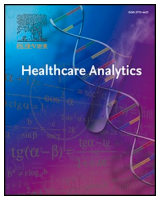
Healthcare Analytics 5 (2024) 100309



Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health





An investigation of the COVID-19 impact on liver cancer using exploratory and predictive analytics

Victor Chang *, Rameshwari Mukeshkumar Patel, Meghana Ashok Ganatra, Qianwen Ariel Xu

Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK

ARTICLE INFO

Handling Editor: Madijd Tavana

Keywords:
Machine learning
Liver cancer
COVID-19
Diagnostic analytics
Artificial intelligence
Health analytics

ABSTRACT

This study presents the influence of COVID-19 and the pandemic on individuals diagnosed with hepatocellular carcinoma and intrahepatic cholangiocarcinoma, the two most common types of primary liver cancer. The study compares the effects before and after the pandemic on these patients. Additionally, it endeavors to predict the likelihood of survival for liver cancer patients. Our research will employ various methodologies to investigate this. Exploratory data analysis techniques are utilized, including univariate analysis, correlation analysis, bivariate analysis, chi-square testing, and T-sample testing. For predictive analytics, machine learning algorithms such as Logistic Regression, Decision Trees, Classification And Regression Tree (CART), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Support Vector Machines (SVMs) will be applied. For our outputs, Logistic Regression and SVMs emerged as top-performing algorithms, boasting a remarkable accuracy rate of 93%. The study reveals that COVID-19 affected all age groups similarly. However, a gender-based difference was observed, indicating that males faced a higher risk of both cancer and mortality. Furthermore, the study found that variables such as year, month, bleeding, cirrhosis, and previously known cirrhosis did not significantly influence patient survival.

1. Introduction

Liver cancer exerts a considerable burden on the resources and healthcare infrastructure of the National Health Service (NHS). Among all types of cancer, the economic Impact of liver cancer (2022) amounted to approximately $1.7 trillion internationally in 2017. Between 2020 and 2040, liver cancer is expected to experience a 55.0% annual incidence increase, potentially resulting in 1.4 million diagnoses and a projected 1.3 million fatalities in 2040 [1]. Liver cancer is typically diagnosed through a combination of physical examinations, blood tests like the Alpha-fetoprotein (AFP) tumor marker and liver function tests, and advanced imaging techniques such as Computed Tomography (CT) scans, MRI, and ultrasound exams. In some cases, a biopsy may be performed, where a sample of liver tissue is collected and examined under a microscope to confirm the presence of cancerous cells. The choice of diagnostic approach depends on the patient's clinical presentation and the need for precise confirmation of liver cancer. Thus, technology plays a crucial role in detecting liver cancer (see Tables 1–3).

Early detection through technology offers several advantages: it enhances treatment outcomes, leading to higher success rates and less aggressive interventions with improved quality of life for patients. Moreover, it significantly increases longevity by extending life expectancy and allowing for the management of liver cancer as a chronic condition [8].

In an era of increasing digitization, vast electronic health data resources hold great potential for valuable applications in predictive analytics. We will explore one of these approaches using machine learning.

1.1. Motivation

Recent figures released by the World Health Organization (WHO) reveal that the total number of deaths linked either directly or indirectly to the COVID-19 pandemic, referred to as "excess mortality," during the period from January 1, 2020, to December 31, 2021, is estimated to be around 14.9 million. The range of this estimate falls between 13.3 million and 16.6 million. Understanding the complex consequences of the COVID-19 pandemic on various facets of healthcare is crucial, which is why the topic " ***Impact of COVID-19 on Liver Cancer through***

* Corresponding author.
E-mail addresses: victorchang.research@gmail.com, v.chang1@aston.ac.uk (V. Chang), rameshwariworks@gmail.com (R.M. Patel), meghana.a.ganatra@gmail.com (M.A. Ganatra), qianwen.ariel.xu@gmail.com, 220247906@aston.ac.uk (Q.A. Xu).

https://doi.org/10.1016/j.health.2024.100309
Received 27 December 2023; Received in revised form 2 February 2024; Accepted 5 February 2024
Available online 12 February 2024
2772-4425/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Evaluation of existing studies on COVID-19's impact on healthcare.

Studies	Machine Learning Model (s)	Results	Gap
Islam et al. (2019)	Random forest classifier	Association between severe cases of COVID-19 and the accelerated progression of liver disease.	The study does not show the impact of COVID-19 on liver cancer.
Delgado-Chaves et al. (2020)	Computational analysis	Delgado-Chaves et al. used computational analysis and gene networks to study the immune response to coronavirus infection.	The study does not directly relate to liver cancer. It provides insights into the use of computational methods but lacks direct findings related to our research focus.
Zimerman et al. [2]	Decision trees	By using decision trees, the study conducted an exploratory investigation of the length of treatment for COVID-19 patients.	The study lacks an evaluation of how liver cancer patients were affected by COVID-19 management regimens.
Miller et al. [3]	Artificial Neural Networks (ANN)	Miller et al. employed real-world data network-based analysis to examine the association between ondansetron use and COVID-19 mortality. Their use of ANN models allowed them to quantify the relationship between medication use and mortality.	This study lacks the potential use of ANN in assessing the Impact of specific interventions on COVID-19 outcomes in liver cancer patients.

Table 2
Statistical analysis results.

Sr.No	Attributes	P-Value	Significance Status
1	Cancer	0.0014	**Statistically Significant**
2	Year	0.1470	**Statistically insignificant**
3	Month	0.5079	**Statistically insignificant**
4	Bleed	1	**Statistically insignificant**
5	Mode of Presentation	7.20 e^-07	**Statistically Significant**
6	Gender	0.0474	**Statistically Significant**
7	Etiology	0.0414	**Statistically Significant**
8	Cirrhosis	0.1111	**Statistically insignificant**
9	HCC_TNM_Stage	4.252 e ^-14	**Statistically Significant**
10	HCC_BCLC_Stage	6.567 e ^-28	**Statistically Significant**
11	Treatment_grps	7.461 e ^ −28	**Statistically Significant**
12	Surveillance_programme	6.567 e ^-05	**Statistically Significant**
13	PS	4.19 e^ −16	**Statistically Significant**
14	Prev_known_cirrhosis	0.5021	**Statistically insignificant**
15	Age	0.0078	**Statistically Significant**
16	Size	3.714 e ^ −07	**Statistically Significant**

Exploratory and Predictive Analytics" was chosen. As the epidemic spread, it became clear that its effects affected not only respiratory health but also other medical disorders. According to Islam et al. [9], due to overburdened healthcare infrastructure and changing patient behavior brought on by the pandemic, liver cancer, a critical worldwide health concern, may encounter interruptions in detection, treatment, and results.

- The COVID-19 pandemic put tremendous strain on healthcare systems worldwide and diverted funds away from normal care, such as cancer screenings and treatments. Patients with liver cancer may experience delayed diagnosis and the start of therapy due to this interruption, which could negatively affect their prognosis.
- Patients' behavior may have changed as a result of lockdown procedures, fear of getting COVID-19, or other factors. Examples include avoiding hospitals or delaying medical appointments. These modifications may lead to the loss of chances for early detection and treatment of liver cancer (Moatar et al., 2022).
- By using exploratory and predictive analytics, this study can identify patterns, trends, and potential risk factors related to liver cancer during the pandemic. Healthcare professionals, decision-makers, and researchers need this information to customize interventions and allocate resources properly [10].
- By using predictive analytics, it is possible to develop models that predict the trajectory of liver cancer patient survival chance.
- This study's findings may help improve methods for controlling the flow of medical care during emergencies. The lessons learned can also improve readiness for future pandemics or other comparable problems.

Table 3
Performance comparison for all algorithms and existing studies.

Algorithm	Accuracy
1. Logistic Regression	93%
2. Random Forest	91%
3. SVM	93%
4. CART	89%
5. KNN	88%
6. ANN	84%
7. CCN with a DenseNet architecture[4]	89%
8. ResNet [5]	76%
9. Random Forest [6]	79%
10. Decision Tree [7]	92%

1.2. Research aim and objective

The main research question is how predictive modeling techniques from business analytics can be applied to optimize healthcare outcomes, specifically to predict the likelihood of survival for patients diagnosed with Hepatocellular Carcinoma (HCC) or Intrahepatic Cholangiocarcinoma (ICC). This question is particularly important in light of the devastation pandemics cause to global healthcare systems.

The main aim of this study is to predict patient survival by utilizing a set of machine learning models. The specific objectives are as follows.

- Examining COVID's (pandemic's) influence on liver cancer-related factors through univariate and bivariate analysis.
- Determining the most crucial survival-related features, analyzing their correlations, and applying machine learning algorithms for patient survival prediction.
- Evaluate various models using performance metrics to identify the most effective one.

1.3. Background

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has triggered a global health crisis on a scale never seen before the COVID-19 pandemic [9]. The severe respiratory difficulties caused by COVID-19 and its highly contagious nature presented major hurdles for healthcare systems around the world, especially in places with few resources. Pharmaceutical companies, governments, and international organizations worked together to produce and distribute many vaccinations against COVID-19 rapidly. The approach to contain the epidemic and achieve herd immunity included widespread vaccination efforts. Variants of SARS-CoV-2 appeared as the pandemic progressed, with some displaying enhanced transmissibility and potential vaccination

escape. This prompted persistent initiatives to monitor the evolving virus landscape using genomic surveillance and vaccination upgrades [2]. The spread of COVID-19 has had far-reaching effects on people's health and the economy, but it has also shed light on deeper societal problems. It became clear that vulnerable groups were disproportionately hit by the virus, highlighting existing health inequalities.

