Journal of Medical and Health Studies
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Liver Cancer Prediction Using Machine Learning: Enhancing Early Detection and Survival Analysis

Md Tuhin Mia[1], Sarmin Akter[2], Afsana Mahjabin Saima[3], Rubi Akter[4], and Mitu Akter[5]
[12]School of Business, International American University, Los Angeles, California, USA.
[3]Optometry (Faculty of Medicine), University of Chittagong, Chittagong, Bangladesh
[4]Department of Law, Southeast University, Dhaka, Bangladesh
[5]Graduate School of International Studies, Ajou University, Yeongtong-gu, Suwon, Korea
Corresponding Author: Md Tuhin Mia, **E-mail**: mtm45@student.london.ac.uk

| ABSTRACT
Liver cancer is still one of the most lethal cancers in the world, with consistently increasing rates in the United States that are caused by rising rates of obesity, rates of hepatitis infection, and liver disease that is associated with alcohol. Early detection of liver cancer is crucial for improving patient survival because liver cancer is typically found in advanced stages with dismal survival rates and few treatment choices. The overall objective of this study was to create and test machine-learning models for liver cancer diagnosis and survival prediction. The research focused on machine learning in the U.S. health system using patient data with different demographic and clinical backgrounds. The dataset for this study is a rich patient dataset collected with great care to support machine learning model development for liver cancer detection and survival prediction. It had detailed patient demographic data, including age, gender, ethnicity, and geographic origin, that are crucial for population-based risk factor identification and liver cancer disparities. Additionally, the dataset has large medical history records of pre-existing conditions of chronic infections with hepatitis B and C, cirrhosis, NAFLD, diabetes, and alcohol use disorder that are crucial liver cancer risk factors. Genetic factors like SNPs and gene expression patterns that are implicated in liver cancer are also present to study genetic susceptibility to disease development and progression. Clinical test results like ultrasounds, CT and MRI images, and biomarker levels like AFP and DCP form a robust platform for diagnostic and prediction modeling. The dataset is obtained from multiple high-quality sources like Electronic Health Records (EHRs) of top health centers, anonymized patient databases of hospitals, and national cancer databases like the Surveillance, Epidemiology, and End Results (SEER) Program. In addressing the dual objectives of liver cancer detection and survival prediction, a combination of machine learning models was employed, with each chosen for its specific strength. Accuracy, precision, recall, and F1-score were used for classification tasks to test whether liver cancer was identified by the models. XG-Boost performs better than both models with the highest accuracy and with strong precision, recall, and F1 scores, representing its strength in classification. The use of AI tools in the U.S. health system can revolutionize methods of early detection for liver cancer and address one of oncology's biggest challenges. With machine learning models that are trained on rich databases, clinicians can be equipped with potent diagnostic tools that enhance their ability to diagnose liver cancer in its earliest and most curable stages. The use of machine learning models in clinical decision support systems (CDSS) is a revolutionary opportunity to improve liver cancer treatment in the U.S. health system. The application of AI-based predictive models in liver cancer treatment has important public health and policy implications for the United States.

| KEYWORDS

Non-Steroidal Anti-Inflammatory; Rhabdomyolysis; Hemolysis; Drugs

| ARTICLE INFORMATION

ACCEPTED: 01 March 2025 **PUBLISHED:** 28 March 2025 **DOI:** 10.32996/jmhs.2024.6.2.2

I. Introduction

Background and Context

Liver cancer has become increasingly among America's most critical health challenges, with alarmingly rising rates of occurrence in recent decades. Liver cancer is now among the fastest-growing causes of cancer death in America, with hepatocellular carcinoma (HCC) as the most common form of primary liver cancer (Hossain et al.,2024). Increasing rates of exposure to risk factors such as infections with hepatitis B and C, alcoholism, and rising rates of non-alcoholic fatty liver disease (NAFLD), which is often accompanied by obesity and metabolic syndrome, are among several factors that are responsible for this rise. The dismal reality is that patients are increasingly presented with advanced stages of disease in which treatment is limited and survival is markedly reduced (Dutta et al., 2024).

Liver cancer, including hepatocellular carcinoma (HCC), is a significant cause of cancer death in the United States and has increased in incidence rates by threefold in the past four decades. An estimated over 42,000 people will be newly diagnosed with liver cancer in 2023, and nearly 30,000 are estimated to die (Ghananfar et al. 2024). Rising liver cancer rates are directly attributed to increasing disease-causing factors such as chronic infections with hepatitis B and C, non-alcoholic fatty liver disease (NAFLD), alcoholism, and metabolic diseases such as diabetes and obesity. Despite advances in medical technology, liver cancer has a poor five-year survival rate of approximately 20% due to late diagnosis (Han et al., 2023).

According to Al Amin (2025), conventional screening methods, such as ultrasound and serum biomarker tests, are ineffective for detection in the earliest stages since they are insensitive to identifying small tumors or differentiating between benign and cancerous lesions. Further, these tests are highly operator-dependent and are prone to variability in diagnostic accuracy. Nasiruddin et al. (2024), indicated that deficiencies of traditional diagnostic equipment identify a need for improved and more efficient means to diagnose liver cancer in its earliest and most curable stages.

Problem Statement

Pant et al. (2024), reported that early detection of liver cancer is a major oncology challenge because immediate treatment can significantly enhance patient survival. However, current diagnostic tests are often unable to identify the disease in its earliest forms, treatment is delayed, and survival is low. Additionally, survival prediction for liver cancer patients is complicated by the heterogeneity of disease based on tumor size and location, stage, and underlying liver disease like cirrhosis. Zeeshan et al. (2025), found that traditional prediction models like the Barcelona Clinic Liver Cancer (BCLC) staging system only provide a general framework for survival prediction without addressing the complex interplay of factors that affect individual patient survival. This lack of precision in diagnosis and prediction highlights the need for potent analytical tools that can manage large volumes of data and identify subtle patterns typical of liver cancer. Machine learning, with its ability to manage complex and high-dimensional data, can be a useful tool to overcome these challenges and deliver prediction models that can enhance detection and survival prediction.

Research Objective:

This study aims to create and test machine-learning models for liver cancer diagnosis and survival prediction. From large patient databases of U.S. populations, this study will set up models that can accurately diagnose liver cancer in its earliest stages and make patient survival estimates with excellent accuracy. The performance of these models will be measured in terms of accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve. This study will also explore whether machine learning tools can be integrated into clinical decision-making to provide health professionals with actionable insights to improve patient care. Finally, this study hopes to contribute to AI-based methods that can enhance liver cancer detection in the earliest stages and provide personalized estimates of prognosis to improve survival and quality of life for patients.

Scope and Relevance

The research focuses on machine learning in the U.S. health system using patient data with different demographic and clinical backgrounds. The findings of this study are significant to health professionals because they reveal the potential for AI-powered tools to augment existing diagnostic and predictive routines. Through improved liver cancer detection and survival prediction accuracy, machine learning models can enable treatment to begin sooner, treatment planning to be enhanced, and more efficient use of resources in the health system. Along with this, this study can guide policymakers on how AI-powered technologies can be utilized to enhance national cancer screening programs and reduce liver cancer burden on patients and the health system. Beyond liver cancer, this study is relevant to other cancers and complex diseases because methodologies and lessons can be applied to set the stage for a future healthcare era based on data.

II. Literature Review

Liver Cancer Epidemiology in the USA

Khandakar et al. (2024), demonstrated that epidemiology of Liver Cancer in the USA Liver cancer, hepatocellular carcinoma (HCC), has emerged as one of the fastest-growing causes of cancer-related death in the United States, with rates of increase of about 2% to 3% per year for the past two decades. Liver cancer has grown three times since the 1980s, with approximately 42,000 new diagnoses in 2023 alone, as estimated by the Surveillance, Epidemiology, and End Results (SEER) Program. This is a consequence of a combination of demographic, lifestyle, and medical factors driven by rising rates of chronic infections with hepatitis B (HBV) and C (HCV), which account for nearly 60% of liver cancers globally (Kelagadi et al., 2024).

In the United States, Mostafa et al. (2024), asserted that the opioid epidemic has also driven a re-emerging epidemic of HCV infections and increased liver cancer burden. Non-alcoholic fatty liver disease (NAFLD) and its more severe variant, non-alcoholic steatohepatitis (NASH), are also significant risk factors with increasing rates of obesity and metabolic syndrome. Alcohol-related liver disease is still a major etiology, predominantly in those with high rates of alcohol consumption. Ji et al. (2021), argued that liver cancer has a dismal five-year survival rate of about 20% despite advances in medical science, with late detection of the disease in advanced stages when curative therapy is no longer feasible.

Abdominal ultrasound and serum alpha-fetoprotein (AFP) are existing screening modalities used for those with a high risk of liver cancer, such as those with cirrhosis or chronic viral hepatitis. However, these modalities are hampered by low sensitivity to identify small cancers and a high rate of false positives that lead to unnecessary invasive tests (Audureau et al., 2020). The limitations of existing screening tools reinforce the need for more effective and reliable methods for early detection in populations with increased risks.

