ENHANCED HEPATOCELLULAR CARCINOMA CLASSIFICATION MODEL WITH RANDOM FOREST AND FUZZY LOGIC APPROACH

TAN YING FEI

UNIVERSITI KEBANGSAAN MALAYSIA

ENHANCED HEPATOCELLULAR CARCINOMA CLASSIFICATION MODEL WITH RANDOM FOREST AND FUZZY LOGIC APPROACH

TAN YING FEI

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF MASTER OF DATA SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI
2024

ENHANCED HEPATOCELLULAR CARCINOMA CLASSIFICATION MODEL WITH RANDOM FOREST AND FUZZY LOGIC APPROACH

TAN YING FEI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

10 July 2024

TAN YING FEI
P121496

ACKNOWLEDGEMENT

I would like to express my highest gratitude to my supervisor Ts. Dr Afzan Adam for her guidance and providing meaningful feedback to improve my thesis throughout the semesters. Besides that, I would like to thank all the lecturers and classmates I met during my studies at the Faculty of Information Science & Technology, UKM. Thank you for sharing your knowledge of data science. Last but not least, I am grateful to have full support from my dearest family.

ABSTRAK

Kanser hati primer, karsinoma hepatoselular (HCC) merupakan kanser keenam penghidap tertinggi di dunia, dan ketiga tertinggi dalam kematian akibat kanser. Walaupun teknologi rawatan semakin maju, kesan HCC masih kekal ketara. Kebanyakan pesakit lewat di diagnos disebabkan tiada gejala yang ditunjukkan pada peringkat awal. Untuk menangani situasi ini, rutin ujian saringan fungsi hati diperlukan untuk pengesanan awal HCC dan dengan itu mengurangkan beban HCC dalam perkidmatan kesihatan. Teknologi kecerdasan buatan telah digunakan secara meluas dalam perkhidmatan kesihatan untuk diagnos dan ramalan penyakit. Walaupun sesetengah penyelidikan telah menggunakan pembelajaran mesin bagi membangunkan klasifikasi model klasifikasi HCC, kebolehtafsirannya adalah terhad. Oleh yang demikian, kajian ini bertujuan untuk (1) mencadangkan model yang terbaik untuk klasifikasi risiko HCC, (2) mengoptimumkan model klasifikasi risiko HCC, dan (3) mempertingkatkan komponen penjelasan dalam model klasifikasi risiko HCC menggunakan fuzzy logik (FL). Data-data ujian fungsi hati ini diambil dari data penanda aras pada pangkalan Kaggle. Analisis, prapemprosesan data dan kejuruteraan fitur telah dijalankan bagi memastikan data yang berkualiti tinggi untuk prestasi model yang lebih baik. Model-model pengelasan seperti pepohon keputusan (DT), hutan rawak (RF), mesin vektor sokongan (SVM), peningkatan kecerunan (GB), regresi logistic (LR), teluk naif gaussian (GNB), k-jiran terdekat (KNN) dan rangkaian saraf (NN) dilatih dengan data-data ini. Keputusan ujian menunjukkan bahawa RF mempunyai prestasi yang terbaik setelah pengoptimuman dengan penalaan hiperparameter. RF mencapai ketepatan 99% dan AUROC 0.9996 Kemudian, FL telah digunakan untuk mentafsir taakulan diagnosis RF tersebut. Hasilnya, terdapat 5 fuzzy rules yang boleh digunakan seperti (1) JIKA Albumin RENDAH TETAPI KEROSAKAN HATI TINGGI (SIROSIS) (0.70) DAN SGPT NORMAL (1.00) DAN Nisbah A/G RENDAH (0.60) DAN Nisbah SGOT/SGPT ADALAH SIROSIS (SGOT/SGPT > 1) (1.00) DAN Jumlah Bilirubin RENDAH (1.00) MAKA kelas 1: HCC, (2) JIKA ALP TINGGI (1.00) DAN Nisbah A/G RENDAH (0.88) DAN SGPT NORMAL (1.00) MAKA kelas 0: non-HCC, (3) JIKA SGPT NORMAL (1.00) DAN Albumin RENDAH TETAPI KEROSAKAN HATI TINGGI (SIROSIS) (0.91) DAN Nisbah A/G RENDAH (0.70) DAN Jumlah Bilirubin RENDAH (0.93) DAN Jumlah Proteins RENDAH (0.36) DAN Nisbah SGOT/SGPT ADALAH HEPATITIS VIRUS (SGOT/SGPT = 1) (0.13), SIROSIS (SGOT/SGPT > 1) (1.00), PENYAKIT HATI ALKOHOLIK (SGOT/SGPT = 2) (0.87) MAKA kelas 1: HCC, (4) JIKA ALP TINGGI (0.50) DAN Jumlah Proteins NORMAL (1.00) DAN Albumin RENDAH TETAPI KEROSAKAN HATI TINGGI (SIROSIS) (0.24), NORMAL (0.52) DAN Nisbah SGOT/SGPT ADALAH SIROSIS (SGOT/SGPT > 1) (0.74) MAKA kelas 1: HCC, dan (5) JIKA ALP TINGGI (0.61) DAN SGPT NORMAL (1.00) Albumin RENDAH TETAPI KEROSAKAN HATI TINGGI (SIROSIS) (0.90) DAN Jumlah Bilirubin RENDAH (1.00) DAN Nisbah A/G RENDAH (0.56) DAN Nisbah SGOT/SGPT ADALAH NAFLD (SGOT/SGPT < 1) (0.43), HEPATITIS VIRUS (SGOT/SGPT = 1) (1.00) MAKA kelas 1: HCC. FL digunakan kerana ia dapat menguruskan ketidakpastian dan ketepatan dalam data. Ini amatlah berguna untuk membantu doktor membuat keputusan dalam diagnostik perubatan. Kesimpulannya, penyelidikan ini mempersembahkan pendekatan baru bagi membangunkan model pengelasan risiko

HCC yang mempunyai kemantapan dan ketepatan yang lebih tinggi serta hasil tafsiran yang lebih mudah difahami.

ABSTRACT

The primary liver cancer, hepatocellular carcinoma (HCC) ranks sixth in cancer incidence and third in cancer-related deaths globally. Despite advancements in treatment, the impact of HCC remains significant. Most patients experienced a delayed diagnosis due to no symptoms in the early stages. In order to tackle this scenario, regular screening of liver function tests is required for early detection and thereby reduce the burden of HCC. Recently, artificial intelligence has been widely applied in healthcare for disease diagnosis and prediction. Some research has used machine learning to develop an HCC classification model, but its interpretability is limited. Hence, this study aims to (1) suggest the optimal model for HCC risk classification, (2) optimize the HCC risk classification model, and (3) improve the explainable component of HCC risk classification model with fuzzy logic (FL). Liver function test data were collected from Kaggle. Data analysis, data preprocessing, and feature engineering have been done to ensure high-quality data for better model performance. The models were trained among different supervised machine learning classifiers such as decision tree, random forest, support vector machine, gradient boosting, logistic regression, gaussian naïve bayes, k-nearest neighbour, and neural network. The testing result showed that the random forest has the best performance after optimization with hyperparameters tuning. Random forest achieved an accuracy of 99% and AUROC of 0.9996. Furthermore, FL has been applied for the interpretation of the 5 extracted rules in random forest such as (1) IF Albumin IS LOW BUT HIGH LIVER DAMAGE (CIRRHOSIS) (0.70) AND SGPT IS NORMAL (1.00) AND A/G Ratio IS LOW (0.60) AND SGOT/SGPT ratio IS CIRRHOSIS (SGOT/SGPT > 1) (1.00) AND Total Bilirubin IS LOW (1.00) THEN class 1: HCC, (2) IF ALP IS HIGH (1.00) AND A/G Ratio IS LOW (0.88) AND SGPT IS NORMAL (1.00) THEN class 0: non-HCC, (3) IF SGPT IS NORMAL (1.00) AND Albumin IS LOW BUT HIGH LIVER DAMAGE (CIRRHOSIS) (0.91) AND A/G Ratio IS LOW (0.70) AND Total Bilirubin IS LOW (0.93) AND Total Proteins IS LOW (0.36) AND SGOT/SGPT Ratio IS VIRAL HEPATITIS (SGOT/SGPT = 1) (0.13), CIRRHOSIS (SGOT/SGPT > 1) (1.00), ALCOHOLIC LIVER DISEASE (SGOT/SGPT = 2) (0.87) THEN class 1: HCC, (4) IF ALP IS HIGH (0.50) AND Total Proteins IS NORMAL (1.00) AND Albumin IS LOW BUT HIGH LIVER DAMAGE (CIRRHOSIS) (0.24), NORMAL (0.52) AND SGOT/SGPT ratio IS CIRRHOSIS (SGOT/SGPT > 1) (0.74) THEN class 1: HCC, and (5) IF ALP IS HIGH (0.61) AND SGPT IS NORMAL (1.00) Albumin IS LOW BUT HIGH LIVER DAMAGE (CIRRHOSIS) (0.90) AND Total Bilirubin IS LOW (1.00) AND A/G Ratio IS LOW (0.56) AND SGOT/SGPT ratio IS NAFLD (SGOT/SGPT < 1) (0.43), VIRAL HEPATITIS (SGOT/SGPT = 1) (1.00) THEN class 1: HCC. FL can manage uncertainty and precision in data. This is especially useful in medical diagnosis helping doctors in decision making. In conclusion, this study presents a novel approach for developing a robust HCC risk classification model with high accuracy and interpreted outcomes that are easy to understand.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

A/G Ratio	Albumin/Globulin Ratio
ADASYN	Adaptive Synthetic Sampling
AI	Artificial Intelligence
ALP	Alkaline Phosphatase
ANN	Artificial Neural Network
AUROC	Area Under the Receiver Operating Characteristic
BCLC	Barcelona Clinic Liver Cancer
CT	Computed Tomography
DT	Decision Tree
FL	Fuzzy Logic
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GB	Gradient Boosting
GGT	Gamma-Glutamyl Transferase
GNB	Gaussian Naïve Bayes
HBV	Hepatitis B Viruses
HCC	Hepatocellular Carcinoma
HCTM	Hospital Canselor Tuanku Muhriz
HCV	Hepatitis C Viruses
IARC	International Agency for Research on Cancer
IQR	Interquartile Range
KNN	K-Nearest Neighbour
LFT	Liver Function Test
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-Layer Perceptron

MNCR	Malaysia National Cancer Registry
MRI	Magnetic Resonance Imaging
NAFLD	Non-alcoholic Fatty Liver Disease
NASH	Non-alcoholic Steatohepatitis
NB	Naïve Bayes
NN	Neural Network
PLC	Primary Liver Cancer
Q1	First Quartile
Q3	Third Quartile
RF	Random Forest
ROC	Receiver Operating Characteristic
RSF	Random Survival Forest
SGOT	Aspartate Aminotransferase
SGPT	Alanine Aminotransferase
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TSM	Treatment Stage Migration
UKM	Universiti Kebangsaan Malaysia
XGBoost	Extreme Gradient Boosting

CHAPTER I

INTRODUCTION

1.1 ARTIFICIAL INTELLIGENCE IN CANCER RESEARCH

Cancer is a chronic disease, one of the leading global health burdens, with a high rate of incidence and mortality. It is estimated that 10 million cancer deaths occur globally and the cancer burden is expected to rise by 60% over the next two decades, reaching 30 million additional cases by 2040, primarily in low and middle-income countries (Anon. 2023). To tackle the current scenario, accurate and timely diagnosis and prognosis are crucial for improving survival rates and avoiding the chance of recurrence.

Over the years, artificial intelligence (AI) has been widely applied in cancer research in conjunction with sophisticated bioinformatics tools due to its feasibility (S. Huang et al. 2020). AI in healthcare refers to computer-coded programs that are similar to human cognition, assist the physician in real-time precision medical diagnosis (Iqbal et al. 2021), provide clinical decision support to reduce diagnostic and therapeutic errors in clinical practice (Jiang et al. 2017).

On the other hand, machine learning (ML) is a subset of AI. ML algorithms are used to extract interesting information based on data trends and make predictions from the complex clinical data itself (Alpaydin 2020; Kourou et al. 2015). The potential of big data has been exploited using a ML technique to find previously hidden insights in medical information (Mostafa et al. 2021) and aid in the decision-making process (B. Zhang et al. 2023) with more accurate and personalized information on diagnosis and treatment. Supervised machine learning is the typical ML technique used in the AI healthcare applications for disease prediction while unsupervised machine learning

such as principal component analysis mostly used in the data preprocessing to reduce the dimensionality while keeping the information of the features (Jiang et al., 2017). ML techniques such as decision tree (DT), gradient boosting (GB), naïve bayes (NB), k–nearest neighbor (KNN), logistic regression (LR), random forest (RF), support vector machine (SVM) and neural networks (NN) have shown promising results in terms of cancer prediction and diagnosis in recent research which will be discussed in Chapter 2.2. Besides that, artificial neural network (ANN) are one of the reliable ML real-time screening tools used to detect high-risk individuals, such as in colorectal cancer, with low misclassification rate in 6% of positive cases misclassified as low risk and 2% of negative cases misclassified as high risk (Nartowt et al. 2020), while pancreatic cancer achieved a sensitivity of 80.7% in the risk assessment (Muhammad et al. 2019).

However, despite the advancements in ML for cancer prediction, which have resulted in higher predictive accuracy, the risk of individuals diagnosed with life-threatening cancer is still growing today, including hepatocellular carcinoma (HCC), which will be discussed in section 1.2. In particular, ML risk prediction models such as ANN, SVM and RF are seldom deployed into clinical practice for decision making due to their *black box* nature and lack of interpretability(Ahmad et al. 2018).