1.4. Research significance

The significance of this research spans multiple dimensions. First and foremost, it is important to note that the research emphasis on vulnerable patient populations and their outcomes in the context of liver cancer during the pandemic is highly persuasive. According to L.F.M. and Cunha (2021), patients with underlying medical disorders, such as those with liver cancer, have been disproportionately affected by the epidemic, reinforcing health inequities [11]. The research can lead to targeted interventions to improve patient care and minimize inequities by examining patient outcomes and identifying relevant factors and access to care.

The study also critically assesses the disruption caused by the pandemic. A thorough understanding of the pandemic's Impact on the treatment of liver cancer depends on the research's effort to identify healthcare service interruptions. The pandemic's extensive Impact extends beyond the field of infectious diseases and has an impact on every facet of healthcare provision. A deep understanding of disruptions can inform future plans for strengthening the resilience of the healthcare system and reducing the effects of unforeseen difficulties, which will promote consistent and efficient patient care [12].

Moreover, the research contribution to preparedness plans for the future will have a lasting impact on healthcare and public health systems worldwide. The pandemic has highlighted the need for adaptable healthcare systems to handle sudden crises [13]. Future preparedness plans may be shaped by the study's findings, making it possible for healthcare systems to anticipate disruptions, maintain essential services, and guarantee the seamless continuity of patient care in emergency situations [12]. Therefore, the research study's multidimensional importance touches on significant healthcare issues related to the COVID-19 pandemic. The significance of this research goes far beyond its immediate application and contributes to a larger conversation about healthcare system readiness and response tactics as healthcare continues to develop.

The paper is organized as follows. Section 2 explores the existing body of knowledge and research relevant to our study. We review the relevant literature on liver cancer, the impact of COVID-19 on healthcare, and predictive analytics. By examining prior work, we identify the context and gaps that our study will address. Section 3 describes the research methodology from data collection to performance evaluation. We discuss the application of data preprocessing, correlation analysis, statistical techniques, and machine learning algorithms. Section 4 and 5 present the results of our research, including data analysis and the performance of machine learning models. Section 6 concludes the paper by summarizing the key takeaways from our research. The contributions of this research, limitations, and future research directions in the field of healthcare analytics are highlighted.

2. Literature review

2.1. Literature background

The COVID-19 pandemic, which first appeared in late 2019, has profoundly impacted many facets of medical research and patient care and generated an unparalleled worldwide healthcare catastrophe. It is critical to consider the larger landscape of healthcare and data-driven research in the pandemic era to contextualize this study.

The COVID-19 pandemic has severely impeded the detection and treatment of liver cancer. There have been reports of screening, biopsy, and imaging operations being delayed due to a lack of available resources and worries about patient safety. The detrimental impact of COVID-19-related delays on patient prognosis has been highlighted by studies emphasizing the necessity of early diagnosis and therapy for better liver cancer outcomes [14]. As a result of the pandemic, telemedicine and remote consultations became practical options for maintaining liver cancer care while reducing the need for in-person contact [15].

A higher risk of death has been linked to COVID-19 infection in patients with advanced liver cancer. Comorbidities related to COVID-19, such as obesity and diabetes, have been linked to worse outcomes for individuals with liver cancer. Long-term prognosis cannot be determined without more research into the effect of COVID-19 on liver cancer recurrence rates and survival outcomes. Data-driven approaches, such as Python-based analysis, have been shown to have promise for studying the correlation between COVID-19 and liver cancer. According to Leenaerts et al. [16], evidence-based decision-making has been aided by using big data analytics to determine risk variables, treatment patterns, and outcomes for patients with COVID-19-positive liver cancer. The severity of COVID-19 among patients with liver cancer has been predicted using machine learning algorithms, which has helped with risk classification. They are reducing the Harm of COVID-19 in Patients with Liver Cancer with Intervention Strategies and intervention options, such as targeted vaccination campaigns and infection prevention measures in healthcare settings. New studies emphasize the need for early COVID-19 diagnosis and the deployment of specialized treatment procedures for patients with liver cancer. Care for patients with liver cancer during the pandemic requires teamwork between oncologists, herpetologists, and infectious disease specialists [17].

According to Yang and Ma [13], patients with cancer and their care were profoundly affected by the disturbance in cancer management induced by hospitals' singular emphasis on COVID-19 patients. Cancer screenings and routine follow-ups, both of which are considered non-urgent medical procedures, were postponed or canceled. Many cancer patients may not benefit as much from their treatment if they have to wait longer to receive crucial therapies. Hospitals also reallocated oncology sections to treat patients with COVID-19, which reduced access to specialized care for the disease. Cancer patients' suffering is compounded when they are forced to postpone lifesaving procedures like surgery (de Graaf et al., 2020). The disturbance also hindered participation in clinical studies, which slowed the development of new cancer medicines and reduced available treatment alternatives. Fear of getting COVID-19 in a hospital setting and ambiguity surrounding the pandemic increased psychological discomfort and anxiety among cancer patients. In addition, cancer patients may not have received sufficient palliative care, which is vital for relieving their symptoms and enhancing their quality of life. Limited access to resources and COVID-19 care discrepancies hampered the cancer management of vulnerable and underprivileged groups [18]. These interruptions may have cumulative effects on cancer survival rates and patient outcomes due to diagnosis, treatment, and specialized care delays. Amidst the difficulties, the need for a well-rounded and flexible approach to healthcare delivery became clear, as did the significance of effective coordination, telemedicine integration, and individualized care plans in the management of cancer patients.

2.2. Literature gap

In our study, the independent variable is Alive_Dead (Alive/Dead) as we try to predict the survival of a patient within the context of Hepatocellular Carcinoma (HCC) and Intrahepatic Cholangiocarcinoma (ICC), and the independent variables are the rest, which will be reduced after feature engineering and finding correlation. It is important to learn from similar literature that uses machine learning models and the kinds of outcomes they produce. In this regard, we examine a number of pertinent research results that make use of various machine learning

techniques.

There are important studies related to this topic that reflect a varied range of machine learning and analytical methodologies. Even though some of these studies do not specifically examine COVID-19's effects on liver cancer, they are still useful models of methodologies and techniques that could be applied to investigate COVID-19's impact on liver cancer and patient survival. A related study conducted by Siyu Zeng et al. [7] aimed to predict the risk of postoperative complications in patients undergoing hepatectomy. Their study emphasized the importance of identifying influencing factors and predicting complications in surgical patients. Using a dataset of 175 liver cancer patients, they employed five machine learning models including logistic regression, decision tree-C5.0, decision tree-CART, support vector machine and random forest. Decision Tree-CART performed the best, obtaining an accuracy of 92.45%. Their findings suggest that variables such as the duration of surgery, the patient's body mass index, and the length of the incision are important influences on the risk of postoperative complications. In addition, Kiani et al. [4] explored the impact of AI assistants on histopathological classification in diagnosing hepatocellular carcinoma. Their study aimed to differentiate between two subtypes of primary hepatocellular carcinoma, hepatocellular carcinoma and cholangiocarcinoma, using hematoxylin and eosin-stained whole--section images (WSI). They developed a deep learning-based assistant to assist pathologists in the diagnostic task, achieving a significant accuracy of 0.885 on the validation set and 0.842 on the WSI independent test set. Importantly, the use of the AI assistant greatly improved the accuracy of a specific subset of nine pathologists with a defined level of experience. In addition to these studies, Rajesh et al. [19] focused on predicting hepatocellular carcinoma (HCC) liver cancer using machine learning algorithms. Their goal was to predict HCC liver cancer using publicly available datasets. They preprocessed the data and used various classification models, including KNN and decision trees. Their results show successful prediction of HCC liver cancer with good accuracy and computational efficiency. Other studies are presented in the following table.

3. Methodology

The study's methodology comprises six primary stages: conducting a review of existing literature, gathering data, conducting exploratory analysis, preparing the data, selecting an appropriate model, and assessing the chosen model's performance [20]. This section will begin by discussing the software and programming language used for our analysis. After that, we will discuss the data-cleaning process that was employed for the analysis.

3.1. Data collection and source

Data collection is a process of gathering and analyzing from different sources with the aim of identifying trends and uncovering answers to potential questions [21]. The data analyzed for this research falls in the category of secondary data. One of the major advantages of using secondary data is its ease of accessibility (Valcheva, 2021) because a considerable amount of this data is hosted on open-source websites. Hence, researchers and interested parties can obtain this data without significant expenses, often for free or at a minimal cost. The data is sourced from a website named Kaggle in CSV (Comma Separated Values) form. Kaggle is a popular online community platform among data enthusiasts, enabling them to collaborate with others, find datasets suitable for projects as well as participate in friendly hackathons [22].

This data was collected in a prospective manner for all patients who were referred to the hepatopancreatobiliary multidisciplinary team (HPB MDT) at Newcastle-upon-Tyne NHS Foundation Trust (NUTH) during the initial 12 months of the pandemic (March 2020 to February 2021). This was then compared to a retrospective observational group of consecutive patients who had presented in the 12 months directly prior (March 2019 to February 2020). The study encompassed newly diagnosed cases of hepatocellular carcinoma (HCC) or intrahepatic cholangiocarcinoma (ICC), which were confirmed through radiological or histological means following international guidelines (Fedesoriano, 2022). Details about the data are in the Appendix.

3.2. Data pre-processing

Data pre-processing involves a stage within the data mining and data analysis process where raw data is converted into a structure that computers and machine learning systems can interpret and analyze effectively (Mesevage, 2021). For our initial analysis, we'll take advantage of Anaconda's in-build library, specifically Pandas.