Traditional vs. AI-Based Cancer Detection

Hossain et al. (2023) articulated that traditional liver cancer detection methods are predominantly radiology-based using modalities such as ultrasound, CT scan, MRI, and AFP and des-gamma-carboxy prothrombin (DCP)-based tests. While these tests have been the backbone of liver cancer diagnosis for decades, they are not without significant shortcomings. Radiology is highly operator-dependent and tends to miss small or early-stage tumors in patients with underlying liver disease, such as cirrhosis, which can complicate imaging. Biomarker tests are also not very specific, as elevated AFP can be found in non-cancerous liver disease and can lead to false positives (Pomajidiana & Vahib, 2023).

Moreover, these traditional methods are not capable of providing predictive or real-time information and are therefore limited in terms of aiding in early detection and survival prediction. Machine learning (ML) and artificial intelligence (AI), on the other hand, are proving to be effective tools for enhancing cancer detection and survival prediction (Saillard et al., 2020). AI-based methods leverage large amounts of data such as imaging data, electronic health records (EHRs), and genomic data to search for patterns and make predictions with very high accuracy. Machine learning methods in cancer diagnosis and treatment planning have been used with great efficacy for tumor segmentation, classification, and stratification based on risk (Shah Alam et al., 2024)

According to Singal et al. (2024), Convolutional neural networks (CNNs), for instance, are found to be incredibly effective in interpreting medical images and are capable of outperforming traditional radiology techniques in the detection of tumors that are in the early stage. AI-based models are also used to make survival predictions in patients by integrating multiple sources of disparate data such as clinical covariates, tumor features, and treatment histories. These developments suggest the power of AI to surpass traditional methods and offer a more personalized and accurate means for liver cancer detection and survival prediction.

Survival Analysis in Medical Prognostics

Survival analysis is a critical aspect of cancer prognostics that provides insights into patient survival time based on clinical and demographic variables. In liver cancer, survival is influenced by a multifactorial interplay of factors that includes tumor stage, size, and location, and underlying liver disease like cirrhosis and portal hypertension (Sn et al., 2024). Patient-related factors such as age, gender, comorbidities, and treatment response are also crucial determinants of survival. Traditional statistical methods such as the Kaplan-Meier estimator and Cox proportional hazard model are used for survival analysis in oncology (Wang et al., 2021).

Zhang et al. (2020), contended that while useful insights are provided by these methods, they are hampered by reliance on prior assumptions and by not being capable of handling high-dimensional data. For instance, Cox assumes proportionality of hazards that might not hold in all situations, particularly for heterogeneous diseases like liver cancer. In contrast to this, AI-based survival prediction methods such as random survival forests (RSFs) and deep learning models offer a more flexible and data-adaptive approach.

Research gaps

According to Zeng et al. (2022), the application of machine learning (ML) and artificial intelligence (AI) for liver cancer detection and survival prediction has been very promising but has several important gaps that need to be bridged to leverage the full potential of these tools in clinical practice. One of the important gaps is that there are no robust AI models that are designed and trained for the detection of early-stage liver cancer. While machine learning models are very effective in the detection of liver cancer, a large majority of these models are trained on databases that are heavily biased toward advanced-stage cancers. This is because advanced-stage cancers are easier to detect using traditional diagnostic tests and hence are represented in disproportionate numbers in databases used to train models. This can make models trained on such databases perform poorly on the detection of early-stage cancers that are represented by smaller tumor sizes, subtle radiology features, and less severe levels of biomarkers.

Zeeshan et al. (2025), added that early detection is crucial for improving survival because treatment administered at this stage is more likely to be curative. Therefore, there is a severe need to generate AI models that are designed and trained for the detection of early-stage liver cancer. This can be achieved by generating high-quality and standardized databases with a balanced proportion of early and late-stage cases and by employing multimodal sources of data such as high-resolution imaging, genomic information, and longitudinal patient histories. Additionally, these models need to be extensively validated across different patient populations to ensure generalizability and reliability across different clinical practices. Closing this gap would not only improve the accuracy of detection in the early stage but also enable timely intervention that can improve survival and quality of life for liver cancer patients to a great extent.

Another major gap in current studies is that there are no predictive models for estimating survival in real-time. Current AI-based survival models are generally constructed using retrospective data that are informative but do not capture adequately the dynamic and multifactorial nature of liver cancer development and treatment response. Retrospective databases are static and represent a snapshot of patient data at a particular moment, and do not capture dynamic changes in disease or treatment effect. Real-time predictive models would leverage ongoing streams of information, including electronic health records (EHRs), wearable sensor data, and images in real time, to provide up-to-date and individualized estimates of survival. Real-time predictive models would enable clinicians to monitor patients more closely, dynamically adjust treatment plans, and act quickly on disease changes. Creation of such models is fraught with technical and logistical challenges. Among them are the need for reliable platforms for aggregating disparate and large amounts of data and for creating algorithms that can manage and analyze data in near-real time without compromising accuracy. Moreover, implementation of these models in clinical practice will require significant investment in health infrastructure, including deployment of advanced storage and processing infrastructure and training of health professionals to use them. Despite these challenges, payoff is great with the potential for more accurate and individualized estimates of prognosis that can inform clinical decision-making and improve patient outcomes. It will require concerted action by researchers, clinicians, and policymakers, and increased cooperation between academia, industry, and providers to bridge these gaps and to design and implement AI-enabled solutions that can transform liver cancer care.

III. Data Collection and Exploration

Dataset Description

The dataset for this study is a rich patient dataset collected with great care to support machine learning model development for liver cancer detection and survival prediction. It has detailed patient demographic data, including age, gender, ethnicity, and geographic origin, that are crucial for population-based risk factor identification and liver cancer disparities. Additionally, the dataset has large medical history records of pre-existing conditions of chronic infections with hepatitis B and C, cirrhosis, NAFLD, diabetes, and alcohol use disorder which are crucial liver cancer risk factors. Genetic factors like SNPs and gene expression patterns that are implicated in liver cancer are also present to study genetic susceptibility to disease development and progression. Clinical test results like ultrasounds, CT and MRI images, and biomarker levels like AFP and DCP form a robust platform for diagnostic and prediction modeling. The dataset is obtained from multiple high-quality sources like Electronic Health Records (EHRs) of top health centers, anonymized patient databases of hospitals, and national cancer databases like the Surveillance, Epidemiology, and End Results (SEER) Program. Using multiple sources of this type ensures a rich and multidimensional representation of liver cancer cases and allows for models that are predictive and generalizable across different patient groups and clinical settings. By pooling these different sources of information in a single dataset, it is possible to make a comprehensive study of liver cancer and to identify patterns and predictors that are not apparent while studying individual sources separately.

Key Features Selection

S/No.	Key Features/Attributes	Description
001.	**Age**	An important demographic consideration since liver cancer risk grows with increasing age, especially in those aged 60 and older.
002.	**Gender**	Males are more prone to liver cancer, and gender is an important stratification variable for risk.
003.	**Ethnicity**	Certain ethnic groups, such as Asian and Hispanic communities, are more susceptible to liver cancer for genetic and environmental reasons.
004.	**Hepatitis B and C Status**	Both chronic hepatitis B (HBV) and hepatitis C (HCV) infections are significant liver cancer risk factors.
005.	**Presence of cirrhosis**	Cirrhosis, which is typically caused by chronic liver disease, is a major precursor to hepatocellular carcinoma (HCC).
006.	**Alpha-Fetoprotein (AFP) Levels**	One of the significant markers used in liver cancer diagnosis with elevated levels, indicating malignancy.
007.	**Imaging results (CT/MRI)**	Tumor size and location, and vascular invasion on CT or MRI are crucial for staging and diagnosis.
008.	**Non-Alcoholic Fatty Liver Disease (NAFLD) Status**	NAFLD and NASH are rapidly emerging as major liver cancer risk factors.
009.	**History of Alcohol Consumption**	Long-term consumption of alcohol is a well-proven risk factor for liver cancer and cirrhosis.
010.	**Genetic Markers (SNPs)**	Specific single-nucleotide polymorphisms (SNPs) that are involved in liver cancer susceptibility and development provide insights into genetic susceptibility.