Furthermore, there is no consensus on standard and unified guidelines among the healthcare community in the diagnosis process. It is challenging for the practitioner to make an early diagnosis of HCC due to the absence of symptoms in the early stages. However, this issue could be resolved through regular health screening, including the liver function test (LFT) to detect HCC at the earliest and most treatable stage, especially for those in the high-risk category. Therefore, HCC risk prediction remains a promising research direction that aims to enhance the current ML model by improving interpretability with a degree of confidence.

1.2 HEPATOCELLULAR CARCINOMA

The liver is the largest organ in the human body which is located below the diaphragm on the right side of the abdomen as shown in Figure 1.1. It is responsible for many

important metabolic functions, it filters toxics from the blood, detoxifies chemicals and metabolizes drugs. Liver responsible in the secretion of bile, which effectively removes waste products from the liver and exits the body through urine or feces. There are two forms of liver cancer: primary and metastatic. Primary liver cancer (PLC) is a type of cancer that originates in the liver, whereas metastatic (secondary) liver cancer is cancer that has progressed to the liver from somewhere else in the body. PLC can be categorized into HCC, cholangiocarcinoma (also known as bile duct cancer) and angiosarcoma (Anon. 2022). HCC is the most common type of PLC accounting for 75 to 85% of the total liver cancer burden worldwide (Bray et al. 2018), ranks sixth globally in cancer incidence and third in cancer-related deaths in 2020 (Chakraborty & Sarkar 2022).

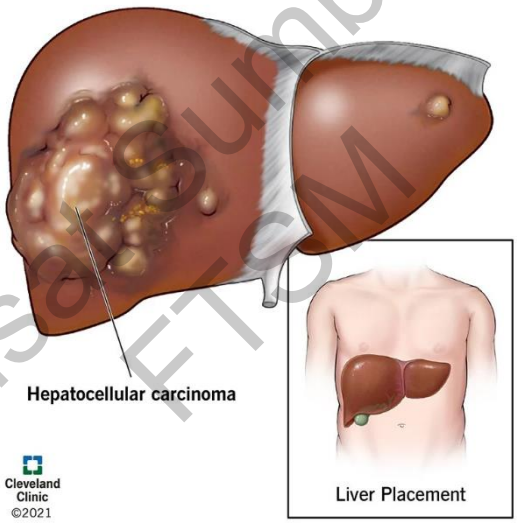

Figure 1.1 Liver placement and HCC

Source: Anon. 2021

According to the statistics, over half of the global liver cancer cases and mortalities were concentrated in Eastern Asia with China accounting for 45.3% of cases and 47.1% of deaths caused by liver cancer (Rumgay et al. 2022). A recent study published by the International Agency for Research on Cancer (IARC) reveals that in 2020, 905,700 people were diagnosed with liver cancer, resulting in 830,200 deaths worldwide (Rumgay et al. 2022). IARC projects an alarming trend in liver cancer and predicts new cases and fatalities might increase by more than 55% by 2040, totaling 1.4

million diagnoses and 1.3 million deaths, if the current trend remains (Rumgay et al. 2022).

In Malaysia, HCC is one of the major causes of the mortality rate which increased by 31.5% since 1990 (Raihan et al. 2018). Data samples collected from 2006 to 2009 at the University Malaya Medical Center on 348 HCC patients profiling by Goh et al. showed that males are at higher risk than female to be diagnosed as HCC with the ratio of 3.4:1 while Chinese (68.7%) are the most afflicted by HCC among 3 races, followed by Malays (20.4%) and Indian (10.9%) with the median age of 62.5 years (B. Norsa'adah 2013)(Goh et al., 2015).

According to the Malaysia National Cancer Registry's (MNCR) quinquennial report, males are more likely than females to be diagnosed with liver cancer as shown in Figure 1.2 (Ministry of Health 2019a). The incidence pattern rate of liver cancer among men in 2012-2016 is comparable to that of 2007-2011, with a positive indicator that the rate of all age groups is decreasing as in Figure 1.3 (Ministry of Health 2019b). However, the analysis showed that the cases diagnosed at stage 3 and stage 4 increased from 78.8% in 2007-2011 to 85.4% in 2012-2016 as in Figure 1.4 (Ministry of Health 2019b). Many studies have further justified that men have a significantly higher likelihood of developing HCC compared to women. For instance, in the USA, the male-to-female ratio is approximately 2.5-3 to 1, whereas in some regions worldwide, this ratio can be up to 6 males to 1 female (Sayiner et al. 2019).

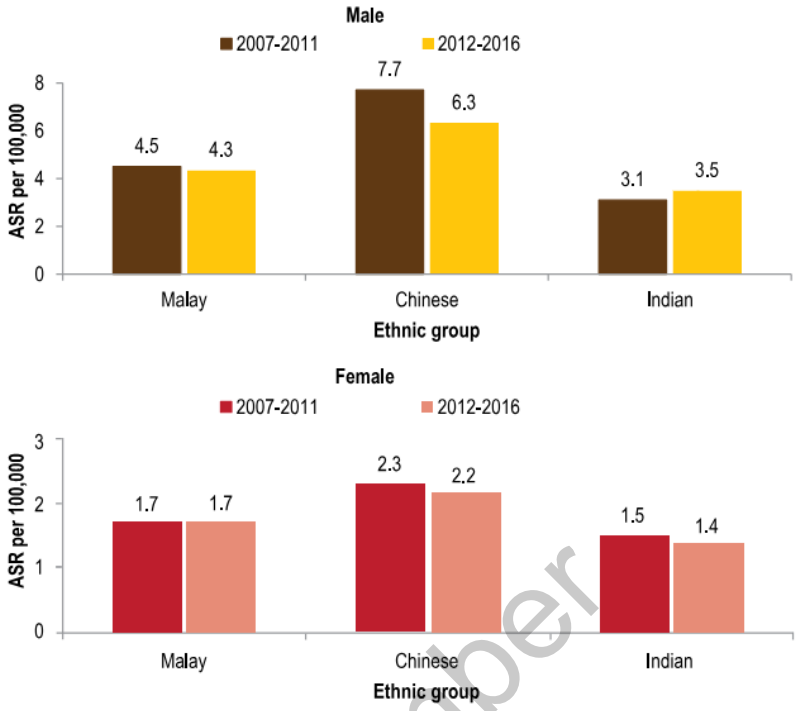

Figure 1.2 Age-standardized incidence (ASR) rate by year, major ethic group and sex in Malaysia, 2007-2011, 2012-2016

Source: Ministry of Health 2019a

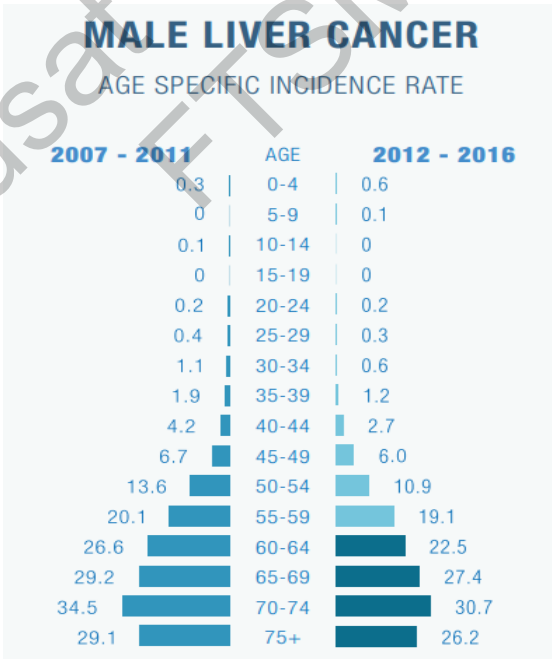

Figure 1.3 Age specific incidence rate in male liver cancer in Malaysia 2007-2011, 2012-2016

Source: Ministry of Health 2019b

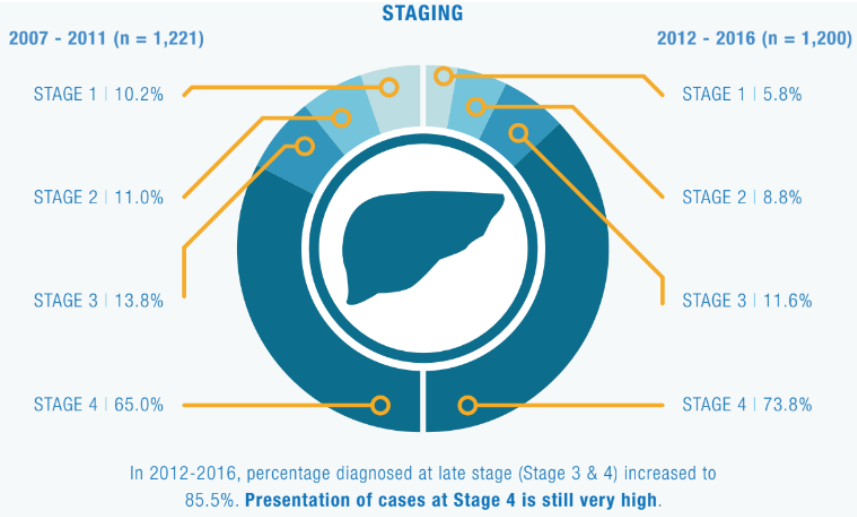

Figure 1.4 Staging in male liver cancer in Malaysia 2007-2011, 2012-2016

Source: Ministry of Health 2019b

The survival rate of liver cancer is relatively low even in high-income countries like Australia, Denmark, Ireland, New Zealand, Canada, Norway, and United Kingdom with highest 1 year and 3-years net survival rate based on the registration on International Cancer Benchmarking Partnership from 1995-2014 (Rutherford et al. 2021). The low long-term survival rate of liver cancer is largely attributed to delayed diagnosis due to asymptomatic early stages, resulting in most patients being diagnosed at an advanced stage with distant metastases, emphasizing the crucial need for improved risk assessment at the early stages (Ginès et al. 2004).

1.2.1 Causes of Hepatocellular Carcinoma

HCC begins in cells called hepatocytes (Anon. 2024) primarily afflicting individuals with pre-existing chronic liver conditions, especially fibrosis and cirrhosis, which are commonly caused by chronic liver inflammation, as shown in Figure 1.5. Extensive scarring of the liver (fibrosis) will eventually cause cirrhosis. In other words, when the liver is inflamed, it looks for ways to heal itself, but this causes the formation of scars (fibrosis), eventually causing irreparable liver damage. Hepatitis B viruses (HBV), hepatitis C viruses (HCV), aflatoxin B1 exposure, excessive alcohol use, and diseases such as diabetes and obesity are all major causes of chronic liver inflammation.

Additionally, aflatoxin B1 exposure, excessive alcohol consumption, and conditions like diabetes and obesity are significant factors contributing to chronic liver inflammation. The relative risk of developing malignancy is highest in those patients with chronic HBV (56%) while HCV with relatively less (20%) (Maucort‐Boulch et al. 2018). Besides that, aflatoxins are powerful carcinogens found on moldy crops like peanuts, corn, and other nuts and seeds, especially when they are stored in warm and humid conditions. High rates of HCC are frequently associated with extensive aflatoxin exposure, and work in combination with chronic HBV infection. (Gouas et al. 2009).

On the other hand, there is substantial evidence to prove that alcohol intake, obesity, and type 2 diabetes work together to elevate the risk of HCC (Bertot & Adams, 2019; Hassan et al. 2002; Marrero et al. 2005). The rising epidemics of obesity and diabetes have been observed in Asia and Western countries. The growth of obesity, diabetes, and, consequently, Non-alcoholic fatty liver disease (NAFLD) in Asians is being caused by the shift towards a sedentary lifestyle and dietary habits leaning towards overnutrition (Bertot and Adams 2019; Hashimoto et al. 2012; Okanoue et al. 2011; Wong et al. 2011). NAFLD is identified by the presence of excess fat accumulated in the liver (steatosis) without any inflammation or damage to the liver cells.

On the other hand, non-alcoholic steatohepatitis (NASH) is indicated by liver inflammation and the potential for development to fibrosis, cirrhosis, and, finally, HCC. The global prevalence of NAFLD including its advanced form, NASH in HCC is becoming significant (Ozakyol 2017) even though the current burden of HCC worldwide focuses on HBV and HCV (Bertot & Adams 2019). Several studies have shown that HBV, followed by cryptogenic causes such as NAFLD and NASH is the predominant cause of HCC in Malaysia which is closely related to the increasing rate of obesity and diabetes (B. Norsa'adah 2013; Goh et al. 2015; Raihan et al. 2018).

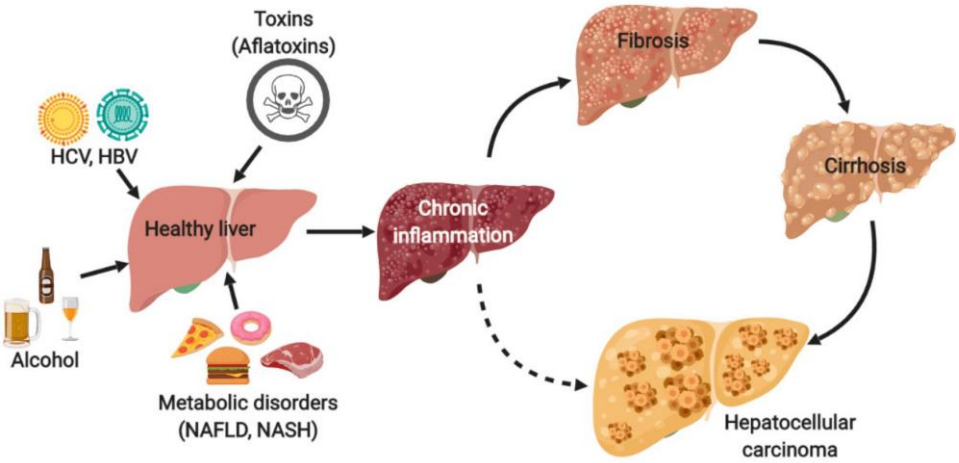

Figure 1.5 Development of HCC

Source: Macek Jilkova et al. 2019

1.3 DIAGNOSIS AND TREATMENT OF HEPATOCELLULAR CARCINOMA

The diagnosis of HCC has been mainly based on blood screening, imaging tests such as ultrasound, computed tomography (CT) scan and magnetic resonance imaging (MRI) scan and also procedure like liver biopsy. The procedure for diagnosing and treating HCC as shown in Figure 1.6.