Pandas serve as a Python package designed for manipulating data, serving a range of purposes, including importing datasets, refining data quality—such as addressing missing values—and performing data aggregation by deriving summary statistics [23]. Different approaches can be used to determine whether data requires cleaning. These methods include examining entries, eliminating duplicates, identifying outliers, and confirming data consistency. For instance, recognizing that "YES", Y″," yes","YeS" convey the same meaning and need to be rectified.

The **info** function was used to gain insights into the attribute types, distinguishing between categorical and numerical variables. The **describe** function is to help understand the general statistics of the data, which provided the following results.

- Most patients showed a symptomatic mode of presentation.
- Age shows a high range of values, with a minimum age of 27–96.
- The majority of patients are suffering from underlying liver disease (Cirrhosis).
- The tumor's largest size is 220 mm.
- Most patients belong to HCC_TNM_Stage 1 and for HCC_BCLC_Stage 'C'.
- The majority of patients belong to Supportive care treatment groups.
- Patients are at a high risk to die.
- The majority were not a part of the Surveillance program.

Missing values refer to values that are absent for variables in a dataset. It is essential to handle missing values as they can influence the quality of a machine learning model [24]. In the case of categorical variables, the missing values are replaced with the mode (most occurring value), while for numeric variables, they are replaced with the mean value. In our analysis, columns with null entries surpassing 50% were eliminated, leading to a reduction in the columns to 18.

3.3. Correlation matrix

In our research, we created a correlation matrix and visualized it with the help of a heat map.

Heatmap is a visualization technique that shows how numeric variables are correlated to each other. The correlation value ranges from −1 to 1, where −1 signifies negative linear correlation and 1 signifies positive linear correlation. A correlation matrix helps in finding variables that can cause multicollinearity. Multicollinearity is a statistical instance that occurs when two or more variables are strongly correlated with each other. (WallStreetMojo, 2019).

- The heatmap (in Fig. 1) reveals a negative correlation between age and survival following Metachronous Distant Metastasis (MDM). This implies that as age rises, the survival rate from MDM tends to decrease.

- Conversely, the remaining variables display neither significant positive nor negative correlations. This absence of strong correlations is favorable as it mitigates concerns about multicollinearity. It suggests

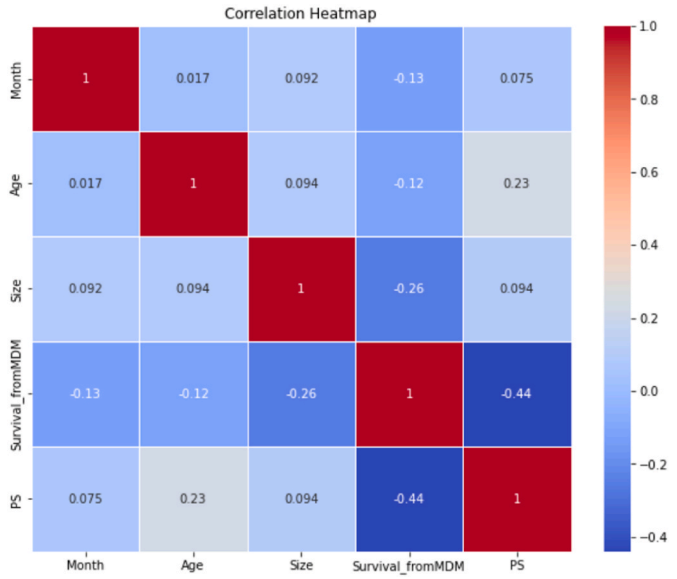


Fig. 1. Heatmap.

that these variables might have an impact when considered collectively and not when considered as an individual.

3.4. Statistical techniques

A statistical hypothesis is a statement that has not been proven yet but can be examined, and a hypothesis test is a method used to determine if this statement holds true or not. In this study, we will define the null hypothesis as that there is no significant effect of the tested features on the survival of the patients with liver cancer and test this hypothesis by interpreting the p-values. The p-value for each variable will be compared to the predetermined significance level (5%). When the p-value falls below 5%, it leads to the rejection of the null hypothesis. This verdict indicates that the collected data offer substantial proof to assert the presence of a noteworthy effect or association between variables. Conversely, when the p-value surpasses α, the null hypothesis remains unaltered. In such circumstances, the data do not furnish enough evidence to affirm that the null hypothesis is incorrect.

3.5. Proposed machine learning algorithms

3.5.1. Logistic regression

Logistic regression is a machine learning algorithm that can be used for both classification and regression problems. In predictive analytics, when the target variable is categorical, it is considered as a classification problem, whereas if it is numeric, it falls under a regression problem. It is particularly important when dealing with categorization target variables such as binary outcomes (e.g., "yes/no" or "1/0″). At its core, logistic regression utilizes sigmoid functions to transform input data into probabilities that lie between 0 and 1. For example, if the predicted probability is greater than 0.5, it is categorized in one category (e.g., "yes"), while a probability less than 0.5 corresponds to another category (e.g., "no").

Logistic regression is known for its ease of implementation, interpretability, and efficiency in training [25]. In addition, logistic regression is highly efficient when it comes to classifying unknown records quickly and tends to deliver good accuracy [25]. In the healthcare industry, logistic regression is widely employed due to its efficacy in addressing binary classification tasks, particularly when dealing with a binary outcome variable featuring two distinct categories. Another advantage of logistic regression is that the data is linearly separable. This means that two categories of data can be effectively separated by a straight line. Consequently, for our binary classification problem, logistic regression stands out as a highly suitable choice.

3.5.2. Decision trees

A decision tree is used for both classification and regression tasks. It is one of the simplest models. It works in a hierarchical way. The data is split repeatedly based on the most crucial feature [26]. The objective of the decision tree is to predict the target variable through simple decisions. The most significant advantage of decision trees is that they are highly intuitive and easy to understand. It requires a lot less data preparation. The decision tree model can be used without going through the usual data preparation steps, such as treating missing values, outliers, data normalization, etc.

Decision trees classify examples through a tree structure, starting from the root and ending at a leaf or terminal node, ultimately assigning the example to one of the categories.

Every node in the tree acts as a condition to assess a particular attribute, and the branches represent potential outcomes or answers related to that condition. This recursive process continues for each subtree originating from the current node.

Decision trees were chosen as a modeling method for several reasons. Decision trees are presented in the form of clear flowcharts, and their interpretability helps in understanding the model's decisions, which is an important factor when communicating results to different stakeholders. In addition, decision trees exhibit robustness when dealing with real-world data challenges. They do not require any additional preprocessing to work efficiently, making them well-suited to handle complex datasets with missing values or outliers. In addition, they have the capability to pinpoint significant features, facilitating the process of feature selection.

3.5.3. Random forest

Random Forest is a supervised machine learning algorithm that can be used for both classification and regression. It is an ensemble algorithm that combines multiple decision tree models to make predictions instead of one model (Rodriguez-Galiano et al., 2015).

The ensemble uses two types of methods: Bagging and Boosting.

- **Bagging:** Bagging is also known as bootstrap aggregating. A number of subsets are created from the training subset with replacements. Further, a number of decision tree models are created, with each model generating its own output. A final prediction is determined through a majority of voting for classification or averaging for regression problems (E R, 2021).
- **Boosting:** Boosting is a process where a number of weak learners are combined together and trained to improve their performance and transform into strong learners. Weak learners are decision trees in the ensemble that perform only slightly better than random chance.

Random Forest has several key advantages that make it a reliable choice for our analysis. Notably, it typically has higher accuracy compared to individual decision trees, thus improving the precision of our predictions. Importantly, Random Forest exhibits robust performance even without extensive hyperparameter tuning, thus simplifying the modeling process. In addition, it is effective in solving the overfitting problem, which is a frequent difficulty with single decision trees. By harnessing the power of ensemble learning, Random Forest mitigates this problem and ensures that our analysis results are more reliable and consistent.

3.5.4. Classification and regression trees (CART)

CART stands for classification and regression trees. Introduced by Breiman et al. [27], it works in a similar manner as decision trees and is one of the algorithms used by decision trees. The only difference between a decision tree and CART is that the latter uses the Gini index to split the nodes. The Gini index is the measure of impurity. It is the degree

of probability that a variable is classified in a category. It ranges between 0 and 1, where 0 indicates a pure homogenous group, whereas 1 indicates there is maximum heterogeneity and data is classified randomly. Hence, while splitting the nodes, the split with the lowest Gini Index is taken into consideration. The cart does not require specific details of data distribution.

3.5.5. Support Vector Machines (SVMs)

SVM stands for Support Vector Machines, which can be used for both classification and regression tasks, but it works best for classification problems. The goal of SVM is to find the best hyperplane with a maximum margin that separates the classes into two distinct groups [26]. The hyperplane needs to have a maximum margin as it indicates the level of confidence in the classification. Margin is the distance between the hyperplane and the closest data point.

Support Vector Machines (SVMs) were chosen for several reasons. SVMs have become an important tool for cancer diagnosis and prognosis, which coincides with the context of our study. Of particular note, SVMs excel in handling high-dimensional data. In addition, SVM reduces susceptibility to overfitting and ensures that our model generalizes well to unseen data.SVM creates a linear function that maximizes the separation of classes using strategically chosen support vectors, thereby improving classification accuracy [28].

3.5.6. Artificial neural networks (ANN)

ANN stands for Artificial Neural Networks. This algorithm is inspired by the structure of the human biological system of neurons. Neural Networks consist of a network of nodes that are connected by arcs that resemble dendrites and synapses [29]. The fundamental structure of an ANN includes three key components: the input layer, hidden layers, and the output layer. While the input layer receives raw data, the hidden layers, which can be multiple in number, perform complex computations. Finally, the output layer produces the network's results. The operation of an ANN involves the propagation of data through its layers. Initially, input neurons receive data, each associated with specific weights. These weighted inputs are summed, and a bias is added. Subsequently, an activation function is applied to this value. Based on the resulting output, neurons decide whether to transmit the signal to the next layer or not. This iterative process continues until the final neuron in the output layer generates the network's output.