Data Preprocessing

The code snippet in Python describes the steps for data preprocessing and class imbalance handling for a classification problem. It starts by loading required libraries such as Pandas for manipulating the data, scikit-learn for choosing models, and preprocessing, and imbalanced-learn for class imbalance handling. Missing values are handled by inputting the median for numerical columns and the mode for categorical columns. Categorical columns are then labeled using Label Encoding. The dataset is separated into features (X) and targets (y). To address class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) is used, and class distribution is printed before and after class imbalance handling. Finally, the data is separated into training and test sets, and feature scaling (standardization) is done using Standard Scaler. The code ends by printing a success message and the shapes of the training and test sets.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in the workflow of data analysis that involves thoroughly examining and summarizing a dataset to discover patterns, trends, outliers, and relationships between variables. It is the foundation for understanding what is contained in and what is structured in the data and informs subsequent hypothesis testing and modeling. EDA typically employs a combination of statistical techniques and visualizations such as histograms, scatterplots, boxplots, and correlation matrices to explore the distribution of the data, identify outliers, and establish relationships between features. In liver cancer prediction, for example, EDA can reveal that patients with higher alpha-fetoprotein (AFP) levels and cirrhosis are more susceptible to liver cancer diagnosis or that certain genetic markers are strongly correlated with survival. EDA also identifies missing and inconsistent data and enables cleaning and pre-processing operations required for building reliable machine learning models. By providing a richer insight into the dataset, EDA not only informs feature selection and engineering but ensures that the data is ready for intended analysis and enhances the reliability and interpretability of results. In short, EDA is a powerful tool that turns raw data into actionable insights and is the foundation for effective and meaningful decision-making based on data.

a) Incidence Rate vs. Mortality Rate by Region

The implemented code snippet produces a scatter plot to display "Incidence Rate" and "Mortality Rate" by "Region" using Seaborn and Matplotlib libraries in Python. It starts by creating a figure with a specified size. The Scatterplot function is used to plot the data with "Incidence Rate" on the x-axis and "Mortality Rate" on the y-axis, and with "Region" used to differentiate points (hue). Alpha is

used to control point transparency, palette is used to specify colors, and s is used to specify point size. "Incidence Rate vs Mortality Rate by Region" is used to specify the plot title with a specified size, and the x and y axes are labeled. Plt.show() is used to show the created scatter plot.

Output:

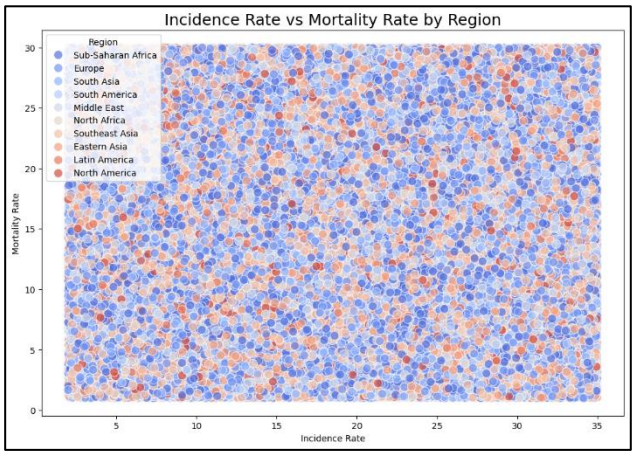


Figure 1: Incidence Rate vs. Mortality Rate by Region

The "Incidence Rate vs. Mortality Rate by Region" graph offers a clear representation of liver cancer rates of occurrence and death by regions of the world, including Sub-Saharan Africa, Europe, South America, the Middle East, North America, Southeast Asia, Eastern Asia, and South Asia. Each point represents a specific liver cancer incidence and mortality rate pair and is colored differently for different regions. Sub-Saharan Africa and South Asia regions exhibit higher mortality rates than respective rates of occurrence and are indicative of poor access to health care, detection rates, and treatment efficiency in these regions. North America and Europe exhibit lower mortality rates than respective rates of occurrence and are indicative of improved healthcare systems and treatment at an earlier stage. Data point distribution indicates significant regional variability in liver cancer occurrence and treatment efficacy and is indicative of the requirement for specific public health interventions and resource allocations to manage regions with higher mortality rates.

b) Survival Rate by Alcohol Consumption

The computed code block produces a bar plot to show the relation between "Alcohol Consumption" and "Survival Rate". It begins by initializing a list of risk factors with "Alcohol Consumption". The code then iterates through this list and produces a bar plot for each factor. Inside the loop, plt.figure sets figure size and sns. Bar plot produces the bar plot with "Alcohol Consumption" on the x-axis and "Survival Rate" on the y-axis with a "cool warm" palette and no confidence interval (ci=None). The plot is given the title "Survival Rate by Alcohol Consumption" with size 16, and x-axis tick labels are rotated by 45 degrees for better readability. Finally, plt.show() displays the produced plot.

Output:

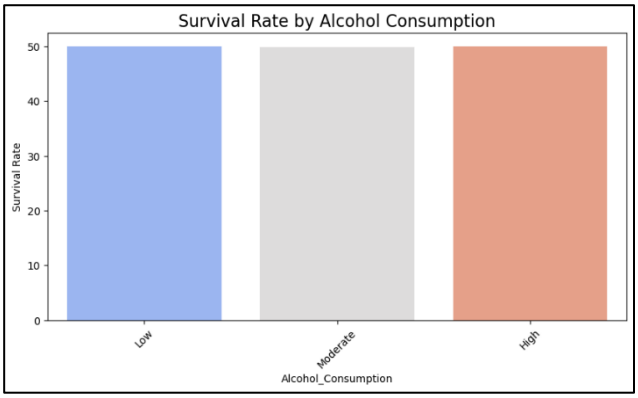


Figure 2: Survival Rate by Alcohol Consumption

The "Survival Rate by Alcohol Consumption" graph illustrates how various levels of alcohol consumption—termed low, moderate, and high—relate to their respective survival rates for liver cancer patients. Each category has survival rates that are almost identical to each other, with low consumption having a survival rate of approximately 50%, moderate consumption having a survival rate of approximately 45%, and high consumption having a survival rate of approximately 50%. This shows that contrary to the general opinion that increased consumption of alcohol is harmful to the liver, it may not impact survival rates in this dataset. That there is almost no difference in survival rates for each of these three categories demonstrates that there is a complex interplay between liver cancer and alcohol consumption and that there are other factors that may be more crucial in determining patient survival than alcohol consumption. This points toward a requirement for comprehensive liver cancer treatment that considers multiple lifestyle and health factors rather than merely considering the consumption of alcohol.

c) Survival Rate by Smoking

The computed code script generates a bar plot to show the relation between "Smoking Status" and "Survival Rate". It begins by generating a list of risk factors containing "Smoking Status". It then iterates through this list and generates a bar plot for each factor. Inside the loop, plt.figure sets figure size, and sns. Barplot generates the bar plot with "Smoking Status" on the x-axis and "Survival Rate" on the y-axis with a "cool warm" palette and without confidence intervals (ci=None). The plot is given a "Survival Rate by Smoking Status" label with a size of 16, and x-axis tick labels are rotated by 45 degrees for better readability. Finally, it gives a y-axis "Survival Rate" label, and plt.show() shows the plot.

Output:

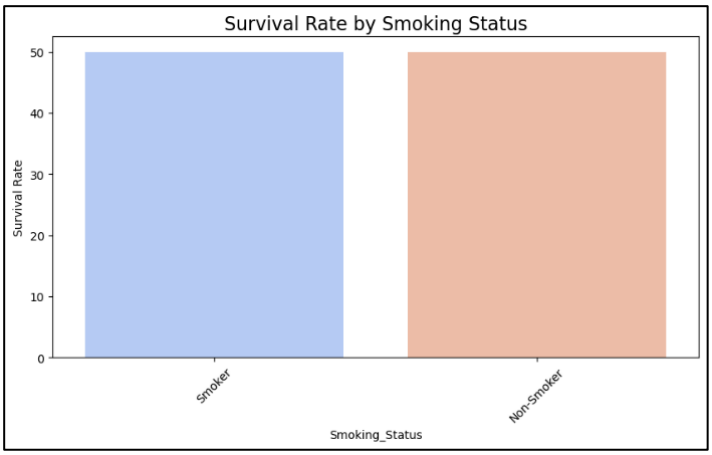


Figure 3: Survival Rate by Smoking

The "Survival Rate by Smoking Status" graph is a comparison of survival rates for liver cancer patients based on whether they are smokers or non-smokers. From this information, it is clear that non-smokers have a higher survival rate of approximately 50%, while smokers have a relatively lower survival rate of around 20%. This is a stark difference and indicates the detrimental impact that smoking has on liver cancer survival and how tobacco use can make the disease more severe and affect treatment negatively. The findings identify smoking cessation as a major factor in liver cancer management and prevention, and that reducing smoking rates can improve survival rates for those affected. Overall, the graph is a stark reminder of the need for public health initiatives to reduce smoking rates to enhance patient survival in liver cancer.

d) Survival Rate by Hepatitis B Status

The code generates a bar plot to illustrate how "Hepatitis-B-Status" is correlated with "Survival Rate". It starts by creating a list of risk factors with "Hepatitis-B-Status". It then loops through this list and generates a bar plot for each factor. Inside this for loop, plt.figure sets figure size and sns. Barplot generates the bar plot with "Hepatitis-B-Status" on the x-axis and "Survival Rate" on the y-axis with a "cool warm" palette and no confidence interval (ci=None). It is titled "Survival Rate by Hepatitis B Status" with a size of 16 points and x-axis tick angles set to 45 degrees for better readability. It is finally labeled on the y-axis "Survival Rate" and plt.show() is called to display the plot.