First and foremost, the routine liver function test (LFT) is the fundamental screening for the early detection of any kind of liver damage and liver disease. LFT included parameters such as total bilirubin, direct bilirubin, alkaline phosphatase (ALP), alanine aminotransferase (SGPT), aspartate aminotransferase (SGOT), albumin, total protein and albumin/globulin ratio (A/G Ratio). Each of the components in LFT plays a crucial role that aids the doctors in monitoring and evaluating the patients' liver conditions. If a patient has an abnormal range of LFT, the practitioners will conduct a risk assessment to identify those who are at high risk of developing HCC.

Patients with cirrhosis, chronic HBV or HCV, a family history of liver disease, excessive alcohol consumption, obesity, and diabetes were among the high-risk groups. These high-risk populations will proceed to imaging diagnosis. For those patients who are not in the high-risk group will have a follow-up with their doctor. The decision on whether to proceed with image diagnostics in low-risk patients depends on the clinical judgement during the follow-up.

Following the risk assessment, an abdominal ultrasound test will be conducted as the following step in HCC diagnosis. When there is a suspicious lesion on the ultrasound. Further investigation such as CT or MRI scan is required to provide more detailed imaging of the tumour. However, if the imaging result is still inconclusive, the invasive approach of liver biopsy needs to be applied. Subsequently, staging and treatment options for HCC will be based on the imaging or liver biopsy result.

Tumour staging and treatment decisions of HCC patients are made based on the Child-Pugh score and Barcelona Clinic Liver Cancer (BCLC). The Child-Pugh score is used to predict mortality during surgery and assess the severity of liver disease in patients with liver cirrhosis (Pugh et al. 1973). The score considers five essential criteria, three of which measure the synthetic function of the liver (bilirubin, albumin, and prothrombin) and two of which are based on clinical assessment (ascites and encephalopathy). Child-Pugh grade is calculated by summing the scores for each criterion, with Grade A representing least severe liver disease (5-8 points), Grade B representing moderately severe liver disease (9-11 points) and Grade C representing most severe liver disease (12-15 points) (Durand & Valla 2005).

Furthermore, BCLC is a classification system that seeks to determine patient prognosis, as well as recommends specific treatment algorithms based on HCC tumour stage, liver function and patient performance (Reig et al. 2022). BCLC applied the treatment stage migration (TSM). TSM emphasizes customizing treatment based on patient profile; when the standard first-line treatment is unfeasible, practitioners may deviate from the standard recommendations on BCLC, transitioning to a more advanced stage treatment (Reig et al. 2022; Tsilimigras et al. 2022).

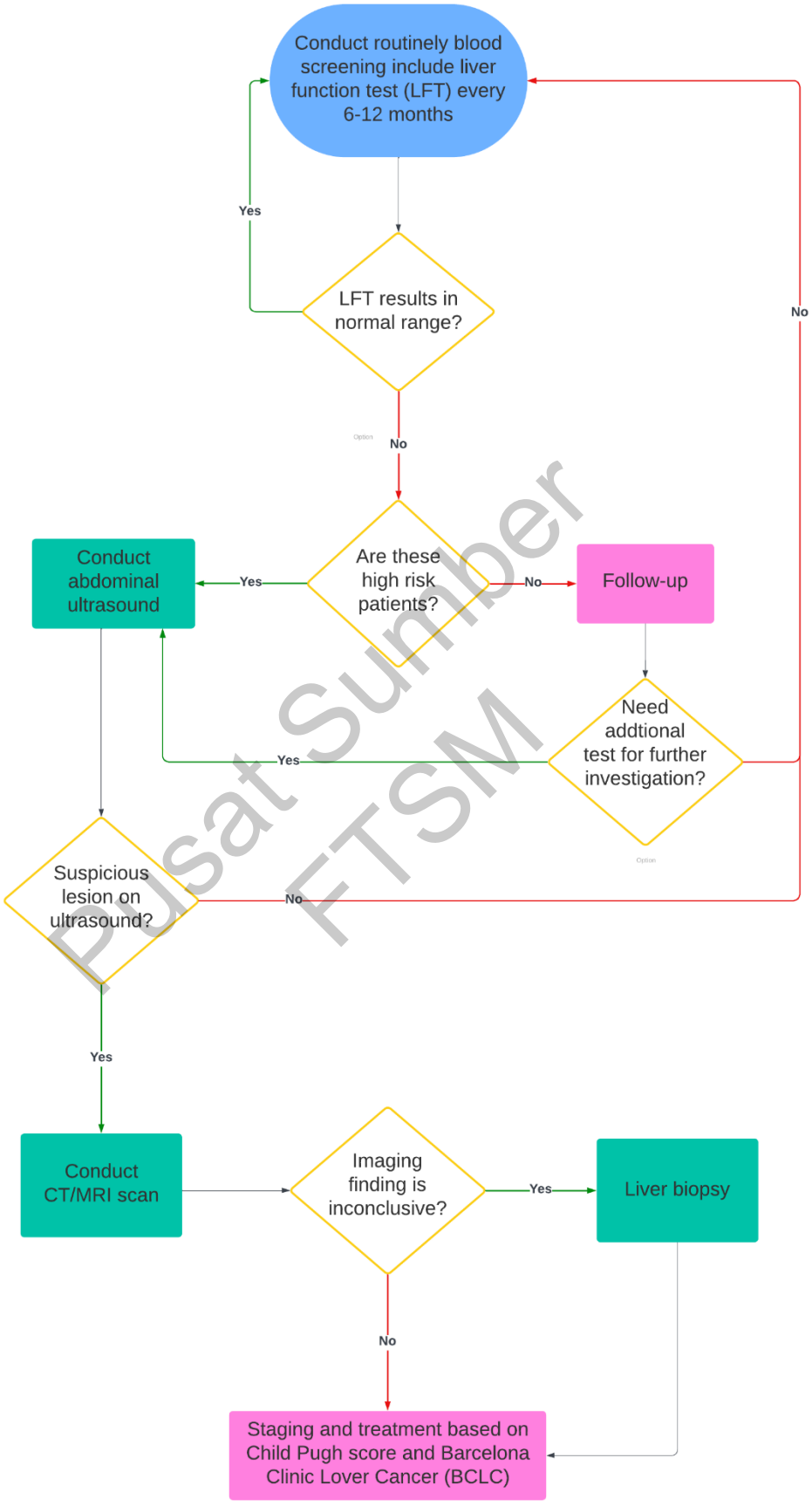

Figure 1.6 Flow diagram on diagnosis and treatment of HCC

1.4 PROBLEM STATEMENT

HCC is a highly fatal tumour and remains a global health concern. It is challenging for physicians to make an early diagnosis of HCC due to the absence of symptoms in the early stages. The overall HCC incidence is predicted to rise in the coming years due to demographic factors in terms of population growth and aging (Akinyemiju et al. 2017). There are a growing number of HCC risk predictions using ML techniques, but they exhibit limitations in terms of robustness and interpretability. In the medical field, lack of interpretability is the bottleneck for clinicians to use ML models as it may cause adverse consequences if the predictive output is not explained accurately (Ahmad et al. 2018). Therefore, enhancing the existing ML predictive model with a degree of confidence in assessing the risk of HCC development is necessary. Each of the features and the predictive output should be comprehensible to the targeted end-user which are the physicians from a domain perspective.

The use of fuzzy logic (FL) as a reasoning tool are known in manufacturing and robotics, but not yet in medical. Therefore, a FL approach could be applied to ML models to address the uncertainties in the classification results. FL acts as a decision support tool, incorporating a clinical prediction rule to improve HCC risk management more effectively. Moreover, FL provides valuable flexibility for reasoning to solve the interobserver agreement among doctors, as clinical decisions are often taken based on doctors' perceptions and experiences. Additionally, FL improves the diagnostic precision in the risk classification model.

1.5 RESEARCH OBJECTIVES

1. To suggest the optimal ML model for HCC risk classification.

2. To optimize the HCC risk classification model.

3. To improve the explainable component of HCC risk classification model with FL.

1.6 RESEARCH SCOPE

This study focused on developing an HCC risk prediction model utilizing blood test results of liver function such as total bilirubin, direct bilirubin, ALP, SGOT, SGPT, albumin, total protein and A/G Ratio, and demographic data such as age and gender. In addition, the research adopts a rule-based system to enhance the prediction with confidence level. The study will utilize an existing dataset from Kaggle to evaluate the proposed approach.

1.7 RESEARCH ORGANIZATION

Chapter 1 introduces the research background of ML in cancer research and HCC with their causes, diagnosis, and treatment process. It also includes the problem statement, research objective, and research scope.

Chapter 2 reviews the relevant past research on HCC risk classification. It covers the rule-based and non-rule-based techniques used in the paper, evaluation metrics to assess the model performance, challenges in the ML classification models, and explores the FL approach to enhance the interpretation of classification outcomes.

Chapter 3 discusses in detail the application of data preprocessing methods to prepare the data for modelling.

Chapter 4 explains in detail the process of data modelling. It covers model training, testing, hyperparameter tuning, evaluation of the model performance, and rules extraction from the best model.

Chapter 5 focuses on the logical reasoning behind the classification results with the FL approach. It provides a detailed explanation of procedures for FL.

Chapter 6 is the summary of the research work with their outcomes. It also addresses the limitations encountered and future works to further enhance the current proposed model.

1.8 CHAPTER SUMMARY

Regular health screening, including the liver function test (LFT) is one of the approaches for all individuals to detect HCC at the earliest and most treatable stage, especially for those in the high-risk category. Nevertheless, the existing classification models often lack interpretation of the outcome. Therefore, this research aims to improve the current machine learning model by using an FL technique to offer understandable predictive results along with a level of confidence.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

Supervised ML is frequently used for disease classification problems (Sharma & Rani 2021) such as classifying tumours based on clinical criteria. Supervised ML allows the model to learn through the labelled data and train the algorithm to accurately predict the output for the new dataset based on the "reference" which is the generalized pattern discovered during the learning process. The application of ML could help practitioners detect symptoms of HCC earlier and make more accurate diagnoses.

In this chapter, we discuss the past and relevant studies on HCC risk classification models using clinical data such as LFT and other variables. A variety of variables have been researched to equip clinicians with a reliable tool for HCC early detection. The parameters used are those easily obtained from routine laboratory data such as alpha-fetoprotein (AFP), albumin, platelet, total bilirubin, alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), aspartate transaminase (SGOT), and others (Feng et al., 2023). Lastly, the chapter concludes the ML techniques with other approaches that will be used in this project.

2.2 PAST RESEARCH ON CLASSIFICATION OF HEPATOCELLAR CARCINOMA RISK

Książek et al. (2020) proposed a novel ML model for HCC early detection that uses ensemble learning to merge seven models into one, including KNN, RF, NB, LR, and three additional classifiers. Książek et al. trained the models based on 165 HCC patients with 49 clinical features including the LFT, demographic, and other variables. Their approach achieved an accuracy of 0.9030 and an F1-score of 0.8857 (Książek et al.

2020). On the other hand, the same dataset has been applied in the study by Santos et al. (2015) to predict the survival rate among HCC patients. LR and NN were tested, and the results showed that NN performed better than LR in the synthetic minority over-sampling technique (SMOTE) oversampling approach with an accuracy of 0.717 and 0.706 respectively.

Two studies have reported that the RF classifier has the best performance among other algorithms with accuracy = 0.762, recall = 0.843, F1-score = 0.775, and AUC = 0.999 in (Ding et al. 2022) and an accuracy score of 98.14% in (Mostafa et al. 2021). Ding et al. used ML algorithms such as regularized regression, LR, RF, DT, and extreme gradient boosting (XGBoost) to identify 14 significant risk factors based on the basic blood tests of 525 patients. The risk factors are listed in descending order based on their importance: total bilirubin, GGT, direct bilirubin, haemoglobin, age, platelet, ALP, SGOT, creatinine, SGPT, cholesterol, albumin, urea nitrogen, and white blood cells (Ding et al 2022). Similarly, the study by Mostafa et al. applied SVM, RF, and ANN and identified five crucial indicators: AST, ALT, GGT, bilirubin, and ALP in predicting liver disease based on the blood results from 615 individuals.

Furthermore, Sato et al. (2019) developed an HCC predictive model based on clinical data collected at the University of Tokyo Hospital. The study involved 4242 patients from two groups: those diagnosed with HCC initially and those who tested positive for HBV and later developed HCC. The clinical data examined included the biomarkers of liver inflammation, liver fibrosis, liver function, and hepatitis virus status (Sato et al. 2019). Various algorithms like LR, SVM, RF, GB, NN, and deep learning were employed. The result showed that GB achieved the highest accuracy (87.34%) and area under the curve (AUC) of 0.940 (Sato et al. 2019).

An et al. (2021) developed an ML-based model to predict the risk of HCC development in the Korean cohort using health screening examination results. An et al. identified several predictors associated with increased or decreased risk, including age, sex, obesity, LFT, family history, chronic liver diseases, and other health condition. ML models used such as Random survival forest (RSF) and XGBoost while XGBoost

showed promising results with an AUC of 0.882 and a standard deviation of 0.013 (An et al. 2021).