There are several key advantages to using artificial neural networks (ANNs) in our research. First, ANNs excel at modeling complicated nonlinear interactions, a capability that linear models frequently struggle with, to capture detailed patterns within data. In addition, ANNs can learn relevant features independently from input data, which lessens the need for manual feature engineering. This works particularly well when dealing with high-dimensional data. ANNs also demonstrate an ability to achieve remarkable accuracy in binary classification tasks. Importantly, ANNs have the ability to generalize from their training data efficiently, allowing them to generate accurate predictions on previously unknown data. This reduces the chance of overfitting and improves their ability to handle datasets from the real world. Finally, the parallel processing efficiency of AI networks speeds up the inference and training phases. When working with complicated models and large datasets, this is quite useful.

3.6. Performance evaluation

For any machine learning algorithm, it is important to evaluate the performance of the model [30]. A number of performance metrics can be used for classification problems. We will have a look at metrics such as Accuracy, Precision, Recall, Specificity and F1 score for our research. However, our primary performance matrix would be accuracy. In order to understand the concept of performance metrics, it is crucial to understand the following terminologies.

True Positive (TP): When the data point is correctly classified as positive, it falls under the category of true positive [31].For instance, if a person takes a test for a disease and the model predicts that the person has the disease. Later, the test confirms that the person has the disease; this situation is considered a "True Positive". In simple terms, the model was able to identify the presence of the disease correctly and that's True Positive.

True Negative (TN): When the data point is correctly classified as negative, it falls under the category of true negative [31]. For instance, when the person takes a test and the model predicts that the person does not have the disease. Later, the test confirms the same; this situation is considered a "True Negative." In simple terms, the model was able to identify the absence of the disease correctly and that's True Negative.

False Positive (FP): When the data point is wrongly placed in the category of being positive, it falls under False Positive. If a person takes the test and, the model predicts the presence of a disease. However, the test results reveal the person is disease-free; this situation is termed as False Positive. In other words, the model made a mistake by wrongly predicting the presence of a disease when the person was, in fact, disease-free. A false positive is considered as Type I error.

False Negative (FN): Let's say the person has a disease and the model predicts they are disease-free. This situation falls in the category of False Negative. In simple terms, the model incorrectly predicted the absence of a disease when, in reality, the person has the disease. A false Negative is considered a Type II error.

Confusion Matrix: A confusion matrix (Fig. 2) provides the terms mentioned above in a tabular form. For a binary target variable, it would be a 2*2 matrix, as shown below.

3.6.1. Performance metrics

1 **Accuracy:** It is one of the most common evaluation metrics used in classification problems [32]. Accuracy is used to evaluate how well the overall model is performing. It is used to check how accurately the model made all predictions. It is a ratio of correctly classified data points over all data points. It is given by the formula.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

2 **Precision:** Precision is a measure that indicates the instances of correctly classified positives in relation to all positives [29]. For instance, it would be the measure of patients that the model correctly identified as having a disease out of all the patients actually having the disease. It's necessary to understand precision for situations where a person is healthy. However, we start treating him for the disease due to the model's incorrect prediction. The formula to calculate precision is

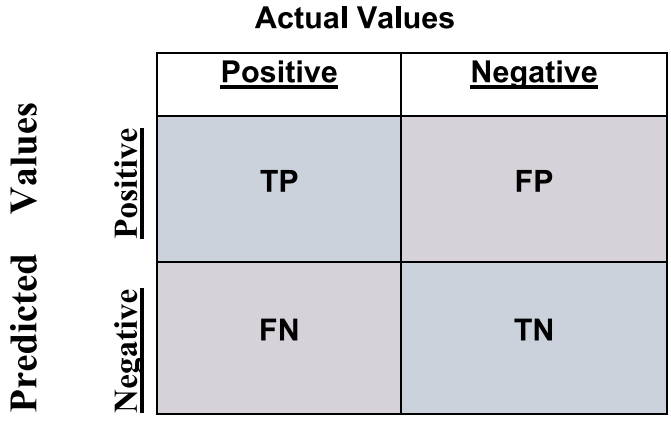


Fig. 2. Confusion matrix.

$$Precision = \frac{TP}{TP + TN} \quad (2)$$

3 **Recall:** Recall is also known as sensitivity. It is the measure of correctly identifying true positives. Imagine a patient has a disease and does not get treatment as the model predicted otherwise [33]. For such instances, it is important to understand recall. It is given by the formula:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

4 **Specificity:** It is the measure of correctly identifying a true negative. For instance, a person is disease-free, and the model predicts the same. It is the ratio of correctly identified negative cases to all actual negative cases. A higher specificity indicates that the model correctly identifies true negatives, while a lower specificity suggests that the model may incorrectly classify some negatives as positives. It is given by the formula:

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

5 **F1- Score**: The F1-score is a metric that combines precision and recall into a single value. It is often referred to as the harmonic mean of precision and recall. It is particularly advantageous when the class distribution is unbalanced. F1 Score is given by the formula

$$F1 = \frac{2 * (precision * recall)}{precision + recall} \quad (5)$$

The selection of a performance metric should be tailored to the unique objectives and limitations of the problem at hand. For instance, when dealing with a cancer diagnosis, placing greater emphasis on achieving high recall may be preferred to reduce the risk of overlooking actual cases, even if it results in more false positives. Nevertheless, it is essential to carefully assess the clinical context and consequences associated with each metric before making a decision in this regard.

3.6.2. Implementation of machine learning algorithms

A few important actions were taken before running the models to ensure they functioned properly.

1 **Feature Selection**:

The process of reducing the number of features provided as inputs to the machine learning algorithm to build a predictive model is known as feature selection [34]. It plays an important role in enhancing the predictive power and accuracy of the model ([35].). Statistical tests like the z-test and chi-square test were conducted to identify the most critical variables for the model. In addition to this, Information Gain was also employed to identify the same. Initially, the dataset consisted of 18 columns, which were reduced to 7 after running the tests. The advantages of feature selection are.

a) **Reduced computational time**: When a model is fed with fewer variables as input, it takes less time to run the model.
b) **Improves accuracy:** Fewer variables means less data to misclassify the datapoints and hence helps improve accuracy.

2. Label Encoding:

The data needs to be pre-processed in a number of ways to make it appropriate for the model to understand. Label encoding is one of the ways of transforming categorical variables into numeric format. Many machine learning algorithms require numerical inputs; label encoding helps in making such data useable for modeling. The process of label encoding involves converting each category of the categorical variable into a unique integer. For example, an attribute named "Shirt" has categories of shirt sizes S, M and L. Label encoding will assign each category with a unique integer label and might convert it to 0,1,2 for S, M and L, respectively. Label encoding is straightforward to implement, making it a popular choice for handling categorical data.

3. Data partitioning:

Data partitioning is an essential step before running a machine learning model. It serves two primary purposes: training the model and assessing its performance. Typically, the dataset is divided into two main subsets – the training set and the test set. The split ratio commonly used is 80:20 or 70:30, depending on the specific problem and dataset characteristics. For our analysis, we have adopted an 80:20 split, with 80% of the data allocated for model training and the remaining 20% reserved for evaluating the model's performance. This division ensures that the model learns from a substantial portion of the data while having an independent dataset to evaluate its predictive capabilities accurately.

Training set: The training data is used to train the model to learn prediction patterns. A significant portion of data is reserved for the training set to ensure better training for the model. A well-trained model results in accurate predictive performance.

Testing set: After the model is trained, it is important to test its performance. This is where the test set is used; it helps evaluate the model on unseen data. This data is not used in training the model and hence it helps us understand how well the accuracy of our model is. 20 % of the data was kept aside for the test set.

The training set is used to train the model, while the testing set is used to test the model. To split the data into train and test, scikit-learn train_test_split function has been used.

4. Results

4.1. Preliminary data analysis

Data Visualization: The process of presenting information in a graphical form with the use of charts, graphs, etc, is known as data visualization [36]. Data visualization means presenting information visually so that it can be easily understood by everyone, quickly absorbed, and highly efficient in conveying the message [37].

1. CountPlot
   - A count plot is a bar graph in Fig. 3 that displays the count of each category in each bar. (Agrawal, 2022).
   - A count plot can be used to show one categorical variable at a time or compare two categorical variables using a parameter called hue, which adds a different color to the category. (TrainDataHub,2022)

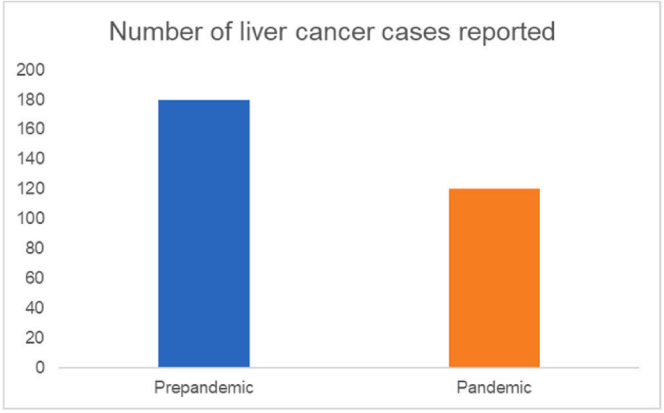


Fig. 3. Countplot of year.

- The count plot illustrates that fewer cancer cases were detected during the pandemic.

- Around 180 cancer cases were identified before the pandemic, represented by the blue bar, while this number decreased to 120 during the pandemic, indicated by the orange bar.