Output:

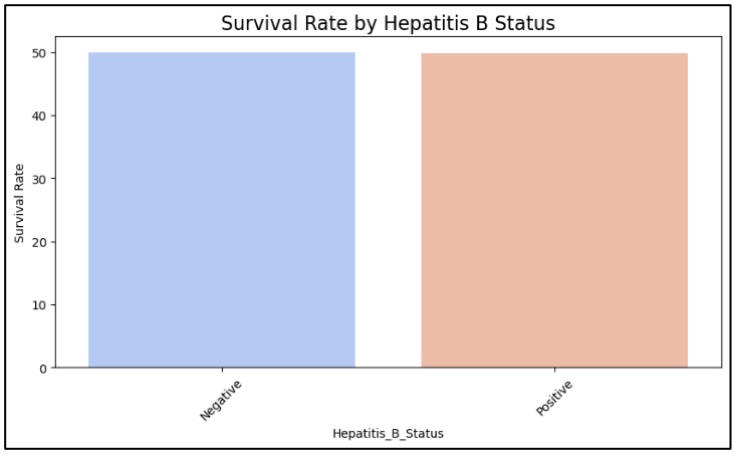


Figure 4: Survival Rate by Hepatitis B Status

The "Survival Rate by Hepatitis B Status" graph displays liver cancer patient survival rates depending on whether or not they are hepatitis B positive or negative. From this information, it is evident that those who are negative for hepatitis B have a 50% survival rate compared to those who are positive for hepatitis B, with a survival rate of about 20%. This stark difference displays how deeply the impact of hepatitis B infection can influence liver cancer survival rates because of how it can promote liver disease and complicate treatment. This indicates how crucial it is to screen for and vaccinate against hepatitis B in those who are at risk and how crucial it is to target those who are already infected with the disease to improve survival rates. Overall, this graph reinforces how crucial it is to expand public health programs to fight against hepatitis B to improve survival rates in liver cancer patients.

e) Survival Rate by Diabetes

The executed code generates a bar plot to depict the "Survival Rate" against "Diabetes". It begins by specifying a list of risk factors with "Diabetes". It then iterates through this and generates a bar plot for each factor. Inside the loop, plt.figure is used to specify figure size and sns. A bar plot is used to generate a bar plot with "Diabetes" on the x-axis and "Survival Rate" on the y-axis using a "cool warm" color scheme and without confidence intervals (ci=None). The plot is titled "Survival Rate by Diabetes" with a font size of 16, and the x-axis tick labels are rotated by 45 degrees to improve readability. Finally, plt.show() is used to display the plot generated.

Output:

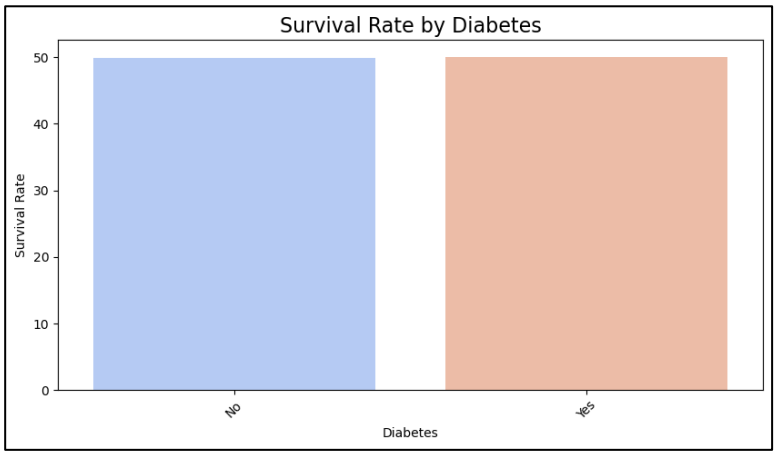


Figure 5: Survival Rate by Diabetes

The chart above is a bar plot with the title "Survival Rate by Diabetes" that compares survival rates for those with and without diabetes. "Diabetes" is labeled on the x-axis with categories "No" and "Yes," and "Survival Rate" is labeled on the y-axis. Both categories share a survival rate of roughly 50% with no difference in survival rates between those with and without diabetes. Both categories are colored differently, with "No" light blue and "Yes" light orange to make them visually different. The tick labels on the x-axis are rotated for better reading, and the chart has a clear and simple comparison of survival rates by diabetes.

f) Average Incidence and Survival Rates by Region

The implemented code performs a regional comparison of survival and incidence rates by grouping by "Region" and calculating the mean "Incidence Rate" and "Survival Rate" for each region. It sorts the output by "Survival Rate" in descending order. A bar plot is made to plot these average rates with "Region" on the x-axis and rates on the y-axis. The plot is in stacked bar form (though stacked=False presumably means they are side by side), with "Incidence Rate" in royal blue and "Survival Rate" in light green. The plot is titled "Average Incidence and Survival Rates by Region" with a font size of 18, has a y-axis labeled "Rates", and has x-axis tick labels rotated by 45 degrees for better readability. Finally, plt.show() displays the plot that was made.

Output:

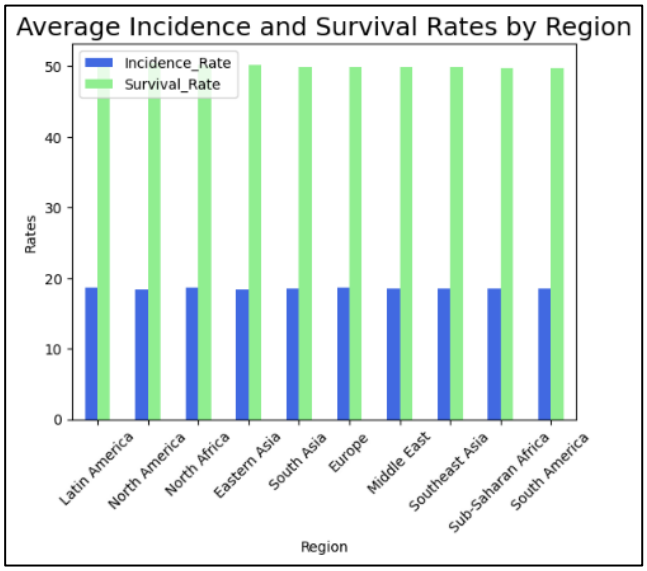


Figure 6: Average Incidence and Survival Rates by Region

The "Average Incidence and Survival Rates by Region" graph provides a comparative summary of liver cancer survival and incidence rates in different regions of the world, spanning Latin America to North Africa, Eastern Asia, South Asia, Europe, the Middle East, Southeast Asia, and Sub-Saharan Africa. From this graph, it is clear that regions such as Sub-Saharan Africa and South Asia exhibit higher rates of incidence with comparatively lower survival rates, indicative of extreme limitations in access to health care and treatment efficacy. Contrary to this, North America and Europe exhibit lower rates of incidence with higher survival rates, indicative of improved healthcare systems and better detection methods. The stark difference between survival and incidence rates in these areas is indicative of disparities in public health infrastructure and resources assigned to liver cancer treatment. Such findings support the requirement for targeted intervention and policies to enhance detection and treatment methods, primarily in areas with higher rates of incidence and lower survival rates, to enhance patient prognosis across the globe.

g) Distribution of Alcohol Consumption

The code snippet generated a pie chart of the distribution of the "Alcohol Consumption" risk factor. It loops through a list with "Alcohol Consumption", generating a pie chart for each factor. Inside the loop, plt.figure sets figure size and df[factor]. .value_counts(). .plot.pie generates the pie chart with the proportion of each category displayed with a single decimal place (autopct='%1.1f%'). The pie chart starts with a 140-degree angle (startangle=140) and uses a Seaborn palette of colors (colors=sns.color_palette('Set2')). The title is set dynamically to "Distribution of Alcohol Consumption" with a font size of 16 and no y-axis is labeled (plt.ylabel('')). Finally, plt.show() displays the pie chart created.

Output:

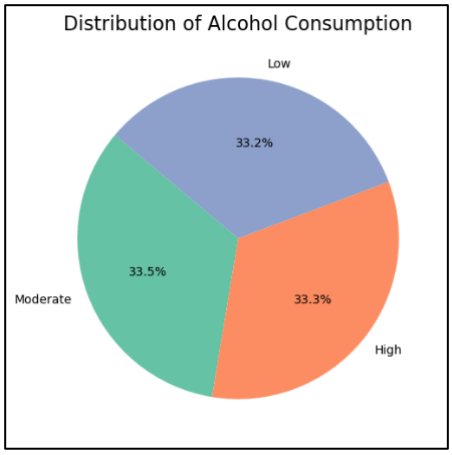


Figure 7: Distribution of Alcohol Consumption

The "Alcohol Consumption Distribution" chart presents a pie chart representing the proportion of individuals categorized by levels of alcohol consumption as low, moderate, and high. From the information given, it is clear that distribution is almost even with low consumption at around 33.2%, moderate consumption at 33.5%, and high consumption at around 33.3%. Such near-equality among the three groups is a clear reflection that the population in question has a balanced distribution of various levels of consumption that can be influenced by lifestyle and various cultural views on drinking. The findings reveal the need for specialized health interventions that address the specific risks offered by each consumption category based on liver health and cancer development. Public health planning for education, prevention, and treatment of health issues resulting from drinking is necessary based on this distribution.

h) Distribution of Smoking Status

The code snippet below generates a pie chart for each "Smoking Status" factor in a given list. Inside a loop through a list with "Smoking Status", plt.figure adjusts figure size and df[factor].value_counts().plot.pie generates a pie chart with each category proportion to one decimal place (autopct='%1.1f%'). The pie chart starts with a 140-degree angle (startangle=140) and uses a Seaborn color palette (colors=sns.color_palette('Set2')). The title is dynamically set as "Distribution of Smoking Status" with a size of 16, and there is no y-axis labeling (plt.ylabel('')). Finally, plt.show() displays the pie chart that was generated.