In addition, Chicco and Oneto (2021) used ML methods such as DT, RF, SVM and multilayer perceptron network with dropout (MLP) to predict survival and identify key clinical factors for HCC from 165 patients with 50 features and concluded that ALP, AFP, haemoglobin to be most crucial predictors. RF achieved accuracy of 0.772 and AUROC of 0.766. However, there are some drawbacks in the research, included only using the data from a hospital and lacking survival time feature (Chicco & Oneto, 2021). Therefore, their future plans involve validating findings with alternative datasets and applying the approach to other diseases and high-throughput sequencing data (Chicco & Oneto, 2021).

Wong et al. (2022) reported that a novel HCC-ridge score model correctly predicted chronic viral hepatitis patients based on a large cohort of 124006 patients in Hong Kong. RF with AUROC of 0.992 outperformed LR, adaptive boosting (AdaBoost), DT, and ridge regression. It is further validated with an external cohort of 4462 Korean patients with an AUROC result greater than 0.8 for all the models except the DT (0.799) result (Wong et al. 2022). Nonetheless, it is challenging to deal with a large amount of data from multiple centres because there consists of lots of missing data and inconsistent intervals of the laboratory measurements which may lead to bias (Wong et al. 2022).

Besides that, clinical data is useful not only for developing a model for HCC early detection and prediction but also for predicting recurrence after surgical resection. Y. Huang et al. (2021) carried out research that utilized clinical information from 7919 post-hepatectomy patients from 2 different hospitals which encompassed demographic details, blood test findings like AFP, GGT, total bilirubin, albumin, hepatitis virus indicator, tumour traits, and additional parameters. Y. Huang et al. utilized a heat map to personalize recurrence risk and identified key prognostic variables specific to different time intervals after surgery. Based on the ML results on Deep Learning-based Survival Model, XGBoost, and RSF. They found that XGBoost to be the most effective with c-index: 0.713, $P < 0.05$.

Table 2.1 Summary of literature on HCC risk classification

Authors	Data Source	Classifiers	Results
Książek et al.	165 HCC patients with 49 clinical features including the LFT, demographic, and other variables.	Ensemble of 7 classifiers such as KNN, RF, NB, LR, and other 3 classifiers.	Accuracy of 0.9030 and F1-score of 0.8857.
Książek et al.	165 HCC patients with 49 clinical features including the LFT, demographic, and other variables.	LR and NN.	Accuracy of 0.717(NN) and 0.706 (LR).
Ding et al.	Basic blood tests of 525 patients.	Regularized regression, LR, RF, DT, and XGBoost.	RF achieved accuracy = 0.762, recall = 0.843, F1-score = 0.775, and AUC = 0.999.
Mostafa et al.	Blood results from 615 individuals.	SVM, RF, and ANN	RF with highest accuracy score of 98.14%
Sato et al.	Blood results of 4242 patients from two groups: those diagnosed with HCC initially and those who tested positive for HBV and later developed HCC.	LR, SVM, RF, GB, NN, and deep learning.	GB achieved the highest accuracy 0.8734 and AUC of 0.940.
An et al.	Health screening results of Korean cohort.	RSF, and XGBoost.	XGBoost showed promising results with an AUC of 0.882 and a standard deviation of 0.013.
Chicco and Oneto	165 HCC patients with 50 features	DT, RF, SVM, and MLP.	RF achieved accuracy of 0.772 and AUROC of 0.766.
Wong et al.	124006 chronic viral hepatitis patients. External cohort of 4462 Korean patients.	RF, LR, AdaBoost, DT, and ridge regression.	RF with AUROC of 0.992. External validation with AUROC result greater than 0.8 for all the models except the DT (0.799).
Y. Huang et al.	Blood test results of 7919 post-hepatectomy patients.	Deep Learning-based Survival Model, XGBoost, and RSF.	XGBoost with c-index: 0.713, $P < 0.05$.

2.3 RULE-BASED ML TECHNIQUES

Based on the past research ON HCC classification discussed in Chapter 2.2. We notice that rule-based classifiers such as DT and RF are applied in most of the studies. Thus, we will discuss each of the rule-based techniques.

2.3.1 Decision Tree Classifier

DT classifier is frequently used in classification systems to attribute type information, as well as predictive systems, where the predictions are based on previous data and contribute to drive the structure of the decision tree and the output (Saouabi & Ezzati 2020). DT requires minimal data preparation; it can handle both numerical and categorical data and is easily interpreted. A DT algorithm is used in classification by dividing the data into classes consisting of three components root node, branch (edge or link), and leaf node. DT starts from the root node at the top of the tree and is split into branches that contain all possible outcomes for the test and then further divided into leaf nodes containing the label of the class to which it belongs.

Techniques such as information gain, Gini index, and entropy are used to split the data into different nodes (Sen et al. 2020). Entropy represents randomness in features, it measures the discriminatory capability of an attribute for classification problems (Sen et al. 2020). Information gain measures the expected reduction in entropy (uncertainty) by calculating the difference of the entropy at the parent node and the entropy of the weighted average of the child node (Krishnan 2021).

2.3.2 Random Forest Classifier

RF classification makes predictions using an ensemble of decision trees (Breiman 2001). It is built by taking random samples of the actual data and then constructing an ongoing series of decisions trees on the subsets and lastly aggregate the results to predict each observation (Speiser et al. 2019). RF collects class votes from each tree and uses a majority vote for classification (Hastie et al. 2009). In RF, the more the decision trees are used with different criteria, the better the RF performs. As a result, RF usually has high predicted accuracy when compared to other models while also reduce overfitting in the training data.

2.4 NON-RULES-BASED ML TECHNIQUES

Non rules-based model are also popular among the disease classification such as SVM, LR, KNN, NB, GB and NN.

2.4.1 Support Vector Machine Classifier

The SVM algorithm is widely used in classification tasks for linear and non-linear data. SVM creates the optimal line or decision boundary known as a hyperplane that can segregate data points in n-dimensional space into two classes in training data, allowing the new data point can be easily classified in the correct category in the future (Sen et al. 2020). The maximum margin creates the greatest possible distance between the separating hyperplane and the nearest instances from either side (Kotsiantis et al. 2007). Support vectors, also known as extreme points, are the data points that lie on the optimal hyperplane as shown in Figure 2.1.

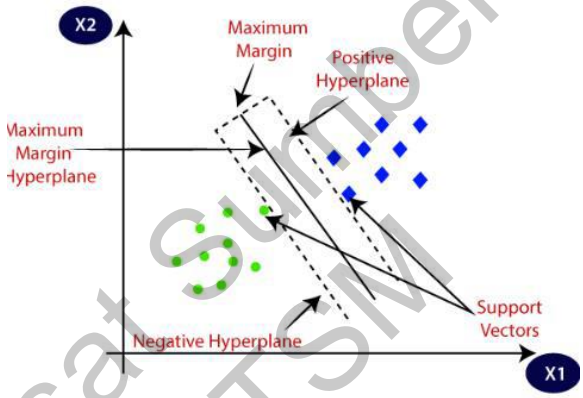

Figure 2.1 Hyperplane in SVM

Source: Saini 2024

SVM is suitable for the large dataset with many features as the SVM only chooses the minimum number of support vectors, hence, the complexity of the SVM is unaffected by the number of features in the training samples. However, the real-world data are always complex and noisy, this will cause a situation where SVM is unable to find a hyperplane to separate the classes correctly. Consequently, this led to misclassifying the instance where the data points are located on the wrong side of the hyperplane. Thus, a more flexible approach called soft margin is introduced to achieve a balance between the maximal margin and some misclassifications on the training data (Kotsiantis et al. 2007).

Besides that, a kernel function is used if the data is not linearly separated. Kernel function transforms the original feature space (non-linear) into a higher dimensional

space where data might become linearly separated (Kotsiantis et al. 2007). By applying kernel function onto non-linear data, SVM can define the optimal hyperplane for classification.

2.4.2 Logistic Regression Classifier

LR is one of the regression methods in classification for predicting the probability of presence or absence of a dichotomous (binary) dependent variable based on one or more predictor independent variables (Kurt et al. 2008).

LR calculate the coefficient (weights) and the intercept (bias term) to linearly combine the input features in the training data. Each weight, w_i is a real number associated with a specific input feature, x_i representing its importance in the classification decision. The bias term, β_i is another real number added to the weighted inputs. The classifier multiplies each feature by its weight, sums up the weighted features, and adds the bias term. The resulting number z represents the weighted sum of evidence for the class is as Equation 2.1.

$$Z = \beta_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \quad (2.1)$$

Maximum likelihood estimation is used estimate the optimal parameter (weights and bias term) in the LR, which maximize the likelihood of observing the data given the model (Hosmer et al. 2013). Then, the probability of the output prediction is determined using logistic function also known as sigmoid function. The s-shaped sigmoid curve fits the predicted probability in the desired range between 0 and 1. Lastly, the classification of the outcome is based on the decision threshold or boundary by comparing the predicted probability to a threshold.

In medical diagnosis, LR can determine what has an influence on whether a certain disease is present or not. For example, we could study the influence of independent variables such as age, gender, and laboratory results on that particular disease and predict how likely a person will have a certain disease.

2.4.3 K-Nearest Neighbour Classifier

KNN is known as a lazy learning algorithm as the generalization beyond the training data is delayed until a new instance is provided to the system (Kotsiantis et al. 2007). It is a pattern recognition method that finds the k closest relatives in future cases by using training datasets. When k = 1, the class of the training tuple closest to the unknown tuple in pattern space is assigned to it. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. Cross-validation is commonly used to find the value of k. KNN is easy to implement and is effective, especially for large training data that contain lots of noise. However, it is quite time-consuming as the calculation of the distance from k neighbours needs to repeat for every new instance.

2.4.4 Naïve Bayes Classifier

NB classifier is a simple probabilistic classifier that uses the Bayes theorem by strongly assuming each attribute variable as an independent variable (Rish 2001). The classification is done by the Bayes principle to calculate the probability of class name Y, given that the particular instance X by the formula as in Equation 2.2.

$$\mathrm{P}\,(\mathrm{Y}|X) = \frac{P\,(X\,|Y) * P\,(Y)}{P\,(X)} \tag{2.2}$$

NB is an effective method because it focuses on identifying the most probable class rather than perfectly modelling the underlying distribution, allowing it to perform well even in scenarios where features are dependent on each other (Rish 2001). Besides that, NB remains strong even when dealing with missing attributes, as it considers all attributes when making predictions and this leads to a gradual decline in performance rather than a sudden drop (Webb et al. 2010).

Furthermore, Gaussian Naïve Bayes (GNB) is the extension of NB. The GNB is suitable for continuous input data that follow a normal or Gaussian distribution where the mean and standard deviation are calculated based on each class (Kamel et al. 2019).

The GNB is effective in supervised learning and complex real-world situations in medical diagnosis (Kamel et al. 2019).

2.4.5 Gradient Boosting Classifier

GB is a well-known algorithm used for classification tasks. GB uses ensemble methods to combine several weak learners into a strong learner through iteration (Bentéjac et al. 2021). It is called GB because it uses gradient descent to minimize the loss when adding new learners to the ensemble. However, the GB classifier may face overfitting issues if the iterative procedure lacks proper regularization (Friedman 2001).

The cycle of GB as illustrated in Figure 2.2. The process start with the prediction based on the naïve model. Then, the gradient of the loss function such as mean squared error is calculated based on the current prediction. Subsequently, a new weak model will be trained based on the calculated loss from the previous prediction. After that, the new model is added to ensemble and make the prediction again. This cycle continues as more models are added to the ensemble to enhance the prediction.

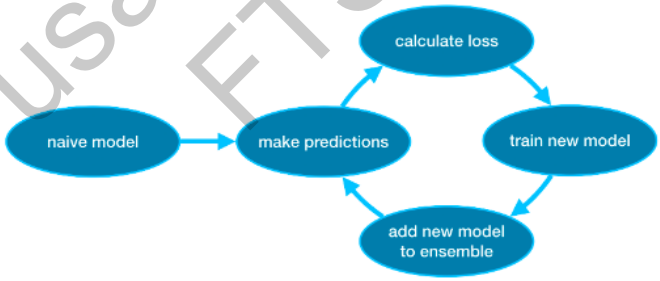

Figure 2.2 GB cycle

Source: Cook 2018

On the other hand, XGBoost is the most popular implementation of GB, it is the top winning solution for many Kaggle competitions (Chen & Guestrin 2016). XGBoost is a decision tree ensemble based on a GB designed to push the extreme of the computation limits of machines to be highly scalable and accurate (Chen & Guestrin 2016). The XGBoost apply regularization technique such as L1 and L2 to avoid overfitting.

2.4.6 Neural Network: Multilayer Perceptron Classifier

The MLP is one of the widely used feedforward NN models (G. P. Zhang 2000). An MLP consists of one input layer, one or multiple hidden layers, and one output layer as shown in Figure 2.3. Backpropagation is a popular supervised learning algorithm for training feedforward NN. Backpropagation is a learning approach that includes iteratively processing a dataset of training tuples, comparing the network's prediction to the actual known target value, and adjusting weights to reduce the mean-squared error between predictions and the target value (Han et al. 2012).

The process starts with the input x arriving through the preconnected path. The input is modelled using randomly selected real weights, W. Next step is to use the activation function to calculate the output for every neuron that passed through the input layer to the hidden layers, to the output layer. Then, error (Error = Actual Output – Desired Output) is calculated in the output using the loss function such as mean-squared error. Lastly, travel back from the output layer to the hidden layer to adjust the weights such that the error is decreased to reduce network error.