2. Box Plot

A box plot is a graph that visualizes numerical data. It uses a box and lines to show the distribution of the data. Box plot helps identify outliers in data. An outlier is a data entry that is significantly different than other entries. Whether it is necessary for an outlier to be dropped or treated varies depending on the circumstances.

The box represents the interquartile range (IQR), with the edges denoting first (Q1) and third quartile (Q3). The line within the box denotes the median in Fig. 4.

- The majority of values in the size (tumor) column are within the interquartile range, but a few values exist outside of this range, suggesting the presence of outliers.
- Some tumors can grow to significant sizes, which might not be a data entry error, and eliminating them might result in losing important information.
- Hence, despite being outliers, we won't be eliminating these entries due to their medical nature.

Visualization helped in addressing the following questions.

a) Comparing cancer cases during a pandemic and pre-pandemic

- From the bar graph in Fig. 5, there has been an approximate 30% decline in the number of reported cases during a pandemic.

b) Comparing Mode of Presentation

- During the pandemic, there was a notable decrease of around 50% in cases detected through incidental and surveillance methods, as shown in Fig. 6.

- The number of symptomatic patients experienced only a slight decline.

c) Did COVID-19 have a more impact on one gender compared to the other?

- The bar graph in Fig. 7 depicts that COVID-19 had more Impact on males than females during the pandemic.

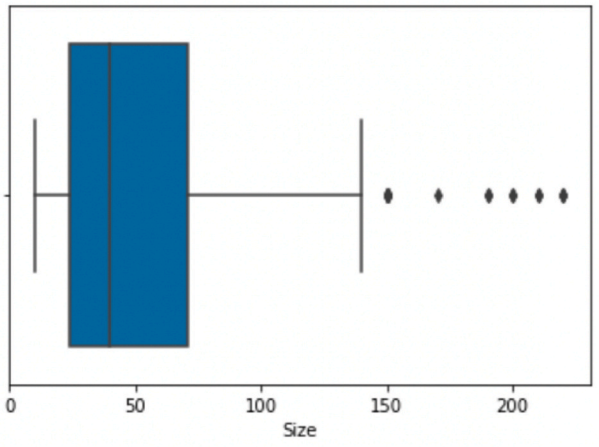


Fig. 4. Boxplot of tumor size.

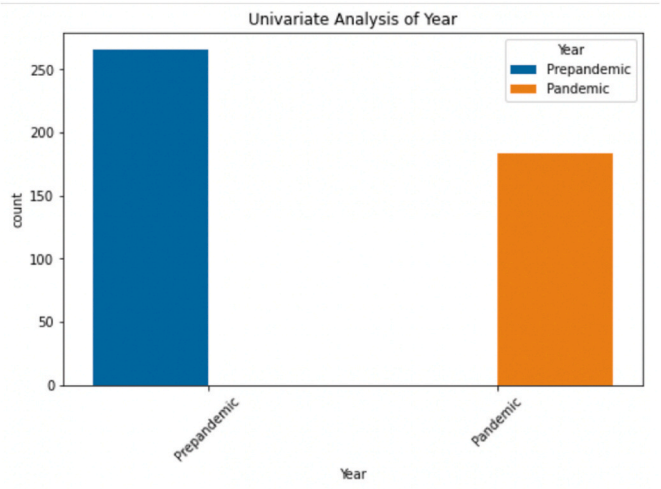


Fig. 5. Count plot of cancer cases during the pandemic and pre-pandemic.

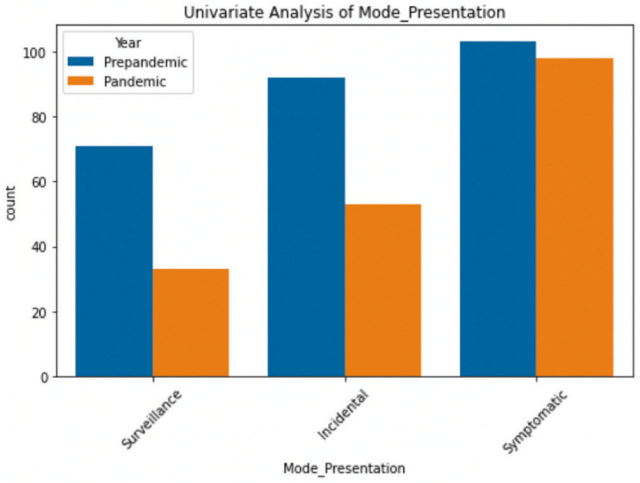


Fig. 6. Count plot of Mode of presentation during pandemic and pre-pandemic.

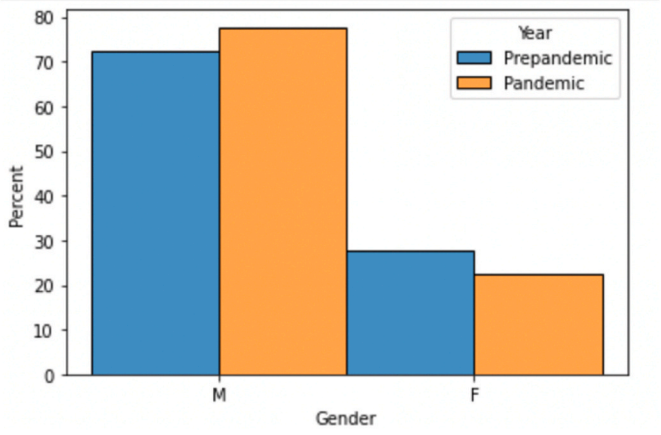


Fig. 7. Bar graph of gender-based comparison to pandemic and pre-pandemic.

- One explanation for this trend can be that alcohol consumption is more culturally accepted among men than women.
- Furthermore, it's possible that during the pandemic, men increased their alcohol intake due to various reasons such as stress, workload, etc, which could have contributed to a rise in liver cancer cases.

d) Did COVID-19 impact Hepatocellular Carcinoma (HCC) at the Barcelona Clinic Liver Cancer (BCLC) Stage (Liver Damage)?

BCLC stands for "Barcelona Clinic Liver Cancer." BCLC stages refer to a staging system used to classify and stage hepatocellular carcinoma (HCC). The BCLC staging system is widely used for assessing the severity of liver cancer and guiding treatment decisions.

The BCLC staging system divides HCC into several stages, from 0 to D. Stage 0 is the Very Early Stage. In this stage, the tumor is small, and the liver function is well-preserved. Curative treatments like surgical resection or ablation may be recommended. Stage A is the Early Stage. This stage includes single or small tumors and well-preserved liver function. Curative treatments are often considered. Stage B is the Intermediate Stage. Patients in this stage may have larger tumors or more extensive involvement of the liver. Treatments such as transarterial chemoembolization (TACE) may be recommended. Stage C is the Advanced Stage (C). This stage includes patients with more advanced liver cancer, compromised liver function, or cancer that has spread to blood vessels or lymph nodes. Treatment options may include targeted therapy or clinical trials. Stage D is the End-Stage. This stage represents patients with advanced liver cancer and significant liver dysfunction. Palliative care and supportive measures are usually the primary focus [38].

- Bar graph in Fig. 8 depicts a rise in detection of stage C BCLC, which falls under the category of severe liver damage.
- There could be a delay in tests and appointments, contributing to the progression to stage C.

e) Did COVID have an impact on the size of tumor?

- The box plot in Fig. 9 shows that the average tumor size during the pandemic was approximately 50 mm, which is larger compared to the pre-pandemic period.
- This suggests there might have been difficulties in timely tumor detection during the pandemic due to overwhelmed hospitals.
- Therefore, larger tumors could have been detected, indicating potential delays in diagnosis and treatment.

f) Did Covid have an impact on the age distribution of patients?

- The histogram in Fig. 10 shows that the average age of patients during and pre-pandemic is around 70 and there is no impact of covid on age distribution.

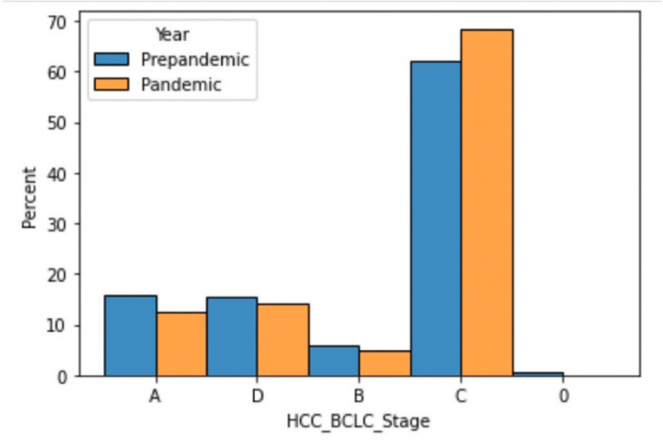


Fig. 8. Bar graph of HCC_BCLC_Stage in comparison to pandemic and pre-pandemic.

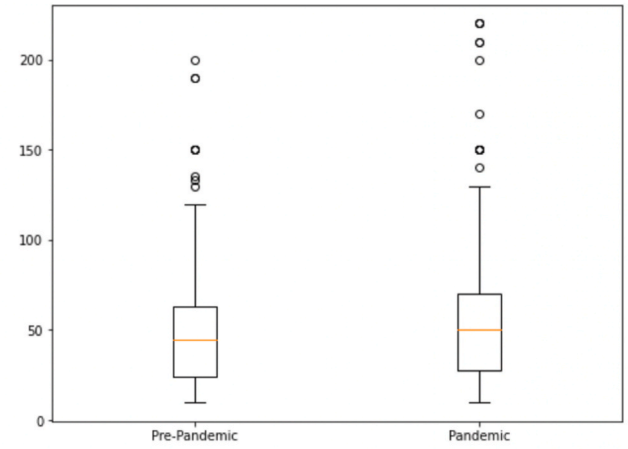


Fig. 9. Boxplot of tumor size in comparison to pandemic and pre-pandemic.