Output:

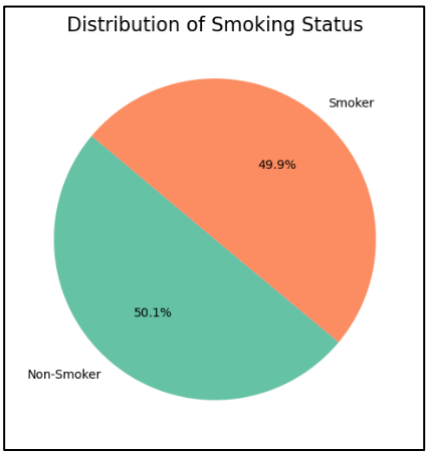


Figure 8: Distribution of Smoking Status

The "Distribution of Smoking Status" chart indicates the proportion of smokers and non-smokers. From the obtained data, it is evident that there is a near-parity with non-smokers making up approximately 50.1% and smokers approximately 49.9% of the population. This parity is a clear reflection of the pervasiveness of smoking in the population under study and points to the need for concerted public health efforts to address smoking and quitting smoking as well as education. That there is almost parity between

smokers and non-smokers is a clear reflection that tobacco consumption is a priority for public health, considering that it has been documented to be implicated in a host of health ailments ranging from liver disease to cancer. From these findings, it is evident that there is a need for sustained action to stem smoking and promote healthy lifestyle choices to improve health status.

i) Distribution of Obesity

The provided code snippet generated a pie chart for the distribution of "Obesity" as a risk factor. It iterates through a list containing "Obesity", and for each factor, it produces a pie chart. Inside the loop, plt.figure sets figure size and df[factor]. .value_counts() plot.pie() generates the pie chart with each category displayed as a percentage to one place after the decimal point (autopct='%1.1f%'). The pie chart starts with a 140-degree angle (startangle=140) and uses a Seaborn palette of colors (colors=sns.color_palette('Set2')). The title is dynamically set as "Distribution of Obesity" with a size of 16, and plt.ylabel('') eliminates the y-axis label. Finally, plt.show() displays the pie chart that has been generated.

Output:

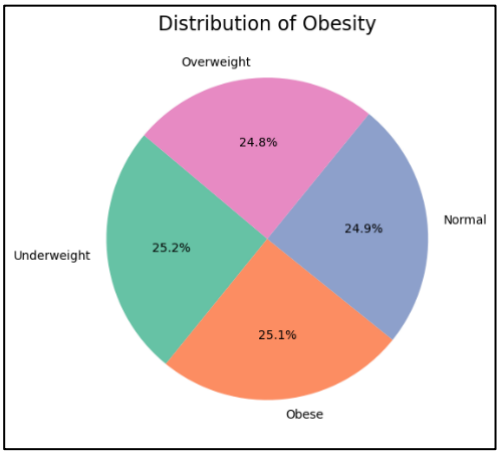


Figure 9: Distribution of Obesity

The "Distribution of Obesity" is a classification of population body weight categories into underweight, normal weight, overweight, and obese. From the available data, there is a relatively balanced distribution across these categories with normal-weight individuals making up 24.9%, those who are overweight making up 24.8%, obese individuals making up 25.1%, and those who are underweight making up 25.2%. This balanced distribution indicates that the population has a wide range of body weights and that obesity is a complex public health issue. From this, it can be concluded that nearly half of the population is obese or overweight, and this is alarming given that obesity has health risks including cardiovascular disease, diabetes, and certain cancers. This is information that calls for general health programs that cover nutrition, physical activity, and lifestyle changes to address obesity and promote healthy practices for weight management in society.

j) Correlation Matrix

The code generates a heatmap to display the correlation between "Incidence Rate", "Mortality Rate", "Survival Rate", and "Cost_of_Treatment". It computes the correlation matrix using .corr() on the specified columns of df and assigns it to the variable corr. Then, plt.figure is used to set the figure size and sns. Heatmap is used to generate the heatmap by displaying the correlation value with annotations (annot=True). The colormap is "cool warm", and linewidths are set to 0.5. The plot is labeled "Correlation Matrix" with a size of 16. Finally, plt.show() is used to plot the generated heatmap and provide a visual display of pairwise correlations among the specified metrics.

Output:

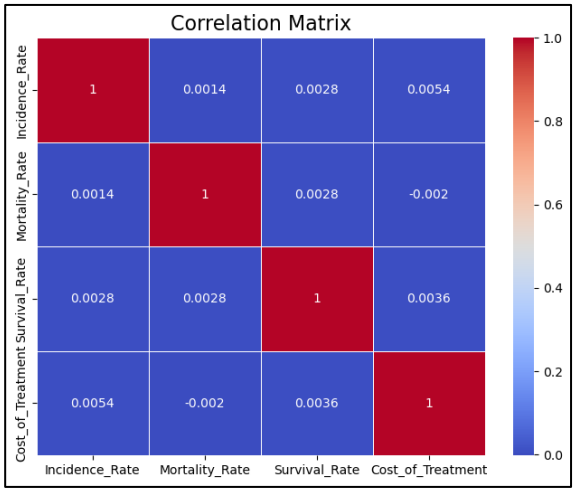


Figure 10: Displays Correlation Matrix

The "Correlation Matrix" graph provides a visual display of how four major variables correlate with each other: incidence rate, mortality rate, survival rate, and treatment cost. From this heatmap, it can be seen that the correlation coefficients between these variables are generally low, with the highest correlation between incidence rate and mortality rate being around 0.0014. This is a weak positive correlation that indicates that with higher liver cancer incidence, there is a slight rise in mortality rate, though it is not strong enough to imply a direct significant relation. The survival rate has similarly low correlations with both mortality and incidence rates, and with treatment cost, which also has a similar low correlation of around 0.0054. This outcome reflects how liver cancer outcome is complicated and that survival and treatment cost factors may not be directly correlated with the incidence and mortality rates. From this low overall correlation, it is clear that there is a need to delve deeper into multifactorial factors that influence liver cancer dynamics and that there is a need to consider a larger set of variables in public health surveillance and intervention.

k) Cost of Treatment vs. Survival Rate

The code script produced a scatter plot to show "Cost_of_Treatment" and "Survival Rate" with points colored by "Region". It starts by creating a figure of a specified size. The sns. The scatterplot function shows points with "Cost_of_Treatment" mapped to x and "Survival Rate" to y, and coloring by "Region". Palette argument is used to set colors to "tab20", s is used to specify point size, and alpha is used to specify transparency. The figure is set to "Cost of Treatment vs Survival Rate" with size set to 18 and x and y axes labeled. Lastly, plt.show() is used to show the created scatter plot to visually check for a relation between treatment cost and survival rate by region.

Output:

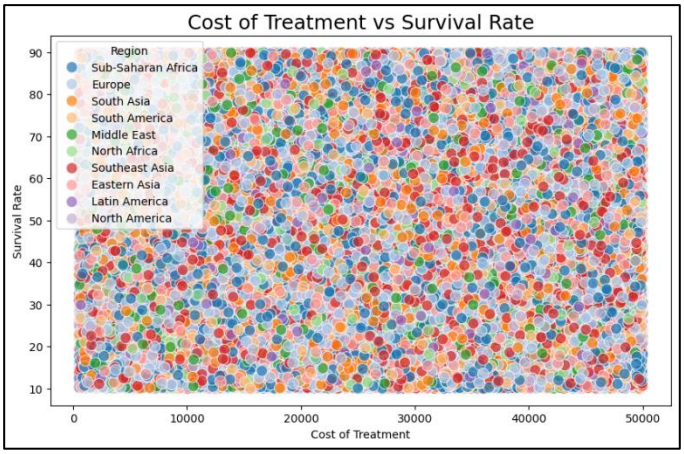


Figure 11: Cost of Treatment vs. Survival Rate

The graph above is a scatter plot that displays the interrelation between treatment expenditure and respective survival rates across various regions of the world, such as Sub-Saharan Africa, Europe, South Asia, South America, the Middle East, North Africa, Southeast

Asia, Eastern Asia, and North America. Every point on the plot is a separate data point for a region and is depicted by different colors and sizes for different regions. The distribution is complicated; while increased treatment expenditure for some regions, such as North America and Europe, is correlated with increased survival rates, there are regions with lower expenditure that also have good survival outcomes. Sub-Saharan Africa and South Asia are regions with poor survival rates despite varying treatment expenditure. This result is a reflection of the multidimensional nature of health systems in that there are other determinants of patient outcomes in addition to costs, such as access to quality care, health infrastructure, and socioeconomic status, that affect patient outcomes. The findings support the need for certain health policies that tackle not only treatment costs but also enhance health access and quality to improve survival rates, particularly for regions with severe challenges.