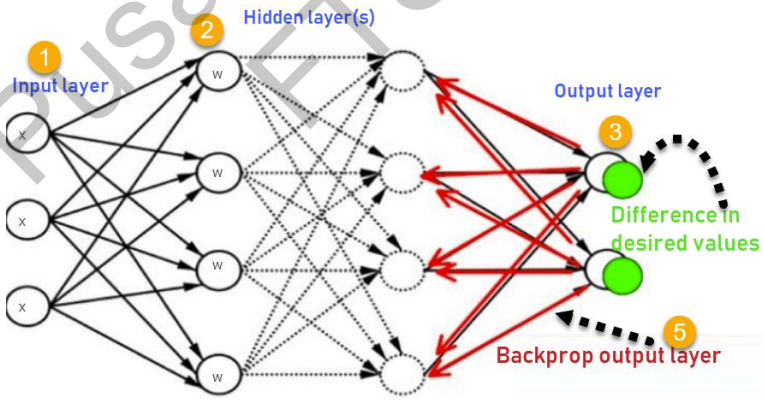

Figure 2.3 Backpropagation in MLP

Source: Johnson 2024

2.5 EVALUATION METRICS IN CLASSIFICATION MODEL

Evaluation metrics such as accuracy, F1-score, recall, precision, confusion matrix and area under the receiver operating characteristic (AUROC) will be used to evaluate the performance of the trained models. While the most effective predictive model is chosen among those with the best overall performance.

1. Accuracy: Total number of the correctly predicted class divided by the total instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.3}$$

2. F1 Score: Combined measure of precision and recall, providing a balanced assessment.

$$\text{F1-score} = 2 \text{ x } \frac{Precison\ x\ Recall}{Precision + Recall} \tag{2.4}$$

3. Recall: Also known as sensitivity or true positive rate, measures the total number of actual positive class that are correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP+ FN} \tag{2.5}$$

4. Precision: Ratio of predicted positive instances that are actual positive to the total number of positive instances predicted by the model.

$$\text{Precision} = \frac{\text{TP}}{TP+ FP} \tag{2.6}$$

5. Confusion Matrix: A thorough summary table as shown in Figure 2.4, consists of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). It is useful for visualizing a model's accuracy, f1-score, recall and precision.

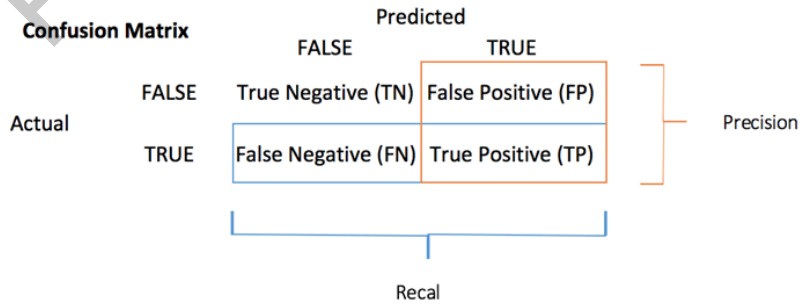

Figure 2.4 Confusion matrix

Source: Arora 2019

6. AUROC is the area under the ROC curve. ROC curve as shown in Figure 2.5 is a comprehensive performance measure that evaluates the accuracy of a model's predictions across all potential classification thresholds: true positive rate (TPR)

and false positive rate (FPR). AUROC is useful in binary classification where it can distinguish between different classes.

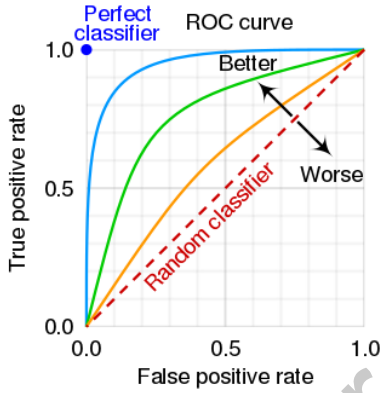

Figure 2.5 ROC curve

Source: Anon 2024

2.6 CHALLENGES IN HCC RISK CLASSIFICATION MODELS

According to Calderaro et al. (2022), the current AI algorithms suffer from significant drawbacks, such as limited interpretability of results, overfitting, and potential poor generalization due to their reliance on training data size and diversity. Besides that, the healthcare sector is currently still reserved in providing incentives for data exchange across different hospitals (Jiang et al. 2017), making obtaining of large-scale real-world datasets difficult. This might be to ensure the confidentiality and privacy of patients.

Furthermore, despite the increasing number of AI studies in healthcare, the implementation of these ML models in clinical practice remains limited due to a lack of comprehension of the predictions. The adoption of an interpretable ML model is important in healthcare as it ensures clinicians understand how the predicted outcomes relate to each of the attributes in the ML models. Moreover, interpretable ML models provide clinicians with reasons to accept or reject forecasts and recommendations by explaining the logic behind them. This is necessary as decisions made by healthcare providers directly impact the well-being of individuals, highlighting the need for thoughtful consideration.

2.7 FL APPROACH

In order to enhance the interpretability of the classification model, we could apply FL. FL was first proposed by Lotti Zadeh in 1965, it is an extension from the traditional Boolean logic with binary true (1) or false (0). FL provide a way to model logical reasoning with a degree of truth that ranges from 0, which is absolutely false, to 1, that is absolutely true. Fuzzy refers to something which is unclear or vague, while FL allows us to design a fuzzy inference system, which is a function that maps a set of inputs to output using human-interpretable rules rather than more abstract mathematics. The architecture of FL system as shown in Figure 2.6.

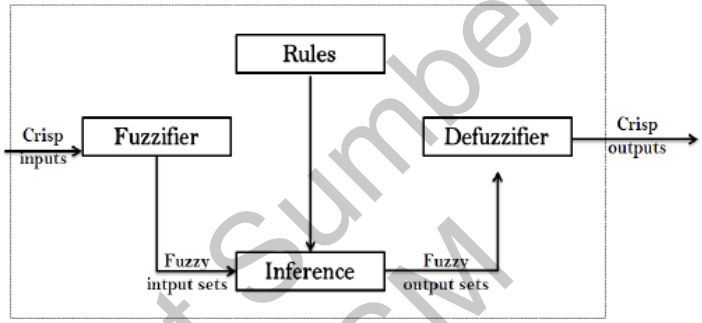

Figure 2.6 FL architecture

Source: Kumar 2019

A FL system is made up of 4 basic components which are fuzzification, rule base, inference engine and defuzzification. Fuzzification is used to transform the crisp input into fuzzy input values. It is done by mapping the crisp values to a value in the fuzzy subset in a membership functions. These membership functions assign a degree of membership in the value range of [0,1] to the crisp value.

On the other hand, the rule base contains a set of fuzzy rules with the if-then conditions defined by the experts based on the basis requirement. The “if” also known as antecedent relates to the input side membership functions while the “then” also known as the consequent relates to the output side membership functions. Then, inference engines evaluate the fuzzy rules by determining the degree of truth or membership of the fuzzy inputs in each rule to generate fuzzy output. There are two common approaches in determining the degree of truth, take the minimum membership degree when the fuzzy operator is AND whereas take the maximum membership degree

when the fuzzy operator is OR. Lastly, defuzzification take place to convert the aggregated fuzzy output values from the inference engine to a single crisp output.

FL has been applied in many sectors such as in chemical science, agriculture, political science, environment science, household, operation research and healthcare industry (Makkar & Makkar 2018). An example of the application of FL in healthcare by Warren et al., who proposed a decision support system to automate clinical practice guidelines such as the imprecision in language description, lack of selectivity and sensitivity in medical examinations (Hayward & Davidson 2003). With fuzzy methods, the likelihood estimates from the test report can be handled as membership values and act as a weighted average in the decision making process (Hayward & Davidson 2003). Moreover, Mammadova et al has proposed a fuzzy rules-based system for HCC staging (Mammadova et al. 2021). Obot and Udoh (2011) applied fuzzy diagnosis on 10 hepatitis A patients by determining the exact degree of hepatitis on a patient. Uzoka et al. (2011); Zahan (2001) also apply fuzzy logic in diagnosis of malaria, and myocardial respectively.

2.8 CHAPTER SUMMARY

In this chapter, we have explored past research papers on HCC classification based on clinical data. Supervised ML techniques included rule-based and non-rule-based used in the previous research as presented in Chapter 2.3 and 2.4. Researchers are using various classifier models classifier models such as SVM, DT, RF, LR, GB, GNB, KNN, and NN to make the prediction and compared the results to get the best predictive models. Most of the studies have shown promising results in terms of accuracy.

Nonetheless, the application of these highly accurate predictive models in clinical practice remains quite restricted. The foremost factor is the fact that these HCC classification models lack interpretation in their results. Therefore, it has motivated us to use the FL approach to improve the interpretability of predictive models because FL can deal with ambiguity in the data. Even though the FL approach has been widely adopted in the field of engineering, it is still quite novel in medicine. By integrating the FL approach with the classification model, we could gain valuable insight into the

model results in a clear and understandable manner based on logical reasoning. FL aids physicians in making better decisions on identifying the risk of HCC.

Aside from that, most of the studies used a lot of variables available in the electronic medical record that consist of a combination of laboratory result and other HCC etiologies. However, developing predictive models with a wide range of variables will increase the model complexity and reduce efficiency. Hence, we aim to simplify the model by focusing on only using the most important and fundamental parameters (liver function results) to predict the risk of an HCC patient.

In addition, it is challenging in acquiring vast volumes of medical data due to the confidentiality of sharing patients' information. Consequently, most of the models are trained with a limited dataset and without external validation. This will bring up concerns regarding overfitting and generalization when dealing with new data, even if the predictive model's accuracy is high. To tackle these issues, our study will use a large amount of data acquired from the Kaggle. In the following chapter, we will discuss the whole development process from data analysis, and preprocessing to model training, testing, and implementation of FL in detail.

CHAPTER III

METHODOLOGY

3.1 INTRODUCTION

The real-world data often gathered from multiple and heterogeneous sources, contains large number of missing values, noises, and outliers. Poor data quality can lead to a decline in the performance of classification models. Accuracy, completeness, consistency, timeliness, credibility, and interpretability are all indicators of good data quality (Han et al. 2012). Data cleaning and pre-processing are fundamental to ensure data quality and prepare for modelling. While data visualization techniques such as plotting are effective techniques to access the distribution of each attribute, gaining insight into the data before cleaning and pre-processing. The flow chart of data preprocessing for this project is illustrated in Figure 3.1.

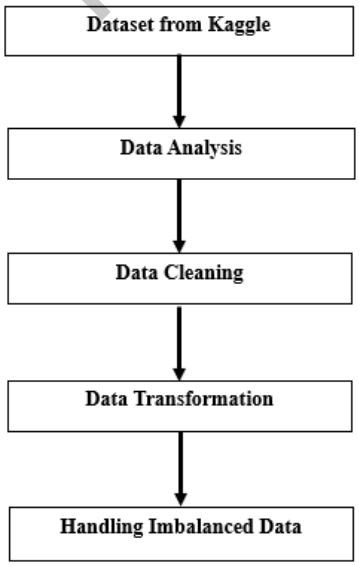

Figure 3.1 Flow chart of data preprocessing

3.2 DATASET

The dataset used in this project is taken from Kaggle (Shrivastava 2021). This dataset consists of 30691 LFT records of HCC and non-HCC patients. There are 583 of the records are collected from Northeast of Andhra Pradesh, India by (Ramana et al.2012).

The feature definition in Table 3.1 provides a more in-depth understanding of each feature's domain knowledge.

Table 3.1 Features definition in HCC dataset

Feature Name	Data Type	Definition
Age_of_the_patient	Discrete	The patient's age
Gender_of_the_patient	Nominal	The patient's gender
Total_Bilirubin	Continuous	Bilirubin is a brownish yellow pigment waste product produced during the normal breakdown of red blood cells. Total bilirubin is the sum of all forms of bilirubin in the blood, including both conjugated (direct) and unconjugated (indirect) bilirubin. It is a general marker of liver health. High level of bilirubin may be caused by the blockage of bile ducts or liver damage. Unconjugated bilirubin is water insoluble. In order for it to move through the circulation it must be bound to albumin.
Direct_Bilirubin	Continuous	It is the water-soluble form of bilirubin that has gone through a process known as conjugation in the liver. The conjugated bilirubin is excreted into bile and then released into the small intestines, finally eliminate as feces.
Alkphos_Alkaline_Phosphotase (ALP)	Discrete	An enzyme located in the bile duct within the biliary system, bones, and placenta of a pregnant women. It is important for breaking down proteins. Raised ALP may indicate cholestasis, which means reduction or blockage in the flow of bile within the biliary system. It may also increase bone breakdown or during pregnancy.
Sgpt_Alamine_Aminotransferase (SGPT)	Discrete	An enzyme mainly found in the liver that converts proteins into energy for the liver cells. SGPT in the blood will be raised when the liver is damaged in the form of hepatitis or inflammation.

to be continued…

… continuation

Sgot_Aspartate_Aminotransferase (SGOT)	Discrete	An enzyme that helps the body break down amino acids. It is presents in various part of human organ like liver, heart, kidneys, brain, muscles and red blood cells. SGOT is usually present in blood at low levels. An increase in SGOT levels may mean liver damage, liver disease or damage in other organs like heart or muscles.
Total_Proteins	Continuous	The sum of all proteins present, including albumin and globulins.
Albumin	Continuous	A specific type of protein that is synthesized by the liver. Albumin levels used as a marker of liver function and nutritional status. Low albumin levels may indicate liver disease (cirrhosis), malnutrition, or kidney disease. High albumin means dehydration.
Albumin_and_Globulin_Ratio (A/G Ratio)	Continuous	A general indicator for checking liver function (albumin) and the immune response of the body (globulin). Low A/G ratio usually associated with liver disease, kidney disease, chronic infection, or malnutrition. High A/G ratio is due to severe dehydration or diarrhoea.
Result	Discrete	1 represents the patient is diagnosed as HCC. 2 represents the patient is not diagnosed as HCC.