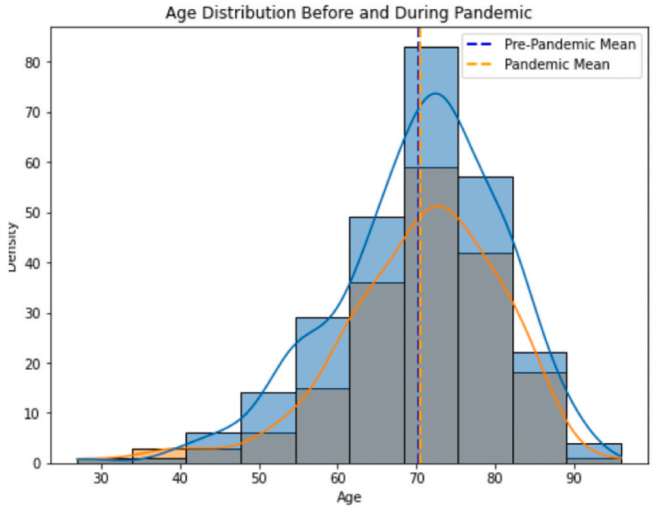


Fig. 10. Histogram of age distribution in comparison to pandemic and pre-pandemic.

4.2. Results of statistical tests

This section discusses the significance values and significance status of each variable in relation to the target variable (Alive/Dead).

The above table shows that all attributes except for Year, Month, Bleed, Cirrhosis and Prev_known_cirrhosis have an influence on the survival of the patient.

This analysis serves as a valuable tool for prioritizing the importance of different attributes when it comes to predicting patient survival.

Evaluating the statistical significance of each attribute enables us to differentiate between factors that are likely to be significant predictors of survival and those that may not have a substantial influence on whether a patient survives.

Attributes categorized as "Statistically Significant" hold particular importance in predicting survival. These attributes are supported by strong statistical evidence, indicating that changes in these factors significantly affect patient outcomes.

On the other hand, attributes categorized as "Statistically Insignificant" may not play a significant role in determining patient survival within the specific context of the study. While they may still be valuable for other purposes or in different scenarios, their impact on survival outcomes in the studied population is limited.

4.3. Results of machine learning algorithm

In this section, we present the results of the machine learning algorithms. In the research, six machine learning algorithms were used to analyze the data, i.e., Logistic Regression, Decision Tree, Random forest, KNN, SVM, and ANN. We will compare the performance of the above-mentioned models. We will evaluate the overall accuracy of the model as well as other metrics such as precision and recall.

1. 4.3.1 Logistic regression confusion matrix:

The logistic regression algorithm achieves an accuracy rate of 93% (Fig. 11), indicating that the model is highly effective at making predictions (see Fig. 12).

A precision of 96% suggests that when the model predicts that an instance will be "Alive," it is usually accurate 96% of the time. In this case, about 96% of the occurrences the model classified as "Alive" were true positives (correctly predicted as such). In contrast, the other 4% were false positives (incorrectly predicted as such when they were actually "Dead"). The overall precision is 93%.

A recall score of 93% indicates that the model correctly labels 93% of the cases that actually fall under class 1 (in this case, "Alive")

A specificity score of 89% indicates that the model is effective at correctly identifying instances of the negative class, which in this case represents individuals who are "Dead". When the model makes a prediction that someone is "Dead," it is correct about 89% of the time.

This indicates that the model can easily avoid false alarms for the "Dead" category, minimizing the instances where it predicts "Dead" incorrectly when individuals are alive.

Overall, the model demonstrates strong performance in the classification task.

2. 4.3.2 Random Forest

The Random Forest algorithm achieves an accuracy of 91%, which is slightly below the accuracy of the logistic regression model. In more detail, the overall precision rate for this algorithm is 90%, indicating that when it predicts a positive outcome, it is correct approximately 90% of the time. Additionally, the recall rate is 92%, which signifies that the algorithm effectively captures about 92% of all true positive cases. Furthermore, the algorithm demonstrates an F1 score of 91%, a harmonic mean of precision and recall. Lastly, the specificity score is 89%, indicating that the model correctly identifies negative instances at an accuracy of 89%.

In summary, while the Random Forest algorithm achieves a slightly lower accuracy compared to logistic regression, it maintains strong performance in terms of precision, recall, F1 score, and specificity, making it a valuable choice for analysis.

4.3.1. SVM

The Support Vector Machine (SVM) delivered results that were on par with the logistic regression model, with both achieving an impressive accuracy rate of 93%, see Fig. 13. Notably, the precision, recall, and F1 score for SVM also stand at 93%. These identical results in accuracy, precision, recall, and F1 score between SVM and logistic regression suggest that both models are highly effective in this context. They exhibit strong performance and are equally well-suited for the specific task at hand, highlighting the robustness of the SVM algorithm in classification tasks.

4.3.2. CART and decision trees

Both CART (Classification and Regression Trees) and Decision Trees (Fig. 14) yielded identical results, so we are presenting the results specifically for CART.

CART achieved an accuracy rate of 89%, coupled with an overall precision score of 88% and a recall rate of 89%. Furthermore, its specificity stood at 88%.

5. KNN

The K-nearest neighbors (KNN) model in Fig. 15 delivered an accuracy rate of 88%, showcasing its ability to make correct predictions in the majority of cases.

When it came to precision, KNN achieved a score of 87%.

The recall rate for KNN was also strong at 88%, implying that it effectively identified about 88% of all true positive cases. Furthermore, the F1 score for KNN stood at 87%, demonstrating a balanced blend of precision and recall.

4.3.3. ANN

The accuracy of the ANN model (Fig. 16) is 84%, which is the lowest among all the models used in this analysis. This could be attributed to the fact that ANNs are generally more effective when dealing with very large datasets.

```
In [64]: from sklearn.metrics import confusion_matrix

         cm = confusion_matrix(y_test,y_pred)
         cm

Out[64]: array([[16,  1],
                [ 2, 26]])

In [65]: # Generate a classification report
         class_report = classification_report(y_test, y_pred)
         print('Classification Report:\n', class_report)

         Classification Report:
                       precision    recall  f1-score   support

                    0       0.89      0.94      0.91        17
                    1       0.96      0.93      0.95        28

             accuracy                           0.93        45
            macro avg       0.93      0.93      0.93        45
         weighted avg       0.93      0.93      0.93        45
```

Fig. 11. Code snippet of the logistic regression confusion matrix.

```
In [68]: from sklearn.metrics import accuracy_score, classification_report

         # Create a Random Forest classifier
         rf_classifier = RandomForestClassifier( oob_score=True,random_state=42)

         # Train the classifier on the training data
         rf_classifier.fit(X_train, y_train)

         # Make predictions on the test data
         y_pred = rf_classifier.predict(X_test)

         # Evaluate the classifier's performance
         accuracy = accuracy_score(y_test, y_pred)
         print(f'Accuracy: {accuracy:.2f}')


         # Generate a classification report
         class_report = classification_report(y_test, y_pred)
         print('Classification Report:\n', class_report)
```

```
Accuracy: 0.91
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.94      0.89        17
           1       0.96      0.89      0.93        28

    accuracy                           0.91        45
   macro avg       0.90      0.92      0.91        45
weighted avg       0.92      0.91      0.91        45
```

Fig. 12. Code snippet of Random Forest algorithm.

```
In [69]: from sklearn.svm import SVC

         # Create an SVM classifier
         svm_classifier = SVC(kernel='rbf', random_state=42)
         # Train the classifier on the training data
         svm_classifier.fit(X_train, y_train)

         # Make predictions on the test data
         y_pred = svm_classifier.predict(X_test)

         # Evaluate the classifier's performance
         accuracy = accuracy_score(y_test, y_pred)
         print(f'Accuracy: {accuracy:.2f}')

         # Generate a classification report
         class_report = classification_report(y_test, y_pred)
         print('Classification Report:\n', class_report)
```

```
Accuracy: 0.93
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.94      0.91        17
           1       0.96      0.93      0.95        28

    accuracy                           0.93        45
   macro avg       0.93      0.93      0.93        45
weighted avg       0.93      0.93      0.93        45
```

Fig. 13. Code snippet of SVM model and accuracy.

The algorithms presented in the comparison table for predicting liver cancer risk utilize clinical variables. From the above table, it can be inferred that both logistic regression and SVM algorithms deliver the best results with the highest accuracy of 93%. Random Forest comes in second with an accuracy of 91%, while CART gives an accuracy of 89 %. KNN provides the lowest accuracy of 88%. Our research, employing logistic regression and SVM, achieved an accuracy of 93%, which is consistent with or even surpasses the accuracy attained by these algorithms in previous studies related to the prediction of liver cancer risk based on clinical variables. This comparison highlights the effectiveness and suitability of our selected algorithms for the specific task of predicting liver cancer risk within the context of clinical variables.