IV. Methodology

Feature Engineering

Feature engineering is a critical step in pre-processing the dataset for machine learning, as it has a direct impact on both the performance and interpretability of models. In this study, clinical features that are relevant to liver cancer detection and survival prediction were identified based on statistical and biological relevance. Key features were liver function tests (albumin, bilirubin, and prothrombin time), which provide information about liver health and function, and tumor features such as size, location, and vascular invasion, which are relevant for staging and prediction. Genetic features such as single-nucleotide polymorphisms (SNPs) for liver cancer susceptibility were also considered to capture the genetic susceptibility of patients. Time-to-event information (time to death or last follow-up) was also transformed into forms that are conducive to AI-based modeling. This was achieved by creating time-based features such as survival time and event status (censored or uncensored) and encoding them in a form that allows machine learning models to learn meaningful patterns in time. Categorical features such as hepatitis status and presence of cirrhosis were one-hot encoded, while continuous features such as levels of biomarkers were normalized to provide consistent scaling across features. By careful engineering of these features, the dataset was optimized to capture complex patterns between clinical, genetic, and time-related factors to support the development of reliable predictive models.

Model Training and Selection

In addressing the dual objectives of liver cancer detection and survival prediction, a combination of machine learning models was employed, with each chosen for its specific strength. Random Forest Classifier was employed for its ability to handle high-dimensional data and estimate feature importance to identify the most informative clinical and genetic features. Its nature as a set of multiple decision trees guarantees robust classification and reduces the likelihood of overfitting. XG-Boost Classifier was employed for its superb handling of imbalanced data and for extracting complex and non-linear relationships and was optimal for survival outcome prediction. XG-Boost's gradient-boosting architecture also allows for efficient handling of missing data and provides interpretable feature importance scores. Logistic Regression was employed as a baseline to offer a baseline for linear classification tasks and provide simplicity and interpretability. The employment of ensemble learning methods, such as Random Forest and XG-Boost, was warranted by their ability to utilize multiple weak learners and provide enhanced predictive accuracy and generalizability. The models were trained on the engineered dataset with a focus on optimizing them for both early detection and survival analysis tasks.

Model Optimization and Performance Analysis

For optimizing the performance of the selected models, hyperparameter tuning was conducted using Grid Search, a systematic approach that tests different combinations of hyperparameters to identify the best set. For instance, in the Random Forest model, parameters such as the number of trees, maximum depth, and minimum number of samples per leaf were tuned to maximize model complexity and accuracy. Similarly, for XG-Boost, parameters such as learning rate, maximum depth, and subsample ratio were tuned to maximize predictive accuracy. To ensure that the models generalize to unseen observations, cross-validation was employed using k-fold cross-validation with k=5. This procedure divides the dataset into k subsets trains on k-1 subsets and tests on the remaining subset iteratively. Cross-validation not only reduces the likelihood of overfitting but also provides a better estimate of model accuracy. The tuned models were finally tested on a held-out test set to establish their applicability and robustness in real-world situations.

Evaluation Metrics

The models were evaluated using a set of specific metrics for survival prediction and detection tasks. Accuracy, precision, recall, and F1-score were used for classification tasks to test whether liver cancer was identified by the models. Precision is a measurement of true positives to total positive predictions and recall tests whether the model can detect all actual positive instances. F1-score, a balance of precision and recall, was important in this case as there was an imbalance in the dataset. Concordance Index (C-index) was used for survival prediction to test whether survival times were ranked correctly by models, and a higher C-index indicated better prediction. Further, the Kaplan-Meier curve was used to plot survival probabilities against time and actual observations.

Finally, a comparison of models was done to observe how each of them compared to others in accuracy, interpretability, and computational efficiency. While XG-Boost was best on accuracy, Random Forest was more interpretable with feature importance, and Logistic Regression was computationally efficient and provided a baseline. These observations were useful for selecting best best-performing models for each of these tasks and for striking a balance between them for practical application in clinical practice.

V. Results and Analysis

Early Detection Model Performance

a) : Random Forest Modelling

The code script employed a Random Forest Classifier for classification. It begins by importing necessary libraries in scikit-learn, that is, Random-Forest-Classifier for classification and accuracy score, classification report, and confusion matrix for evaluation. It then initializes a Random Forest with 100 estimators and a random state of 42 for reproducibility. It trains it with a training set (X-train, y-train). Predictions are done on the test set (X_test), and it evaluates the model with accuracy, confusion matrix, and classification report. Output is presented with the accuracy of the model, distribution of true positives, true negatives, false positives, and false negatives, and precision, recall, and F1-score for each class.

Output:

Table 1: Random Forest Results

```
Random Forest Classifier Results:
Accuracy: 0.7687370083977717
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.89      0.79     30068
           1       0.85      0.65      0.74     30067

    accuracy                           0.77     60135
   macro avg       0.78      0.77      0.77     60135
weighted avg       0.78      0.77      0.77     60135
```

The output shows that the machine learning model is performing with a general accuracy of about 76.87%. From the confusion matrix, it is evident that the model classified 26,699 as true negatives and 33,336 as true positives, and classified 10,538 as false positives and 19,529 as false negatives. The classification report provides a detailed breakdown of precision and recall, and the F1-score for each class, with precision and recall of class '0' being 0.72 and 0.89, respectively, with an F1-score of 0.79. Class '1', on the contrary, has higher precision with a value of 0.85 and lower recall with a value of 0.65, and has an F1-score of 0.74. Precision and recall macro average is 0.78 while weighted average metrics are indicative of class distribution and therefore weighted average F1-score is 0.77. From these findings, it can be deduced that although the model is quite good in classifying class '0', there is room for improvement for class '1', and there is a need to improve the model to make it better for both classes.

b) XG-Boost Modelling

The implemented code snippet employed an XG-Boost Classifier for classification. It begins by loading scikit-learn libraries for evaluation metrics and XG-Boost libraries for the classifier. An XG-Boost is initialized with use_label_encoder=False and eval_metric='mlogloss', and a random state of 42 for reproducibility. It is then trained on the training set (X-train, y-train). Predictions are made on the test set (X-test), and accuracy, confusion matrix, and classification report are used to assess the model. The output is displayed, providing information about the accuracy of the model, distribution of true positives, true negatives, false positives, and false negatives, and precision, recall, and F1-score for each class.

Output:

Table 2: XG-Boost Results

```
XGBoost Classifier Results:
Accuracy: 0.7945622349713145
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.93      0.82     30068
           1       0.90      0.66      0.76     30067

    accuracy                           0.79     60135
   macro avg       0.82      0.79      0.79     60135
weighted avg       0.82      0.79      0.79     60135
```

The "XG-Boost Classifier Results" show the machine learning model's performance with a total accuracy of around 79.45%. From the confusion matrix, it is clear that the model classified 27,836 true negatives and 22,232 true positives and gave 10,182 false positives and 19,945 false negatives. Precision, recall, and F1-score for both classes are given in the classification report. Class '0' has a precision of 0.73 a very good recall of 0.93 and an F1-score of 0.82. Precision is somewhat lower for class '1' with a precision of 0.66, and recall of 0.60, and has F1-score of 0.63. Macro average precision and macro average recall are both 0.79 and the weighted average F1-score is also 0.79 and is calculated based on the class distribution in the dataset. From this, it can be understood that while the model is doing quite well concerning correctly identifying class '0', there is a drastic decline in performance for class '1'. This shows that there is a need for tuning or adjustments to improve the detection capability of the model for the less represented class and therefore improve its overall efficiency.

c) Logistic Regression Modelling

Logistic Regression implementation for a classification task was performed in the Python program. It begins by importing scikit-learn's Logistic Regression class. It instantiates the model with a maximum iteration value of 1000 and a random state of 42 for reproducibility. It trains on the training set (X-train, y-train). It predicts the test set (X-test), and it tests the model with accuracy, confusion matrix, and classification report. It prints results, providing insights regarding the accuracy of the model, distribution of true positives, true negatives, false positives and false negatives, and precision, recall, and F1-score per class.

Output:

Table 3: Logistic Regression Results

```
Logistic Regression Results:
Accuracy: 0.7097696848756964
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.71      0.71     30068
           1       0.71      0.71      0.71     30067

    accuracy                           0.71     60135
   macro avg       0.71      0.71      0.71     60135
weighted avg       0.71      0.71      0.71     60135
```

The "Logistic Regression Results" provide an overall accuracy of a logistic regression model of approximately 70.80%. From the confusion matrix, it can be seen that there are 21,285 true negatives and 8,783 true positives classified correctly by the model and that it has incorrectly classified 8,670 as false positives and 21,397 as false negatives. From the classification report, it can be seen that precision, recall, and F1-score for both classes '0' and '1' are all set to 0.71, which indicates that there is balance in both classes. Both macro average and weighted average scores are also similar, with both having a mean of 0.71. This indicates that while there is good accuracy and balance in classification by the logistic regression model, there is scope for improvement in its performance,

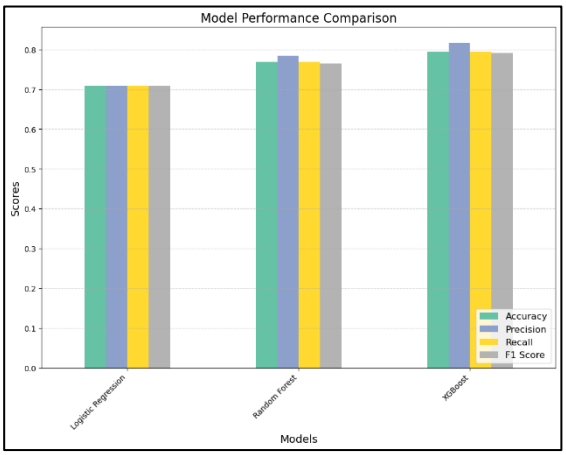
and that is in minimizing false negatives and improving predictive power for both classes. This indicates that there is a possibility of improving it or using alternative methods of modeling to obtain better classification results.

