140.2022.2074404.

Patil, D.R. and Patil, J.B. (2018). Malicious URLs Detection Using Decision Tree Classifiers and Majority Voting Technique. Cybernetics and Information Technologies, 18(1), pp.11–29. doi:https://doi.org/10.2478/cait-2018-0002.

Pires, J. (2019). Hepatitis C Virus for Egyptian Patients Data Set. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/joelpires/hepatitis-c-virus-for-egyptian-patients-data-set [Accessed 7 Sep. 2024].

Sasidharan, A. (2021). Support Vector Machine Algorithm. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

Scikit-learn.org. (2012). 1.11. Ensemble methods — scikit-learn 0.22.1 documentation. [online] Available at: https://scikit-learn.org/stable/modules/ensemble.html.

Sharma, S. (2021). KNN - The Distance Based Machine Learning Algorithm. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/.

Visani, G. et al. (2022). A Survey of Methods for Explaining Black Box Models. ACM Computing

Surveys, 51(5), pp.1–42.

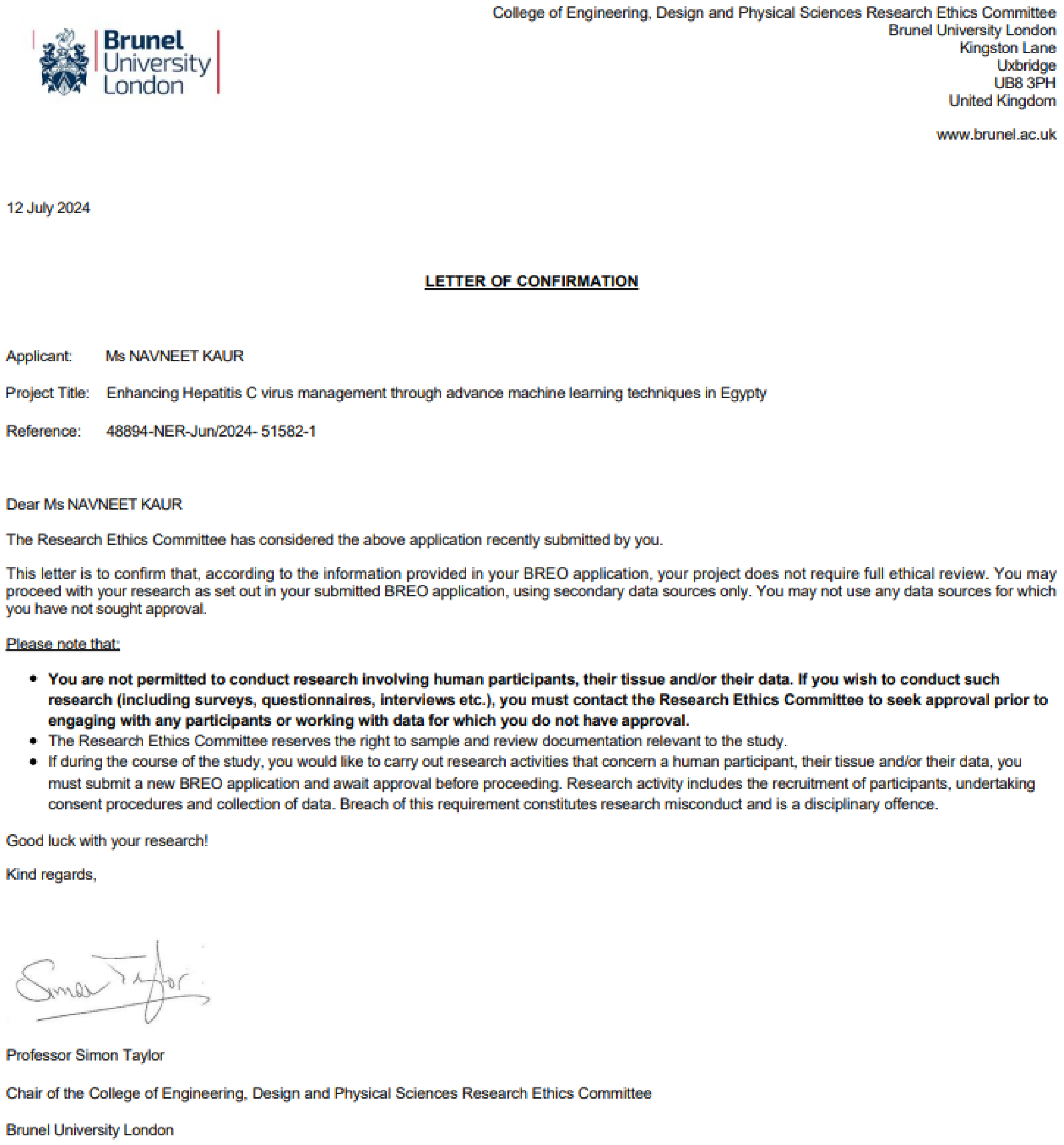
Wang, F., Wang, Q., Nie, F., Li, Z., Yu, W. and Ren, F. (2020). A linear multivariate binary decision tree classifier based on K-means splitting. Pattern Recognition, 107, p.107521. doi:https://doi.org/10.1016/j.patcog.2020.107521.

World Health Organization (WHO), 2021. Hepatitis C. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c [Accessed 6 July 2024].

Yağanoğlu, M. (2022). Hepatitis C virus data analysis and prediction using machine learning. Data & Knowledge Engineering, p.102087. doi:https://doi.org/10.1016/j.datak.2022.102087.

REFERENCES

APPENDIX A: ETHICAL APPROVAL



Brunel University London

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

www.brunel.ac.uk

12 July 2024

LETTER OF CONFIRMATION

Applicant: Ms NAVNEET KAUR

Project Title: Enhancing Hepatitis C virus management through advance machine learning techniques in Egypty

Reference: 48894-NER-Jun/2024- 51582-1

Dear Ms NAVNEET KAUR

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