Examining the literature review, the most suitable model for our research was the SVM, which emerged as the most commonly employed

```
In [70]: from sklearn.tree import DecisionTreeClassifier

In [71]: # Create a CART classifier
         cart_classifier = DecisionTreeClassifier(criterion='gini',random_state=42)

         # Train the classifier on the training data
         cart_classifier.fit(X_train, y_train)

         # Make predictions on the test data
         y_pred = cart_classifier.predict(X_test)

         # Evaluate the classifier's performance
         accuracy = accuracy_score(y_test, y_pred)
         print(f'Accuracy: {accuracy:.2f}')

         # Generate a classification report
         class_report = classification_report(y_test, y_pred)
         print('Classification Report:\n', class_report)
```

```
Accuracy: 0.89
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.88      0.86        17
           1       0.93      0.89      0.91        28

    accuracy                           0.89        45
   macro avg       0.88      0.89      0.88        45
weighted avg       0.89      0.89      0.89        45
```

Fig. 14. Code snippet of CART model and accuracy.

```
In [69]: from sklearn.neighbors import KNeighborsClassifier

In [70]: # Create a KNN classifier with a specified number of neighbors (k)
         k = 4  # You can adjust the value of k
         knn_classifier = KNeighborsClassifier(n_neighbors=k)

         # Train the classifier on the training data
         knn_classifier.fit(X_train, y_train)

         # Make predictions on the test data
         y_pred = knn_classifier.predict(X_test)

         # Evaluate the classifier's performance
         accuracy = accuracy_score(y_test, y_pred)
         print(f'Accuracy: {accuracy:.2f}')

         # Generate a classification report
         class_report = classification_report(y_test, y_pred)
         print('Classification Report:\n', class_report)
```

```
Accuracy: 0.88
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.91      0.85        35
           1       0.94      0.85      0.90        55

    accuracy                           0.88        90
   macro avg       0.87      0.88      0.87        90
weighted avg       0.89      0.88      0.88        90
```

Fig. 15. Code snippet of KNN model and accuracy.

algorithm in prior cancer prediction studies.

5. Conclusion

Following a thorough assessment of multiple machine learning algorithms, logistic regression and SVM emerge as the most appropriate options for our analysis. These two algorithms consistently showcase the highest level of accuracy when it comes to predicting patient outcomes. Their consistent and strong performance, particularly when handling large datasets, positions them as ideal for ensuring scalability and dependability.

Our analysis supports the idea that SVM is indeed a significant algorithm in this context. In other words, our research validated the insights derived from the literature review. During the pandemic, several significant trends were observed in healthcare compared to the pre-pandemic period as follows.

1. There was an approximate 30% decline in the number of reported cases during the pandemic, indicating the potential influence of COVID-19 on healthcare-seeking behavior.

```python
# Define the ANN model
model = keras.Sequential([
    layers.Input(shape=(X_train.shape[1],)),  # Input layer with the number of features
    layers.Dense(64, activation='relu'),      # Hidden layer with 64 neurons and ReLU activation
    layers.Dense(1, activation='sigmoid')     # Output layer with 1 neuron and sigmoid activation for binary classification
])

# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=10, batch_size=32, validation_split=0.2)

# Evaluate the model on the test data
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Loss: {loss:.4f}, Test Accuracy: {accuracy:.4f}")

# Make predictions
predictions = model.predict(X_test)
```

```
Epoch 1/10
11/11 [==============================] - 1s 41ms/step - loss: 1.0728 - accuracy: 0.5648 - val_loss: 0.6775 - val_accuracy: 0.67
90
Epoch 2/10
11/11 [==============================] - 0s 11ms/step - loss: 0.6375 - accuracy: 0.7160 - val_loss: 0.4765 - val_accuracy: 0.79
01
Epoch 3/10
11/11 [==============================] - 0s 10ms/step - loss: 0.5367 - accuracy: 0.7778 - val_loss: 0.3795 - val_accuracy: 0.82
72
```

```python
# Evaluate the model on the test data
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Loss: {loss:.4f}, Test Accuracy: {accuracy:.4f}")
```

```
2/2 [==============================] - 0s 7ms/step - loss: 0.2786 - accuracy: 0.8444
Test Loss: 0.2786, Test Accuracy: 0.8444
```

Fig. 16. Code snippet of ANN model and accuracy.

2. During the pandemic, cases detected through incidental and surveillance methods decreased substantially by around 50%, which suggests a disruption in regular healthcare services and a shift in diagnostic approaches.
3. During the pandemic, there was an observed increase in the detection of stage C BCLC (Barcelona Clinic Liver Cancer), indicating potential delays in cancer diagnosis and treatment.

One concerning trend was the average tumor size during the pandemic, which was approximately 50 mm, significantly larger compared to the pre-pandemic period. This points to potential delays in cancer detection and diagnosis, resulting in more advanced stages at presentation. Interestingly, the age distribution remained largely unaffected by COVID-19 during and before the pandemic, indicating that age groups were similarly impacted. However, a noteworthy gender-based difference emerged, with a more substantial impact on males than females.

5.1. Potential research contribution

According to studies in Section 2, a variety of machine learning methods are utilized in the healthcare sector to boost productivity. This study has discovered that it can predict a patient's survival rates. This assists in resource allocation and workload reduction of healthcare staff. Due to the restricted utilization of hospitals, this study was able to determine the effect the pandemic had on patients with liver cancer. The analysis will enable the healthcare industry to better prepare for such circumstances in the future. In other words, our research finding was able to improve the accuracy of the current analysis and also increase the efficiency of the healthcare industry.

5.2. Limitations

- Our research findings and analyses are constrained to the population of patients in the United Kingdom, specifically in the Newcastle upon Tyne region.
- It's important to acknowledge that these results may not be universally applicable due to regional variations.
- For example, the predictive models that work well for patients in Newcastle upon Tyne may not easily apply to patients in other regions or countries with unique healthcare systems and patient populations. Substantial variations in cancer survival rates, treatment approaches, and access to healthcare services across different geographic areas can also occur.

5.3. Recommendations

While we cannot provide a single set of definitive recommendations for pandemic situations, the insights mentioned above can serve as valuable lessons for addressing similar circumstances in the future.

1. **Healthcare Infrastructure Enhancement:** Consider investments in healthcare infrastructure, such as expanding hospital capacity, procuring critical medical resources, and ensuring an adequate supply of ventilators and oxygen.
2. **Telehealth Promotion:** Encourage the growth and utilization of telehealth services to reduce in-person medical visits, minimize exposure risks, and enhance the accessibility of medical consultations.
3. **Preparedness for Emergencies:** Establish and regularly update pandemic readiness strategies at local, national, and global levels to ensure effective responses to future health crises.
4. **Future Healthcare Planning:** Formulate comprehensive, long-term plans for healthcare delivery that incorporate valuable insights gained from experiences during the pandemic.

5. **Encourage Research Collaboration:** Foster partnerships and cooperation among researchers, institutions, and entities to exchange data, methodologies, and knowledge. This collaborative effort can drive progress in the field of cancer survival prediction.
6. **Utilize Cutting-Edge Machine Learning:** Consider using advanced machine learning techniques such as deep learning neural networks, which excel at identifying hidden data linkages and can potentially improve prediction accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgement

This research is partly supported by VC Research (VCR 0000223) for Prof Chang.

Appendix

Data Understanding.
Following is the data description of all the attributes of the dataset.

Sr. No	Attribute	Description	Value
1.	Cancer	Cancer flag	Yes/No
2.	Year	Patient referred during pre-pandemic (March 2019–February 2020) or Post-pandemic (March 2020–February 2021)]	Prepandemic/Postpandemic
3.	Month	Month of the year	1–12
4.	Bleed	Spontaneous tumor hemorrhage	y/N
5.	Mode Presentation	Presentation mode	Surveillance, Incidental, or Symptomatic
6.	Age	Age of the patient	eg: 45,65
7.	Gender	The gender of the patient	M/F
8.	Etiology	Manner of causation of a disease or condition	"No established CLD" (chronic liver disease)/"ARLD" (alcohol-related liver disease)/ "NAFLD" (non-alcoholic fatty liver disease)/"HCV" (hepatitis C virus)/"HH" (hereditary haemochromatosis)/"PBC/AIH" (primary biliary cholangitis/ autoimmune hepatitis)/"HBV" (hepatitis B virus)/"Other"
9.	Cirrhosis	Whether a patient has underlying liver disease	y/n
10.	Size	Tumor diameter in mm	eg: 0.53, 1.33
11.	Hcc Tnm Stage	Hepatocellular carcinoma tumor node metastasis stage	"I", "II", "IIIA + IIIB", "IV"
12.	Hcc Bclc Stage	Hepatocellular carcinoma barcelona clinic for liver cancer stage	"0″, "A", "B", "C", "D"
13.	Icc Tnm Stage	Intrahepatic cholangiocarcinoma tumor node metastasis stage	"I", "II", "III", "IV"
14.	Treatment Groups	First-line treatment received	"OLTx" (orthotopic liver transplantation), "Resection", "Ablation", "TACE"" (transarterial chemoembolization), "SIRT" (selective internal radiation therapy), "Medical", "Supportive care"
15.	Survival From Mdm	Survival from multidisciplinary meeting	eg: 2, 1.98
16.	Alive Dead	Whether the patient survived or dead	Alive/Dead [Target]
17.	Type Of Incidental Finding		"Primary care-routine", "Secondary care-routine", "Primary care-acute", "Secondary care-acute"
18.	Surveillance Programme	Patient in a formal surveillance program	y/n
19.	Surveillance Effectiveness	Surveillance adherence over the previous year	"Consistent", "Inconsistent", "Missed"
20.	Mode Of Surveillance Detection	Mode of incident surveillance test	"US" (ultrasound), "AFP alone" (alpha-fetoprotein alone), "CT/MRI"
21.	Time Diagnosis 1st Tx	The time interval between and the initiation of the first treatment	eg: 5, 3
22.	Date Incident Surveillance Scan	Whether a Surveillance scan was performed	y/n
23.	Ps	Performance status	0,1, 2, 3, 4
24.	Time Mdm 1st Treatment	Time to multidisciplinary meeting 1st treatment	3.
25.	Time Decision To Treat 1st treatment	Time decision to treat 1st treatment	6 ….
26.	Prev Known Cirrhosis	Suffering from Cirrhosis	Y/N
27.	Months From the last Surveillance	Months from the last surveillance	1,2, …

The target variable chosen for our predictive modeling is "Alive_Dead" which will help determine whether a patient will survive or not based on the other attributes. The raw data comprises 450 rows and 27 columns. Raw data is data that is not processed or analyzed in any way. This data can be inconsistent and hence needs cleaning.