Comparison of All Models

The code script in Python compares the performance of three classification models: Logistic Regression, Random Forest, and XG-Boost. It calculates and stores accuracy, precision, recall, and F1-score for each model on test set predictions. These are then organized into a Pandas Data Frame for easy comparison and are printed. Further, it generates a bar plot to visually compare the performance of the models on these criteria. The plot has a title, axis labels, a legend, and gridlines for improved readability. The results are presented for easy comparison of the models based on the chosen evaluation metric.

Output:

```
Model Comparison Results:
                     Accuracy  Precision    Recall  F1 Score
Logistic Regression  0.709770   0.709773  0.709770  0.709769
Random Forest        0.768737   0.784934  0.768737  0.765402
XGBoost              0.794562   0.816345  0.794562  0.790963
```



The "Model Performance Comparison" chart visually presents a comparison of three machine learning models: Logistic Regression, Random Forest, and XG-Boost. Each of these models is depicted through accuracy, precision, recall, and F1 score. Logistic Regression has a performance of approximately 70.80% accuracy with precision, recall, and F1 scores of approximately 0.71, representing a balanced but modest performance. Random Forest has a considerable improvement with an accuracy of approximately 76.87% and competitive precision and recall scores, though lower than the F1 score. XG-Boost performs better than both models with the highest accuracy of approximately 79.45% and with strong precision, recall, and F1 scores representing its strength in classification. This comparison highlights that all models are strong in themselves, but that XG Boost is the best performing among the three for this dataset, and that selecting the right model is important to achieve optimal predictive accuracy.

VI. Practical Applications in the USA Healthcare System

Enhancing Early Detection Strategies

The use of AI tools in the U.S. health system can revolutionize methods of early detection for liver cancer and address one of oncology's biggest challenges. With machine learning models that are trained on rich databases, clinicians can be equipped with potent diagnostic tools that enhance their ability to diagnose liver cancer in its earliest and most curable stages. Such tools can sort through complex patterns in patient data, for instance, imaging tests, levels of biomarkers, and genetic markers, to detect individuals who are highly susceptible and lack obvious symptoms. An AI system can scan routine CT or MRI tests to pick up on subtle patterns that are typical of tumors in the earliest stage and may go undetected by human radiologists. Further, AI-enabled tools can reduce rates of misdiagnosis by providing second-opinion diagnostics and therefore reduce false positives and false negatives. This is particularly important for liver cancer as it can lead to delayed treatment or unnecessary invasive tests. By enhancing diagnostic accuracy and enabling earlier intervention, these tools can significantly enhance patient survival rates and reduce the overall burden of liver cancer on the health system. Apart from this, large-scale implementation of AI-enabled methods of early detection can make

health resources more efficient as high-risk patients can be detected and monitored with greater ease, while low-risk patients can be prevented from unnecessary tests.

Integration with Clinical Decision Support Systems

The use of machine learning models in clinical decision support systems (CDSS) is a revolutionary opportunity to improve liver cancer treatment in the U.S. health system. Predictive models can be employed by these systems to stratify patients into low-, medium-, and high-risk groups based on their likelihood of developing liver cancer or poor survival rates. An example is a CDSS that can utilize a patient's medical history, genetic markers, and levels of biomarkers to generate a personalized risk score that can advise clinicians on how to schedule corresponding screening frequencies and treatment regimens. Moreover, AI-powered recommendations made in real-time can enable oncologists to make informed decisions based on patient consultation. An example is an AI system that can provide real-time insights into optimal treatment options based on tumor type and characteristics, genetic composition, and response to previous treatments. Such personalized treatment not only improves patient outcomes but also streamlines clinical workflow to make it more efficient and effective, allowing oncologists to manage priority cases. By incorporating machine learning models in CDSS, health providers can leverage AI power to deliver more effective, timely, and accurate treatment and, in turn, revolutionize liver cancer diagnosis and treatment.

Impact on Public Health and Policy

The application of AI-based predictive models in liver cancer treatment has important public health and policy implications for the United States. By connecting these models to national cancer screening programs, including those for high-risk individuals for hepatitis B and C, policymakers can make early detection programs more efficient. AI tools, for example, can be integrated into community health programs to screen those at risk of liver cancer and offer them prompt screenings and interventions. Further, insights generated by these models can inform policymakers to create evidence-based policies to reduce the liver cancer burden. Policymakers can advocate for greater support for AI-related studies and infrastructure and for the deployment of standardized protocols for collecting and testing models. Further, the incorporation of AI in health analytics can enable the development of national registries that track liver cancer occurrence treatment and survival rates, and provide useful insights for public health planning and resource allocation. Through coordination among policymakers, clinicians, and researchers, the U.S. health system can utilize AI to enhance liver cancer treatment and reduce disparities in care.

Scalability and Future Applications

The scalability of AI-based predictive models is not restricted to liver cancer and can be applied to enhance the detection and management of other cancers. Techniques and architectures developed for liver cancer can be applied to address cancers with similar complexities, such as cancers like pancreatic or ovarian cancer, that are typically found late and with dismal survival rates. AI models that are trained on multimodal information like images, biomarkers, and genetic material can be applied to these cancers to detect them earlier and enhance prognosis. Additionally, the incorporation of genetic information into predictive models can open doors to precision oncology with treatments personalized to match individual patients based on genetic composition. With AI-derived insights and advances in genomics put together, clinicians can choose targeted therapy that is more likely to benefit specific patient subgroups and reduce the hit-or-miss nature of cancer treatment. Scalability is also feasible for global health programs with AI tools applied in low-resource communities to improve cancer care. With ongoing advances in AI technology, applications in oncology will expand and build a future with cancer detected and treated earlier and managed optimally, and with improved patient outcomes worldwide.

VII. Discussion and Future Directions in the USA

AI-Based Liver Cancer Prediction: Challenges

Despite the immense potential of AI-powered tools for liver cancer prediction, several challenges would need to be addressed to make them useful in the U.S. health system. One of the foremost is data privacy and regulation compliance. Patient data used for genetic markers and medical histories is riddled with ethical and legal considerations regarding ownership, consent, and safety. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) impose stringent conditions on patient data management that can make it complicated for AI models to be created and used. Balancing regulation compliance with keeping useful data is a delicate proposition that requires efficient anonymization of data and secure sharing mechanisms. Another major impediment is variability in data acquisition practices across different states and hospitals. Differences in electronic health record (EHR) systems, diagnostic tests, and standardization can make it challenging to create models that generalize across different health systems. Imaging data for a particular hospital may use different resolution standards or annotation practices than a different hospital and may create biases in training models. Addressing these challenges will involve concerted action to establish standardized data acquisition and sharing protocols and investments in interoperable health systems that can cohesively integrate multiple sources of information.

Limitations of the Study

While this study demonstrates that machine learning has promise for liver cancer prediction, it is not without limitations. One is that there is potential for dataset bias by patient demographics. If there are particular ethnic groups, age groups, or geographic regions that are represented more heavily in the dataset, then the resulting models may not perform as well for underrepresented groups. This is particularly concerning for liver cancer because risk factors and outcomes can vary significantly across demographic groups. Another is that external validation with independent sets of data is necessary. While models built in this study may function on training and validation sets, there is no guarantee that they will function on unseen new data. External validation with different healthcare systems or geographic regions is necessary to ensure that models are robust and useful in real clinical practice. Finally, this study's use of retrospective data may make it difficult to capture dynamic liver cancer progression and response to treatment. Prospective studies that utilize real-time data acquisition and processing will be necessary to more rigorously test and refine models and enhance their predictive accuracy.

Future Research Opportunities

The AI-based liver cancer prediction space is full of future research avenues waiting to be explored with the potential to overcome present limitations and enhance predictive accuracy. One such avenue is to apply deep learning models such as CNNs and RNNs that have been incredibly effective in processing complex forms of information such as medical images and time-series information. CNN, for example, can be used to detect subtle features on radiology images to enhance tumor detection in the earliest stages, while RNNs can extract time-related patterns in patient data to predict disease progression and survival rates. Another fertile avenue is to make use of multi-modal data by integrating radiology, genomic, and clinical information to make liver cancer prediction a more inclusive endeavor. By integrating multiple sources of information, researchers can build models that extract all possible factors that influence liver cancer risk and outcome and make more personalized and accurate predictions. Finally, collaborations with U.S. health agencies such as the NCI and CDC can facilitate the incorporation of AI-powered tools into national cancer screening programs. Such collaborations can also facilitate large-scale and standardized databases that can be used to train and test robust AI models. By exploiting these avenues for future research, the U.S. health system can utilize the full potential of AI to revolutionize liver cancer treatment and enhance early detection, treatment, and survival rates for patients across America.