3.3 DATA ANALYSIS

3.3.1 Visualization of Numeric Features

a. Box Plot

Box plot visually presents the distribution and skewness of the numerical features by measuring the quartiles, interquartile range (IQR), percentiles and central tendency such as median (Han et al. 2012). IQR measured the distance between first quartile (Q1, 25% of the data) and third quartile (Q3, 75% of the data). In other words, IQR shows the middle 50% of the data. The line that split the box is the median of the data. Furthermore, the horizontal line extended from Q1 is known as lower whisker while extended from Q3 is known as upper whisker, The circle outside both the lower and upper whisker are the outlier.

The box plot of 9 numeric features as shown in Figure 3.2. Age_of_the_patient with the mean age of 44.11 and Total_Proteins were slightly negatively left-skewed distribution as the median closer to Q3. Both features have very few outliers.

On the other hand, Albumin showed slightly right-skewed distribution with longer upper whisker. Total_Bilirubin, Direct_Bilirubin, ALP, SGPT and SGOT were positively right-skewed distribution where the median is significantly closer to Q1. These five variables have a wider dispersion of the data with many outliers beyond the upper whisker. Extreme outliers in the right-skewed data distribution may indicate that a group of patients with abnormally high blood results have a higher risk of being diagnosed with HCC because majority of data is in the lower (normal) range. Furthermore, due to the standard deviation is highly sensitive to extreme outliers, liver enzymes such as ALP, SGPT, SGOT with extreme outlier of over 1000 have extremely high standard deviation.

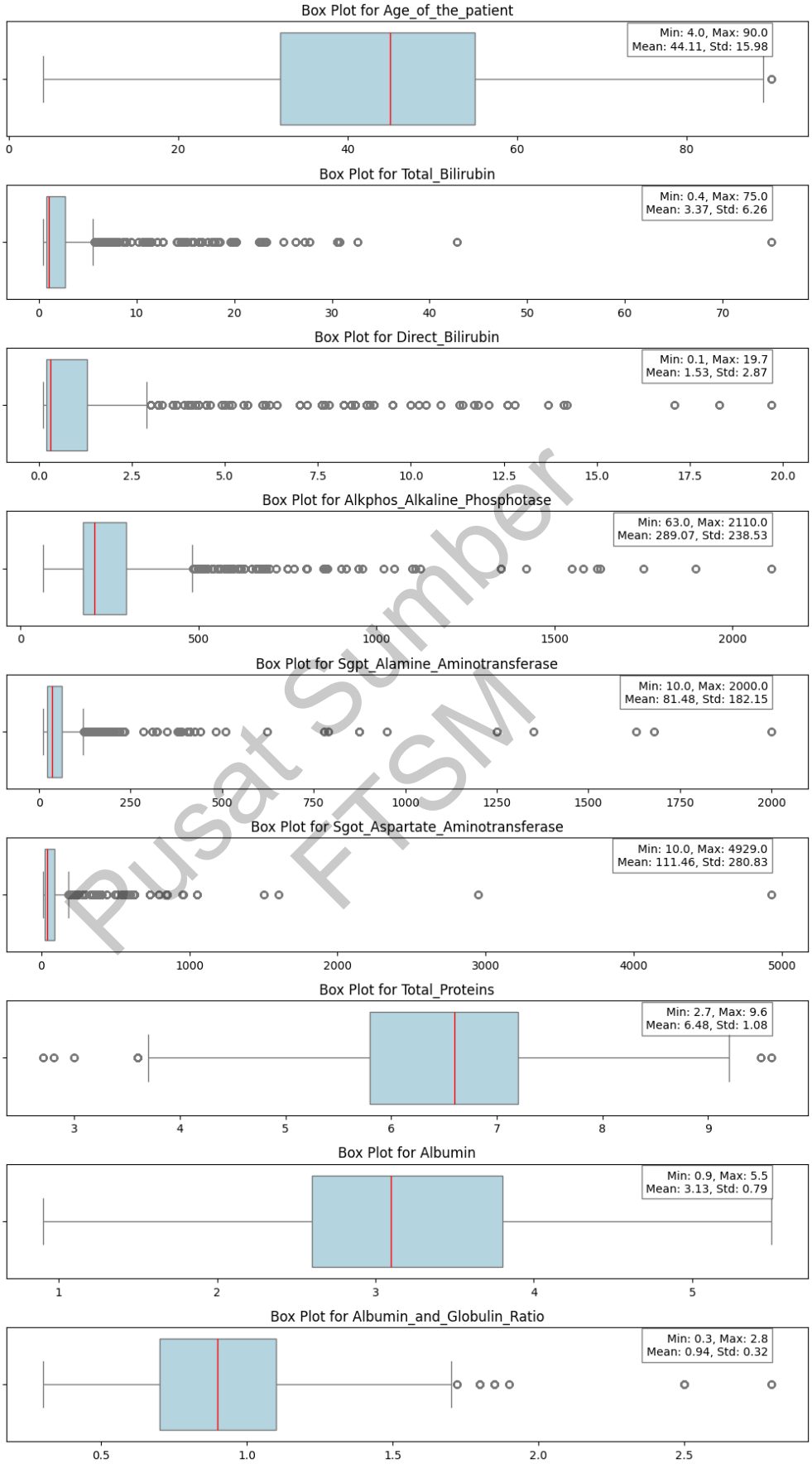

Figure 3.2 Boxplot of numeric features

3.3.2 Visualization of Categorical Feature

This HCC dataset as shown in Figure 3.3 consists of gender bias with more male patients than female patients.

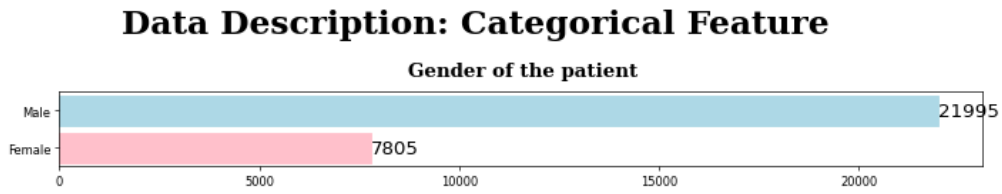

Figure 3.3 Data visualization for categorical feature

3.3.3 Visualization of Class Distribution

The class distribution in the dataset as shown in Figures 3.4 and 3.5 are highly imbalanced with 71.4% (21917) of class 1: HCC and 28.6% (8774) of class 2: non-HCC.

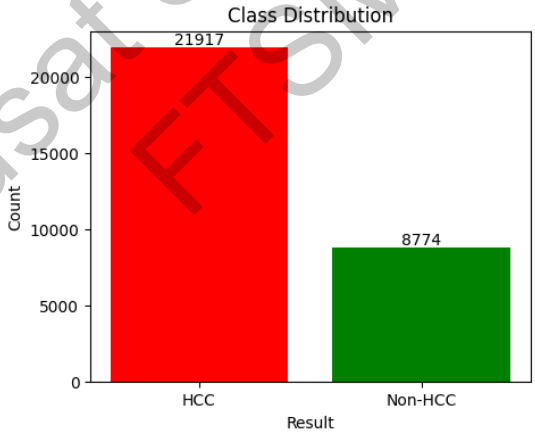

Figure 3.4 Class distribution of HCC dataset

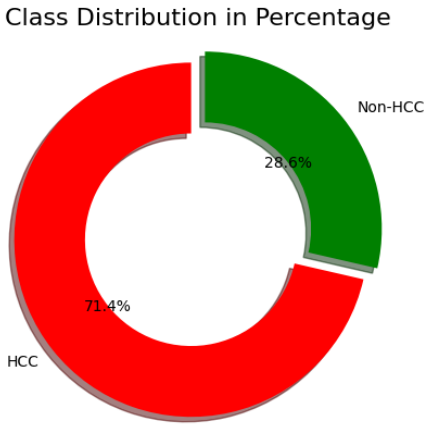

Figure 3.5 Class distribution in percentage

3.4 DATA CLEANING

3.4.1 Missing Values

Figure 3.6 showed that the total missing values are relatively low, with all variables having a missing value less than 1000. Age_of_the_patient, SGOT, Total_Proteins, Albumin have missing values below 500. SGPT, Direct_Bilirubin and A/G Ratio have missing values within the range of 500-600. Meanwhile, Total_Bilirubin, ALP and Gender_of_the_patient have missing values within the range from 600-900. Age_of_the_patient has the lowest missing value of 2. Gender_of_the_patient has the highest missing value of 891.

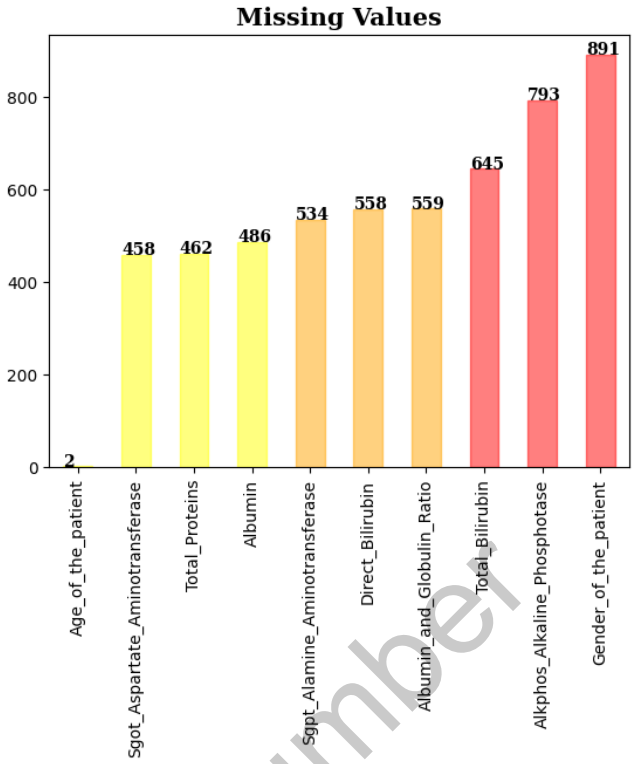

Figure 3.6 Summary table of missing values

3.4.2 Handling Missing Values

a. Row Deletion

The Gender_of_the_patient with highest number of missing values (891) will be directly removed from the dataset. The incomplete medical records on gender might be due to privacy concerns and refusal to disclose gender information for medical records. Therefore, the original dataset with 30691 rows of data is reduced to 29800.

b. Mean Imputation

Mean imputation is used to replace the missing values of Age_of_the_patient, Albumin, Total_Proteins, and A/G Ratio as these features only have limited outliers with a slightly skewed distribution.

c. Median Imputation

The missing values of Total_Bilirubin, Direct_Bilirubin, ALP, SGPT and SGOT are imputed using median as they have positively right-skewed distribution with significant

and many extreme outliers. Median imputation can reduce the data variability and maintain the data distribution shape because it is less sensitive to outliers compared to mean imputation.

d. Box Plot Comparison Before and After Imputation

The imputation of the missing values preserves the overall characteristics and patterns of the original data as shown in Figure 3.7. The main reason for retaining the original data without removing the outliers is that in the real world, those extremely high values in LFT indicate a patient with acute liver failure.

The box plot before and after imputation are quite similar. There are some slightly different in the median of Total_Proteins, Albumin and A/G Ratio after imputation. The median of Total_Proteins shifted slightly to the center of the box whereas the median of A/G Ratio are slightly closer to Q3 This suggests that data distribution of Total_Proteins become less negatively left-skewed and more symmetric whereas A/G ratio become more positively skewed. On the other hand, the median of A/G Ratio are slightly shifted closer to Q3. Furthermore, Albumin has a few outliers after mean imputation, but the outliers have a low impact because the median remains approximately the same as before.

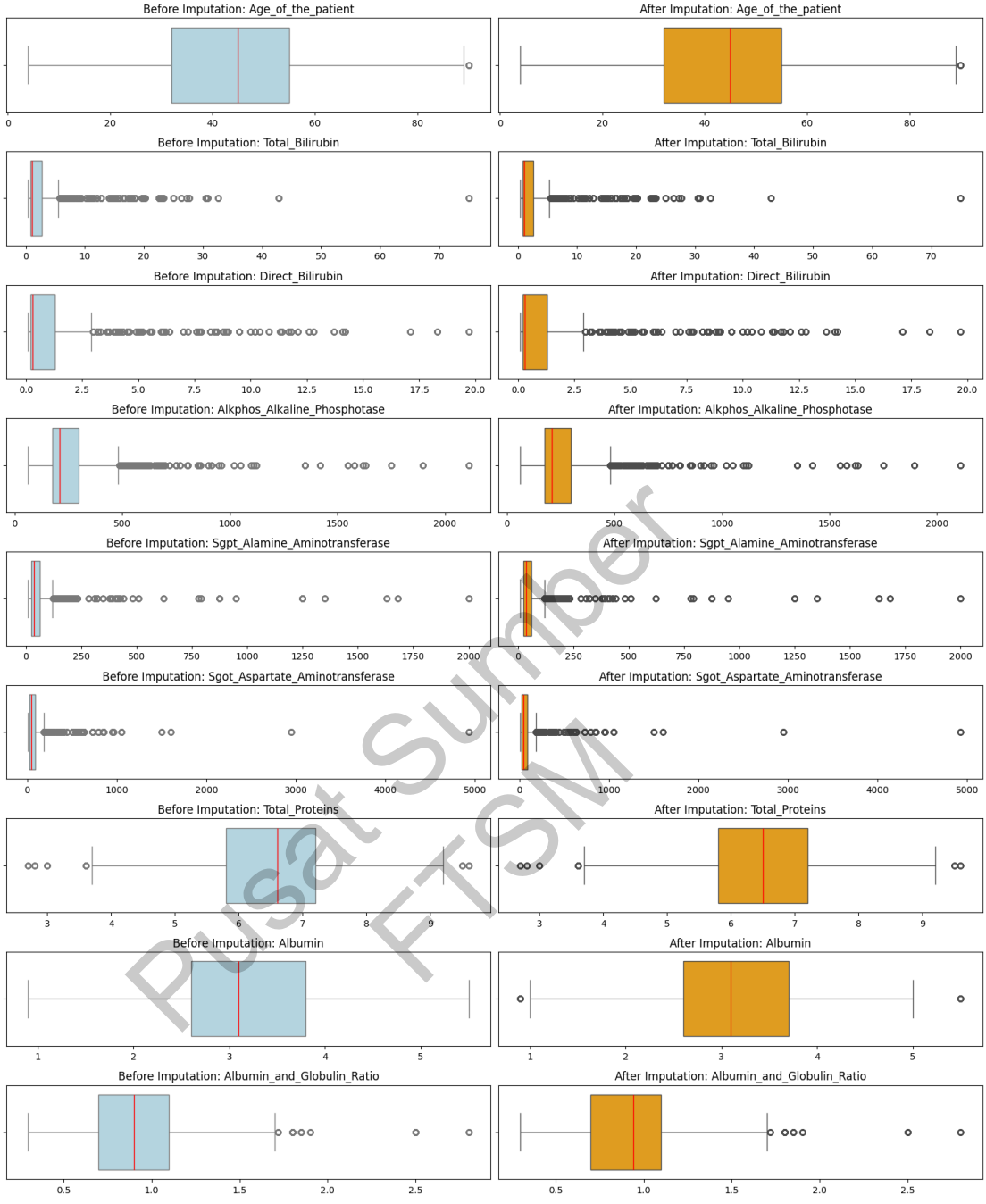

Figure 3.7 Boxplot comparison before and after imputation

e. Handling Duplicate Rows

There are 11241 rows of duplicate data found in the dataset and were deleted. Thus, the total number of data points has been reduced from 29800 to 18559.

f. Rename the Output Column

The column of the target output has been changed from Result to HCC for better understanding.

g. Convert the Output Label

The target class for “non-HCC” has been converted from 2 to 0. This is to ensure the ML algorithms can execute on the binary classification model.