References

[1] H. Rumgay, M. Arnold, J. Ferlay, O. Lesi, C.J. Cabasag, J. Vignat, M. Laversanne, K. A. McGlynn, I. Soerjomataram, Global burden of primary liver cancer in 2020 and predictions to 2040, J. Hepatol. 77 (6) (2022), https://doi.org/10.1016/j.jhep.2022.08.021.

[2] R.A. Zimerman, D. do Nascimento Fonseca, M. do Nascimento Correia, R.N. Barros, D.C. Onety, K.C.P. Israel, E.O. Guerreiro, J.E.M. Medeiros, R.N. Nicolau, L.F. M. Nicolau, R.X. Cunha, Proxalutamide (GT0918) reduction of mortality rate in hospitalized COVID-19 patients depends on treatment duration–an exploratory analysis of the proxa-rescue androCoV trial, medRxiv (2021) 2021, 06.

[3] G.M. Miller, J.A. Ellis, R. Sarangarajan, A. Parikh, L.O. Rodrigues, C. Bruce, N. M. Chand, S.R. Smith, K. Richardson, R. Vazquez, M.A. Kiebish, Ondansetron use is associated with lower COVID-19 mortality in a Real-World Data network-based analysis, medRxiv (2021) 2021, 10.

[4] A. Kiani, B. Uyumazturk, P. Rajpurkar, A. Wang, R. Gao, E. Jones, J. Shen, Impact of a deep learning assistant on the histopathologic classification of liver cancer, NPJ digital medicine 3 (1) (2020) 23.

[5] J.Y. Nam, D.H. Sinn, J. Bae, E.S. Jang, J.W. Kim, S.H. Jeong, Deep learning model for prediction of hepatocellular carcinoma in patients with HBV-related cirrhosis on antiviral therapy, JHEP Rep 2 (2020) 100175.

[6] S. Rajesh, N.A. Choudhury, S. Moulik, Hepatocellular carcinoma (HCC) liver cancer prediction using machine learning algorithms, in: 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1–5, https://doi.org/10.1109/INDICON49873.2020.9342443. New Delhi, India.

[7] S. Zeng, L. Li, Y. Hu, L. Luo, Y. Fang, Machine learning approaches for the prediction of postoperative complication risk in liver resection patients, BMC Med. Inf. Decis. Making 21 (1) (2021) 1–10.

[8] B. McMahon, C. Cohen, Brown, H. El-Serag, G.N. Ioannou, A.S. Lok, T. Block, Opportunities to address gaps in early detection and improve outcomes of liver cancer, JNCI Cancer Spectrum 7 (3) (2023) kad034.

[9] J.Y. Islam, D.C. Vidot, M. Camacho-Rivera, Evaluating mental health–related symptoms among cancer survivors during the COVID-19 pandemic: an analysis of the COVID impact survey, JCO Oncology Practice (2021), https://doi.org/10.1200/op.20.00752 p.OP.20.00752.

[10] R. Wu, W. Guo, X. Qiu, S. Wang, C. Sui, Q. Lian, J. Wu, Y. Shan, Z. Yang, S. Yang, T. Wu, K. Wang, Y. Zhu, S. Wang, C. Liu, Y. Zhang, B. Zheng, Z. Li, Y. Zhang, S. Shen, Comprehensive analysis of spatial architecture in primary liver cancer, Sci. Adv. 7 (51) (2021), https://doi.org/10.1126/sciadv.abg3750.

[11] T.C. Rose, A. Pennington, C. Kypridemos, T. Chen, M. Subhani, J. Hanefeld, L. Ricciardiello, B. Barr, Analysis of the burden and economic impact of digestive diseases and investigation of research gaps and priorities in the field of digestive health in the European Region-White Book 2: executive summary, United European gastroenterology journal (7) (2022) 657–662, https://doi.org/10.1002/ueg2.12298 [online] 10.

[12] N. Darapaneni, M. Arul, A.R. Paduri, V. Meenakshisundaram, P. Srinivasan, N. Joseph, D. Sharma, Comorbidity impact on covid-19, in: 2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), IEEE, 2020, December, pp. 1–6.

[13] B. Yang, J. Ma, Retrospective Analysis of Prognosis of Tumor Patients Co-infected with COVID-19, 2023.

[14] E. Hernandez-Garcia, E. Chrysikou, A.Z. Kalea, The interplay between housing environmental attributes and design exposures and psychoneuroimmunology profile—an exploratory review and analysis paper in the cancer survivors' mental health morbidity context, Int. J. Environ. Res. Publ. Health 18 (20) (2021) 10891, https://doi.org/10.3390/ijerph182010891.

[15] A. Agaal, M. Essgaer, An exploratory data analysis of breast cancer features in south of Libya, J. Pure Appl. Sci. (Ankara) 21 (4) (2022) 57–64.

[16] N. Leenaerts, T. Vaessen, J. Ceccarini, E. Vrieze, How COVID-19 lockdown measures could impact patients with bulimia nervosa: exploratory results from an ongoing experience sampling method study, Eat. Behav. 41 (2021) 101505, https://doi.org/10.1016/j.eatbeh.2021.101505.

[17] O. Ramirez, A Survival Analysis Project of Two Biomedical Diseases: Identifying Factors Associated with Mortality of Patients with Primary Biliary Cholangitis (Pbc) and Coronavirus Disease 2019, 2022 covid-2019).

[18] A.A. Schnitzbauer, N. Filmann, R. Adam, Philippe Bachellier, W.O. Bechstein, T. Becker, S. Bhoori, I. Bilbao, J. Brockmann, O. Patrizia Burra, Chazoullières, U. Cillo, M. Colledan, C. Duvoux, T.M. Ganten, J. Gugenheim, M. Heise, Bart van Hoek, N.V. Jamieson, Jong, mTOR inhibition is most beneficial after liver transplantation for hepatocellular carcinoma in patients with active tumors, Ann. Surg. 272 (5) (2020) 855–862, https://doi.org/10.1097/sla.0000000000004280.

[19] S. Rajesh, N.A. Choudhury, S. Moulik, Hepatocellular carcinoma (HCC) liver cancer prediction using machine learning algorithms, in: 2020 IEEE 17th India Council International Conference (INDICON), IEEE, 2020, pp. 1–5.

[20] V. Chang, M.A. Ganatra, K. Hall, L. Golightly, Q.A. Xu, An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators, Healthcare Analytics (2022) 100118, https://doi.org/10.1016/j.health.2022.100118.

[21] Simplilearn, What Is Data Collection: Methods, Types, Tools, and Techniques, Simplilearn, 2021 [online], https://www.simplilearn.com/what-is-data-collection-article.

[22] A.Y. Wang, D. Wang, J. Drozdal, X. Liu, S. Park, S. Oney, C. Brooks, What makes a well-documented notebook? a case study of data scientists' documentation practices in kaggle, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–7.

[23] W. McKinney, Python for Data Analysis, O'Reilly Media, Inc., 2022.

[24] J.C. Jakobsen, C. Gluud, J. Wetterslev, P. Winkel, When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts, BMC Medical Research Methodology 17 (1) (2017) 1–10.

[25] M.A. Mansournia, A. Geroldinger, S. Greenland, G. Heinze, Separation in logistic regression: causes, consequences, and control, American Journal of Epidemiology 187 (4) (2018) 864–870.

[26] A. Saini, Decision Tree Algorithm - A Complete Guide, 2021 [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm. (Accessed 29 September 2023).

[27] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees, Wadsworth Int. Group 37 (15) (1984) 237–251.

[28] D. Bazazeh, R. Shubair, Comparative study of machine learning algorithms for breast cancer detection and diagnosis, in: 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, https://doi.org/10.1109/icedsa.2016.7818560.

[29] S. Sharma, Artificial Neural Network (ANN) in Machine Learning - DataScienceCentral.Com, 2017 [online] Data Science Central. Available at: https://www.datasciencecentral.com/artificial-neural-network-ann-in-machine-learning/.

[30] A. Mishra, Metrics to Evaluate Your Machine Learning Algorithm, 2018 [online] Medium. Available at: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

[31] E. Zuccarelli, Performance Metrics in Machine Learning — Part 1: Classification (2021). Accessed: Jul, 23, 2022. [online] Medium. Available at: https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92.

[32] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, International Journal of Data Mining & Knowledge Management Process 5 (2) (2015) 1.

[33] B. Juba, H.S. Le, in: Precision-recall versus accuracy and the role of large data sets, 33, Proceedings of the AAAI conference on artificial intelligence, 2019, pp. 4039–4048.

[34] J. Brownlee, How to choose a feature selection method for machine learning, Machine Learning Mastery (2019), 10. Available at: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/.

[35] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Computing Surveys (CSUR) 50 (6) (2017) 1–45.

[36] Tableau, Data Visualisation Beginner's Guide: a Definition, Examples and Learning Resources, Tableau Software, 2020. . Available at: https://www.tableau.com/en-gb/learn/articles/data-visualization.

[37] A. Unwin, Why is data visualization important? what is important in data visualization? Harvard Data Science Review 2 (1) (2020) 1.

[38] J. Cha, J. Seong, Application of radiotherapeutic strategies in the BCLC-defined stages of hepatocellular carcinoma, Liver Cancer 1 (3–4) (2012) 216–225.