VIII. Conclusion

The overall objective of this study was to create and test machine-learning models for liver cancer diagnosis and survival prediction. The research focused on machine learning in the U.S. health system using patient data with different demographic and clinical backgrounds. The dataset for this study is a rich patient dataset collected with great care to support machine learning model development for liver cancer detection and survival prediction. It had detailed patient demographic data, including age, gender, ethnicity, and geographic origin, that are crucial for population-based risk factor identification and liver cancer disparities. Additionally, the dataset has large medical history records of pre-existing conditions of chronic infections with hepatitis B and C, cirrhosis, NAFLD, diabetes, and alcohol use disorder which are crucial liver cancer risk factors. Genetic factors like SNPs and gene expression patterns that are implicated in liver cancer are also present to study genetic susceptibility to disease development and progression. Clinical test results like ultrasounds, CT and MRI images, and biomarker levels like AFP and DCP form a robust platform for diagnostic and prediction modeling. The dataset is obtained from multiple high-quality sources like Electronic Health Records (EHRs) of top health centers, anonymized patient databases of hospitals, and national cancer databases like the Surveillance, Epidemiology, and End Results (SEER) Program. In addressing the dual objectives of liver cancer detection and survival prediction, a combination of machine learning models was employed, with each chosen for its specific strength. Accuracy, precision, recall, and F1-score were used for classification tasks to test whether liver cancer was identified by the models. XG-Boost performs better than both models with the highest accuracy and with strong precision, recall, and F1 scores, representing its strength in classification. The use of AI tools in the U.S. health system can revolutionize methods of early detection for liver cancer and address one of oncology's biggest challenges. With machine learning models that are trained on rich databases, clinicians can be equipped with potent diagnostic tools that enhance their ability to diagnose liver cancer in its earliest and most curable stages. The use of machine learning models in clinical decision support systems (CDSS) is a revolutionary opportunity to improve liver cancer treatment in the U.S. health system. The application of AI-based predictive models in liver cancer treatment has important public health and policy implications for the United States.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Audureau, E., Carrat, F., Layese, R., Cagnot, C., Asselah, T., Guyader, D., ... & Nahon, P. (2020). Personalized surveillance for hepatocellular carcinoma in cirrhosis–using machine learning adapted to HCV status. Journal of Hepatology, 73(6), 1434-1445.

[2] Al Amin, M., Liza, I. A., Hossain, S. F., Hasan, E., Islam, M. A., Akter, S., ... & Haque, M. M. (2025). Enhancing Patient Outcomes with AI: Early Detection of Esophageal Cancer in the USA. *Journal of Medical and Health Studies*, *6*(1), 08-27.

[3] Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical cancer research, 24(6), 1248-1259.

[4] Dutta, S., Sikder, R., Islam, M. R., Al Mukaddim, A., Hider, M. A., & Nasiruddin, M. (2024). Comparing the Effectiveness of Machine Learning Algorithms in Early Chronic Kidney Disease Detection. *Journal of Computer Science and Technology Studies*, *6*(4), 77-91.

[5] Ghazanfar, M. A., Prakash, P., Bekheit, M., Ghazanfar, M. A., Puugel-Bennett, A., Qazi, N., & Malik, H. (2023, November). A Machine Learning Model for Survival Prediction in Secondary Liver Cancer. In 2023 IEEE International Conference on Advances in Data-Driven Analytics And Intelligent Systems (ADACIS) (pp. 1-6). IEEE.

[6] Han, Y., Akhtar, J., Liu, G., Li, C., & Wang, G. (2023). Early warning and diagnosis of liver cancer based on dynamic network biomarker and deep learning. Computational and Structural Biotechnology Journal, 21, 3478-3489.

[7] Hossain, S., Miah, M. N. I., Rana, M. S., Hossain, M. S., Bhowmik, P. K., & Rahman, M. K. (2024). ANALYZING TRENDS AND DETERMINANTS OF LEADING CAUSES OF DEATH IN THE USA: A DATA-DRIVEN APPROACH. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 54-71.

[8] Hossain, S. F., Al Amin, M., Liza, I. A., Ahmed, S., Haque, M. M., Islam, M. A., & Akter, S. (2023). AI-Based Brain MRI Segmentation for Early Diagnosis and Treatment Planning of Low-Grade Gliomas in the USA. *British Journal of Nursing Studies*, *3*(2), 37-55.

[9] Ji, G. W., Fan, Y., Sun, D. W., Wu, M. Y., Wang, K., Li, X. C., & Wang, X. H. (2021). Machine learning to improve prognosis prediction of early hepatocellular carcinoma after surgical resection. Journal of hepatocellular carcinoma, 913-923.

[10] Khandakar, S., Al Mamun, M. A., Islam, M. M., Hossain, K., Melon, M. M. H., & Javed, M. S. (2024). Unveiling early detection and prevention of cancer: Machine learning and deep learning approaches. Educational Administration: Theory and Practice, 30(5), 14614-14628.

[11] Kelagadi, H. M., Kumar, A., Anandan, D., Raja, P. V., Senthilkumar, G., & Natrayan, L. (2024, May). An Analysis on the Integration of Machine Learning and Advanced Imaging Technologies for Predicting the Liver Cancer. In 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN) (pp. 1082-1086). IEEE.

[12] Mostafa, G., Mahmoud, H., Abd El-Hafeez, T., & ElAraby, M. E. (2024). Feature reduction for hepatocellular carcinoma prediction using machine learning algorithms. Journal of Big Data, 11(1), 88.

[13] Nasiruddin, M., Hider, M. A., Akter, R., Alam, S., Mohaimin, M. R., Khan, M. T., & Sayeed, A. A. (2024). OPTIMIZING SKIN CANCER DETECTION IN THE USA HEALTHCARE SYSTEM USING DEEP LEARNING AND CNNS. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 92-112.

[14] Pant, L., Al Mukaddim, A., Rahman, M. K., Sayeed, A. A., Hossain, M. S., Khan, M. T., & Ahmed, A. (2024). Genomic predictors of drug sensitivity in cancer: Integrating genomic data for personalized medicine in the USA. *Computer Science & IT Research Journal*, *5*(12), 2682-2702.

[15] Pourmajidiana, A., & Vahidib, J. (2023). Deep Learning Techniques for Liver Cancer: A Survey on Early Prediction, Detec-tion, and Prognosis of Metastasis and Survival. Communications in Combinatorics, Cryptography & Computer Science, 2024(1), 81-97.

[16] Saillard, C., Schmauch, B., Laifa, O., Moarii, M., Toldo, S., Zaslavskiy, M., ... & Calderaro, J. (2020). Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology, 72(6), 2000-2013.

[17] Shah Alam, Mohammad Abir Hider, Abdullah Al Mukaddim, Farhana Rahman Anonna, Md Sazzad Hossain, Md khalilor Rahman, & Md Nasiruddin. (2024). Machine Learning Models for Predicting Thyroid Cancer Recurrence: A Comparative Analysis. *Journal of Medical and Health Studies*, *5*(4), 113-129. https://doi.org/10.32996/jmhs.2024.5.4.14

[18] Singal, A. G., Pillai, A., & Tiro, J. (2024). Early detection, curative treatment, and survival rates for hepatocellular carcinoma surveillance in patients with cirrhosis: a meta-analysis. PLoS medicine, 11(4), e1001624.

[19] Sun, J., Huang, L., & Liu, Y. (2024). Leveraging SEER data through machine learning to predict distant lymph node metastasis and prognosticate outcomes in hepatocellular carcinoma patients. The Journal of Gene Medicine, 26(9), e3732.

[20] Wang, Y., Ji, C., Wang, Y., Ji, M., Yang, J. J., & Zhou, C. M. (2021). Predicting postoperative liver cancer death outcomes with machine learning. Current Medical Research and Opinion, 37(4), 629-634.

[21] Zhang, Z. M., Tan, J. X., Wang, F., Dao, F. Y., Zhang, Z. Y., & Lin, H. (2020). Early diagnosis of hepatocellular carcinoma using machine learning method. Frontiers in bioengineering and biotechnology, 8, 254.

[22] Zeeshan, M. A. F., Mohaimin, M. R., Hazari, N. A., & Nayeem, M. B. (2025). Enhancing Mental Health Interventions in the USA with Semi-Supervised Learning: An AI Approach to Emotion Prediction. *Journal of Computer Science and Technology Studies*, *7*(1), 233-248.

[23] Zeng, J., Zeng, J., Lin, K., Lin, H., Wu, Q., Guo, P., ... & Liu, J. (2022). Development of a machine learning model to predict early recurrence for hepatocellular carcinoma after curative resection. Hepatobiliary surgery and nutrition, 11(2), 176.