3.5 UNIVARIATE ANALYSIS

According to the univariate analysis on Figure 3.8, most of the HCC patients are middle age of 30 - 59 and more prevalence in males. Patients with lower levels of Total_Bilirubin, Direct_Bilirubin, Albumin and A/G Ratio are at higher risk of being diagnosed with HCC. On the other hand, elevated liver enzymes such as ALP, SGPT, SGOT might indicate liver injury or damage, which can lead to liver cancer. Elevated bilirubin levels in an LFT can indicate potential liver problems. This is because when liver is damaged, it might have difficulty processing and excreting bilirubin properly, consequently, cause the bilirubin to accumulate in the blood.

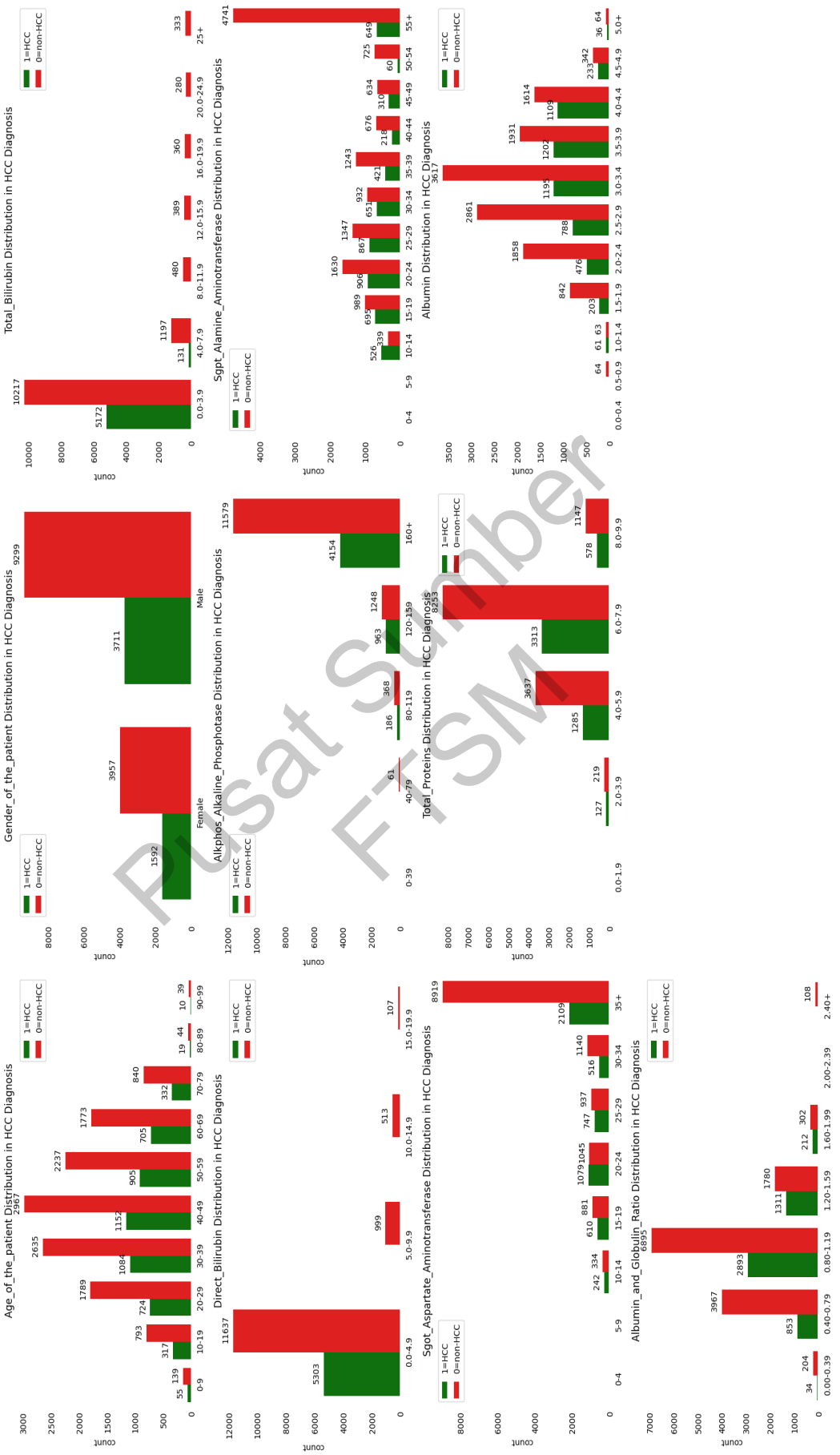

Figure 3.8 Univariate analysis of HCC

3.6 DATA PREPROCESSING

3.6.1 One Hot Encoding for Categorical Feature

One hot encoding is a technique used to transform nominal types of categorical data into numerical data to use in ML algorithms. In one hot encoding, an array that has as many elements as the number of categories is created. In order to represent the category, we have an array that consists of 0 everywhere except the element that corresponds to that particular category with 1. This is to avoid introducing ordinal relationship between the numerical data and ensure that the model does not interpret numerical values as having a meaningful order.

For instance, Gender_of_the_patient is converted into a numerical value represented by a binary (0 or 1) for each categorical variable such as Gender_Male and Gender_Female. Hence, when Gender_of_the_patient in the original dataset is male, it will become “0” in the column of Gender_Female and “1” in the column of Gender_Male.

3.7 BIVARIATE ANALYSIS

3.7.1 Pair Plot

Based on the pair plot on Figure 3.9, the Age_of_the_patient has no linear relationship with other variables as the points randomly scattered. It suggests that Age_of_the_patient is not a strong predictor of the expected values on the blood test results.

The 3 liver enzymes (SGPT, SGOT, ALP) are positively correlated with Total_Bilirubin and Direct_Bilirubin. This suggest that the elevated reading of Total_Bilirubin and Direct_Bilirubin is due to the value increase on the (SGPT, SGOT, ALP). Furthermore, (SGPT, SGOT, ALP) values are negatively correlated with Total_Proteins Albumin and A/G Ratio. The reading of Total_Proteins Albumin and A/G Ratio decline when the values of (SGPT, SGOT, ALP) rise.

Figure 3.9 Pair plot of HCC dataset

3.7.2 Correlation Matrix Heatmap

Based on the correlation matrix heatmap on Figure 3.10, there are strong positive correlation between some variables:

1. Total_Bilirubin & Direct_Bilirubin (0.8915)
2. SGPT & SGOT (0.7561)
3. Total_Proteins & Albumin (0.7675)
4. Albumin & A/G Ratio (0.6683)

On the other hand, Total_Bilirubin, Direct_Bilirubin, and the 3 liver enzymes (SGPT, SGOT, ALP) have positive correlation with the target variable. This indicates that when the value of these variables increase, the higher the chance it is classified as positive class (class 1: HCC). Whereas the lower the reading of Total_Proteins, Albumin and A/G Ratio, the more likely the chance diagnosed as (class 1: HCC).

Total_Bilirubin have a negative correlation with Total_Proteins, Albumin and A/G Ratio which means that the higher Total_Bilirubin level might cause the decrease of the of Albumin and A/G Ratio.

Most of the variables have negative correlation with the target variable (HCC) except Gender_Female, Total_Proteins, Albumin and A/G Ratio. Age_of_the_patient, Gender_Male, Gender_Female have very weak correlation (almost 0) with other features and the target class.

Figure 3.10 Correlation matrix heatmap of HCC dataset

3.8 DATA PREPROCESSING

3.8.1 Create New Feature

SGOT is not a specific marker for liver damage as they will cause potential damage to other organs like muscles, heart, and kidney. Therefore, SGOT usually measured together with SGPT to check for liver problems where SGOT is divided by the SGPT to get the ratio. The SGOT/SGPT ratio was referred to (Hexahealth, 2024) as shown in Figure 3.11.

For example:

1. An SGOT/SGPT ratio of greater than 1 where SGOT is higher than SGPT suggests potentially liver damage such as cirrhosis which will lead to HCC.

2. An SGOT/SGPT ratio of less than 1 where SGOT is lower than SGPT suggests non-alcoholic fatty liver disease.
3. An SGOT/SGPT ratio of equal to 1 where SGOT is equal to SGPT suggests primarily liver cell damage such as viral hepatitis, drug-induced injury such as liver toxicity.
4. An SGOT/SGPT ratio of 2:1 where SGOT is double the value of SGPT suggests alcoholic liver disease.

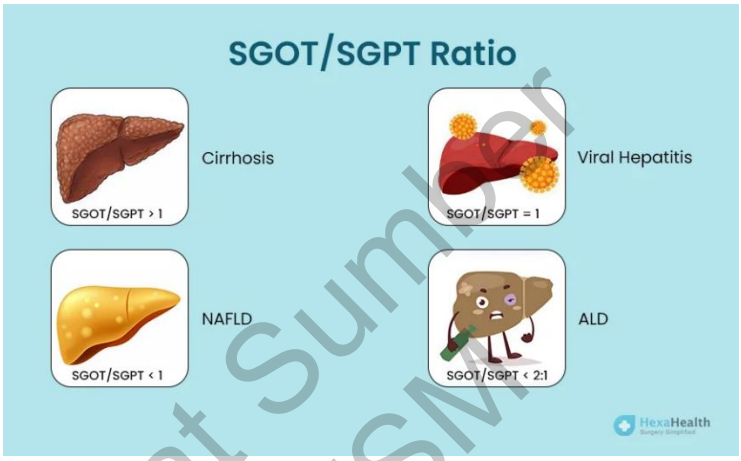

Figure 3.11 SGOT/SGPT ratio

Source: Hexahealth 2024

3.8.2 Remove Feature

1. Direct_Bilirubin will be removed since Total_Bilirubin includes the sum of all forms of bilirubin in the blood which are conjugated (direct) and unconjugated (indirect). Besides that, most of the laboratory tests only take Total_Bilirubin as a measurement in the liver function test.
2. Age_of_the_patient, Gender_Male, Gender_Female will be removed because they have extremely weak correlated with target class with almost 0 based on the correlation heatmap.
3. SGOT will be removed as it is not a specific indicator for early prediction in HCC. It is an enzyme present in various tissues including liver, muscle, brain, and heart. A high value in SGPT does not mean that the person has a high risk

of HCC, it could also mean a high risk of other diseases in the muscle, brain, heart, kidney, and pancreas.

3.8.3 Robust Scaler

Feature scaling is an important step in the data preprocessing to ensure that the features is transformed into a similar scale before fitting into ML models. This is to avoid biased model performance where the algorithms are more sensitive to the larger values of the features. Robust scaling is used in HCC dataset as it consists of features with extreme outliers such as the liver enzymes (ALP, SGPT) as shown in Figure 3.12. Robust scaler normalizes the scale of the features based on the median and IQR. The use of IQR is to reduce the effect of skewed data on the ML models as shown in Figure 3.13.

index	Total_Bilirubin	Alkphos_Alkaline_Phosphotase	Sgpt_Alamine_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio	SGOT_SGPT_Ratio	HCC
0	0.7	187.0	16.0	6.8	3.3	0.9	1.12	1
1	10.9	699.0	64.0	7.5	3.2	0.74	1.56	1
2	7.3	490.0	60.0	7.0	3.3	0.89	1.13	1
3	1.0	182.0	14.0	6.8	3.4	1.0	1.43	1
4	3.9	195.0	27.0	7.3	2.4	0.4	2.19	1
5	1.8	208.0	19.0	7.6	4.4	1.3	0.74	1
6	0.9	154.0	16.0	7.0	3.5	1.0	0.75	1
7	0.9	202.0	14.0	6.7	3.6	1.1	0.79	1
8	0.9	202.0	22.0	7.4	4.1	1.2	0.86	0
9	0.7	290.0	53.0	6.8	3.4	1.0	1.09	1

Figure 3.12 HCC dataset before scaling

	Total_Bilirubin	Alkphos_Alkaline_Phosphotase	Sgpt_Alamine_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio	SGOT_SGPT_Ratio	HCC
0	-0.166667	-0.189655	-0.513514	0.214286	0.166667	-0.107524	-0.076923	1
1	5.500000	4.224138	0.783784	0.714286	0.083333	-0.507524	0.406593	1
2	3.500000	2.422414	0.675676	0.357143	0.166667	-0.132524	-0.065934	1
3	0.000000	-0.232759	-0.567568	0.214286	0.250000	0.142476	0.263736	1
4	1.611111	-0.120690	-0.216216	0.571429	-0.583333	-1.357524	1.098901	1
5	0.444444	-0.008621	-0.432432	0.785714	1.083333	0.892476	-0.494505	1
6	-0.055556	-0.474138	-0.513514	0.357143	0.333333	0.142476	-0.483516	1
7	-0.055556	-0.060345	-0.567568	0.142857	0.416667	0.392476	-0.439560	1
8	-0.055556	-0.060345	-0.351351	0.642857	0.833333	0.642476	-0.362637	0
9	-0.166667	0.698276	0.486486	0.214286	0.250000	0.142476	-0.109890	1

Figure 3.13 Scaled HCC dataset

3.8.4 Adaptive Synthetic Sampling (ADASYN)

Imbalanced class distribution is one of the natures in most of the medical dataset that usually comes with one of the classes is in small minority. This issue is the major drawback of most ML algorithms in present a good performance of the classification model on the minority class. Therefore, ADASYN sampling method is applied to balance the class distribution before proceeding on model training and evaluation.

ADASYN focused on adaptively generating synthetic samples for minority class that are more difficult for the model to learn. ADASYN generate synthetic examples based on data distribution of the minority class, with more samples generated in areas of the feature space where the class density is low. This adaptive approach ensure that the synthetic instances are more representative of the minority class and less likely to overfit.

ADASYN is different from Synthetic Minority Oversampling Technique (SMOTE) which is also a synthetic data generation. SMOTE generate synthetic instance by randomly selects a minority class data (a) and its k nearest minority class neighbours (b). Then, both points a and b will form a line segment. The synthetics data generate in between the line segment. This might cause overfitting problem in the dense area of the feature space as the synthetic samples generated evenly between the original minority data.

3.9 CHAPTER SUMMARY

This chapter has demonstrated the techniques used in data visualization, data cleaning and data preprocessing. The missing values of the attributes were imputed with a mean and median, categorical variable is encoded, data is normalized using robust scaling and ADASYN sampling method is applied to balance the classification output. Finally, the original dirty data is cleaned, preprocessed, and well prepared for ML classification training and modelling.

CHAPTER IV

DATA MODELLING

4.1 INTRODUCTION

This chapter emphasizes on model development process starting from training and testing using the preprocessed data from the previous chapter, followed by a final evaluation among the models to choose the best performance model and then applying the FL to the best performance model with the objective to improve the interpretability of the classification model.

4.2 MODEL DEVELOPMENT APPROACH

The flow chart of model development is illustrated in Figure 4.1. The ML model development is an end-to-end iterative looping process that act as a guideline for developing a successful ML project. Data collection is done based on the understanding of the research objective. The second step is preparing the data for modelling. This includes data exploration analysis to get familiar with the data characteristics; data cleaning to handle inconsistent data, missing values, and outliers; data pre-processing such as encoding the categorical variables, feature scaling, and feature engineering. The details of data preprocessing can be referred in Chapter 3.

The preprocessed data is then divided into three groups: training (80%), and testing (20%). The train-test split to make sure that the selected models are learning the underlying patterns in the training data and generalize well on the new and unseen data. While the validation set (20%) split from the training (80%) is used for model optimization with hyperparameter tuning that helps to determine the best classifier for testing. After data splitting, the supervised ML models are trained with cross-validation.

Then, model evaluation is done based on accuracy, precision, sensitivity, and F1-score. In the real-world situation, the model performance on the training data is not always a good fit to the data. Hence, hyperparameter tuning is needed to optimize the overall performance of the model and its generalizability to unseen data.

Since we are focus on finding the model with interpretable results. Hence, only the rule-based model is chosen for optimization. The tuned rule-based model is then trained on validation data with cross-validation. Evaluation of tuned rule-based model to determine whether any improvement compared to the untuned model. The final evaluation of the tuned model on the test data is conducted only if the model's performance improves and overfitting issues are reduced. This is to ensure the model prediction achieve the research objective. Training, evaluating, and tuning are iterative processes in order to find the optimal model.

Lastly, we will integrate fuzzy logic into the final classification model for result interpretation. This section will be discussed in Chapter 5.

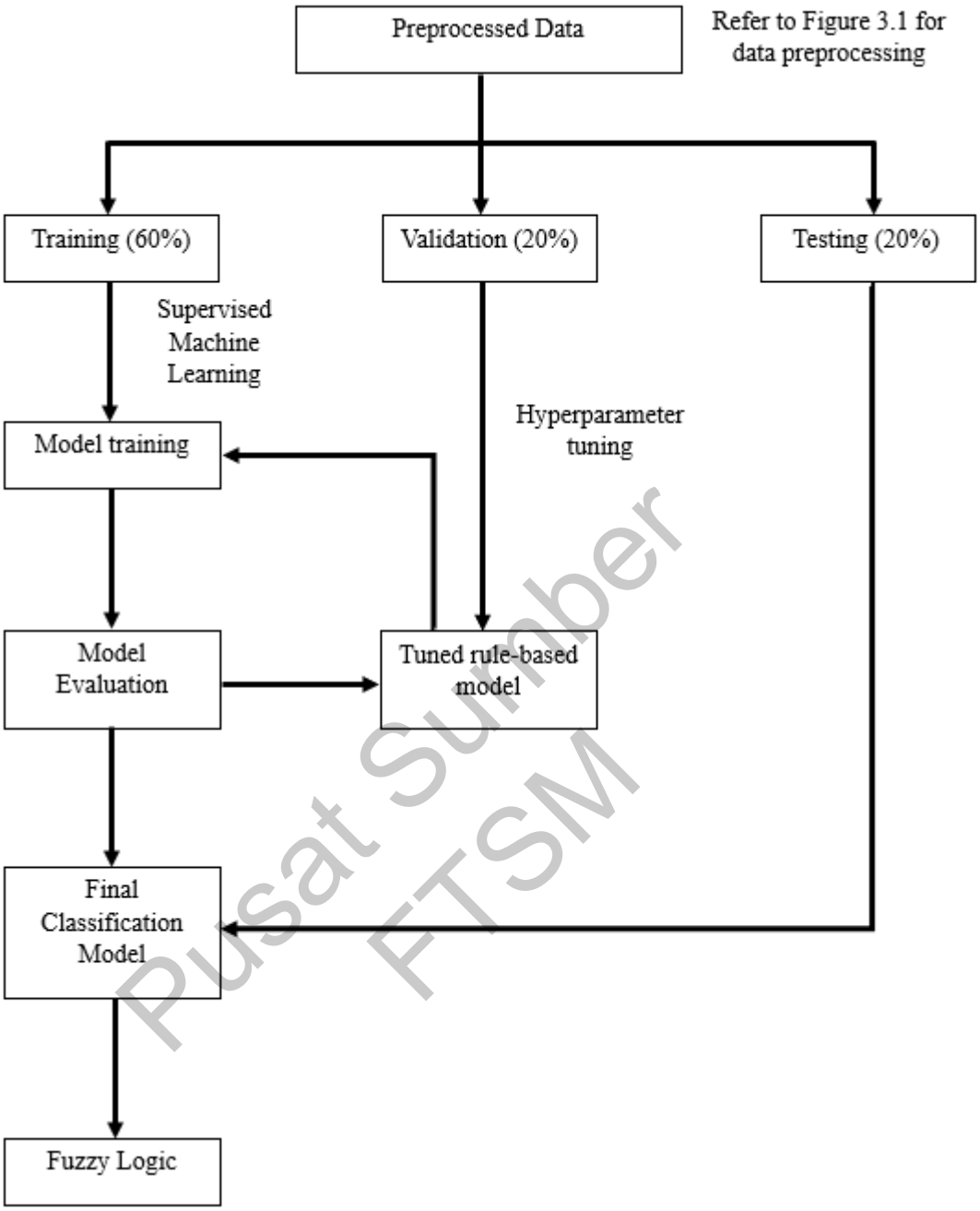

Figure 4.1 Model development process

4.3 MODEL TRAINING

The HCC risk prediction classifier models chosen for training are SVM, DT, RF, LR, GB, GNB, KNN and NN. These models have discussed earlier in section 2.3 and 2.4.

4.3.1 Train Test Data Split

The dataset is split into training (80%) and testing (20%). While 80% of training data is further split into training (80%) and validation (20%). The purpose of further splitting the training data into a validation set is to further improve the model if required. This is crucial to evaluate the generalization of the tuned model on unseen validation data before deciding on the final model evaluation on testing data.

4.3.2 Stratified K-Fold Cross-Validation

Cross-validation is a crucial technique used to assess the performance and generalizability of machine learning models on the new data. Stratified K-Fold cross-validation is useful for imbalanced classification tasks. It maintains the distribution of the target variable's classes in each fold as shown in Figure 4.2. In each iteration: k-1 folds are used for training the model. The remaining 1 fold acts as the validation set for that iteration. This process is repeated k times while ensuring each fold is used for validation once. 10 fold cross validation is used in this project for models training.

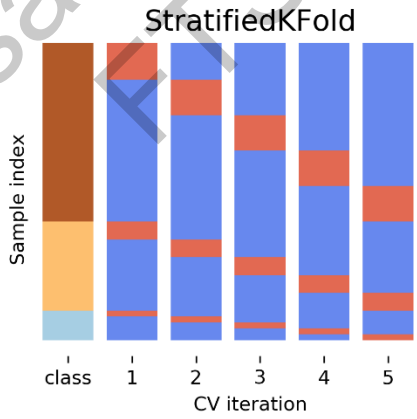

Figure 4.2 Stratified K-Fold validation

Source: Müller, 2020

4.3.3 Check for Overfitting

a. Cross Validation Score

Cross-validation scores function as a criterion to assess the risk of the overfitting of a model by estimating the prediction on unseen data based on the mean accuracy and

standard deviation of the model. It provides insights into the model performance in terms of consistency and variability. Mean accuracy shows how well the model performs on average across different folds of the training data, whereas standard deviation measures the variability or spread of the accuracy scores across the different folds.

1. High mean accuracy and low standard deviation means the model is consistent in their performance and has a good generalizability on unseen data.
2. Overfitting occurs when the mean accuracy and standard deviation of the model exhibit high values. This means that the model's performance displays high variability among distinct folds and have good performance in the training dataset, potentially showing sensitivity to specific subsets while lacking generalizability to unseen data.
3. Underfitting, conversely, happens when both the mean accuracy and standard deviation of the model are low. This indicates that the model has a poor performance with inconsistent predictions. The model is not learning the underlying patterns in the dataset.

```
Cross-validation scores in SVM model training
[0.71175428 0.7241583  0.71825162 0.74069699 0.70761961 0.70998228
 0.73715298 0.71707029 0.72399527 0.72044917]
0.7211 accuracy with a standard deviation of 0.0104
--------------------------------------------------------------------------------
Cross-validation scores in Decision Tree model training
[0.99173066 0.98405198 0.98700532 0.98936799 0.99054932 0.98759598
 0.98759598 0.98287064 0.9893617  0.98995272]
0.9880 accuracy with a standard deviation of 0.0027
--------------------------------------------------------------------------------
Cross-validation scores in Random Forest model training
[0.99763733 0.99586533 0.996456   0.99468399 0.99763733 0.99763733
 0.99704666 0.99291199 0.99763593 0.99468085]
0.9962 accuracy with a standard deviation of 0.0016
--------------------------------------------------------------------------------
Cross-validation scores in Logistic Regression model training
[0.70525694 0.71234495 0.6969876  0.73479031 0.6987596  0.70466627
 0.71411695 0.70821028 0.71217494 0.71926714]
0.7107 accuracy with a standard deviation of 0.0104
--------------------------------------------------------------------------------
Cross-validation scores in Gradient Boosting model training
[0.87241583 0.86887183 0.87123449 0.87714117 0.85646781 0.8747785
 0.87418783 0.88481985 0.88947991 0.87765957]
0.8747 accuracy with a standard deviation of 0.0085
--------------------------------------------------------------------------------
Cross-validation scores in Gaussian Naive Bayes model training
[0.65623154 0.67099823 0.66095688 0.68103957 0.65386887 0.65564087
 0.66450089 0.65564087 0.66489362 0.66016548]
0.6624 accuracy with a standard deviation of 0.0080
--------------------------------------------------------------------------------
Cross-validation scores in K-Nearest Neighbors model training
[0.98582398 0.98759598 0.98523331 0.99054932 0.98700532 0.98464265
 0.98523331 0.98523331 0.98995272 0.98699764]
0.9868 accuracy with a standard deviation of 0.0019
--------------------------------------------------------------------------------
Cross-validation scores in Neural Network model training
[0.86769049 0.89072652 0.89545186 0.88777318 0.89781453 0.88836385
 0.88304784 0.89427053 0.89066194 0.9036643 ]
0.8899 accuracy with a standard deviation of 0.0092
--------------------------------------------------------------------------------
```

Figure 4.3 Cross validation scores of the classification models

Table 4.1 Accuracy and standard deviation of classifiers in model training

Classifier	Accuracy	Standard Deviation
DT	0.9880	0.0027
RF	0.9962	0.0016
KNN	0.9868	0.0019

Figure 4.3 showed the cross validation scores for SVM, DT, RF, LR, GB, GNB, KNN and NN in each fold during the model training. Models such as DT, RF, KNN are among the best models. The highlighted result can be seen in Table 4.1 where the mean