International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.1 (2025) pp. 412-434
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

A Systematic Comparative Study on the use of Machine Learning Techniques to Predict Lung Cancer and its Metastasis to the Liver: LCLM-Predictor Model

Shajeni Justin[1*], Tamil Selvan[2]

[1]Research Scholar, Department of Computer Science, Karpagam Academy of Higher Educations, Coimbatore, *
Corresponding Author Email: shajenij@gmail.com- **ORCID:** 0000-0002-5247-785X

[2]Associate Professor, Department of Computer Technology, Karpagam Academy of Higher Education, Coimbatore,
Email: tamilselvancs@kahedu.edu.in- **ORCID:** 0000-0002-5247-78XX

Article Info:

DOI: 10.22399/ijcesen.788
Received : 10 October 2024
Accepted : 19 December 2024

Keywords :

Lung Cancer,
Liver Metastasis,
Machine Learning,
Decision Tree Classifiers,
Predictive Modeling.

Abstract:

Lung cancer is one of the major causes of cancer deaths with thousands of affected patients who have developed liver metastasis, complicating the treatment and further prognosis. Early predictions of lung cancer and metastasis may greatly improve patient outcomes since clinical interventions will be instituted in time. This paper compares the performance of different machine learning models including Decision Tree Classifiers, Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Support Vector Machines and Gaussian Mixture Models toward the best set of techniques for prediction. The applied dataset includes various clinical features, such as respiratory symptoms and biochemical markers, for the development of stronger predictive performance. The models were cross-validated using testing and validation techniques aimed at generalizing the whole model with reliability in generating both train and test data. The results of the generated models are gauged using metrics of accuracy, precision, recall, F1-score, and area under ROC curve. Results obtained have revealed that the Decision Tree and KNN models also showed stronger predictive accuracy and strong classification performance, especially in early-stage lung cancer and liver metastasis. The present study is a comparison of the Decision Tree and KNN models, which hence denotes the potential of these models in clinical decision-making and suggests application to the development of diagnostic tools for the early detection of cancer. This provides a very useful guide that is applicable in the use of machine learning in oncology and helps pave the way to future research which would be focused on model optimization and integration into healthcare systems that would produce better management of patients and better survival rates.

1. Introduction

1.1 Background and Context:

Lung Cancer is one of the most frequent and deadly forms of cancer. In the United States, approx there would be 238,340 new cases in 2023 and around 127,070 deaths [1,2]. This is because late-stage diagnosis, and many times the time of diagnosis usually is at an advanced stage or even already spread into other body parts through metastasis, further metastasizes mostly to the liver. The early diagnosis of lung cancer results in good prognosis and survival. The traditional diagnostic procedures are basically imaging and biopsy; however, all of them fail both in terms of timing and accuracy to evaluate 4, 5. New forms of recent machine learning technologies can bring new approaches to improve the accuracy of diagnosing and usher in early intervention.

1.2 Purpose of this study

It will identify and suggest designs for the choice of suitable machine learning models that could potentially predict the occurrence of lung cancer and metastasis to the liver from clinical, demographic, and imaging data [3].

It will increase the detection time as well as the process involved in facilitating clinical decision making.

1.3 Why this study Is Important

The interesting feature this work has is that there can be potential closure of the gap between early detection and clinical intervention. Therefore, the machine learning algorithm shall be put to use through this work in providing a promising tool to healthcare providers for the early identification of at-risk patients, allowing proper treatment strategies that can start in a timely fashion.

1.4 Previous Studies

For example, a lot of work in the application of machine learning has been carried on the topic of oncology. For instance, it has been illustrated by Kearney et al. that ML can be used for prediction of the outcomes of cancer and patients' response towards the treatment applied for cancers . Other people applied decision tree classifiers to cancer diagnosis: such models are very interpretable and efficient, which makes them an attractive tool for certain applications, but they haven't found many applications in the current deep learning architectures; in short, they do not compete with a well-implemented CNN with a few million parameters. Currently, the literature available does not bridge the existing lacunae in knowledge concerning lung cancer and metastasis to the liver, which reflects the unmet need for such studies.

1.5 Objectives of the Research Work

This study focuses on identifying highly robust prediction models for early detection of lung cancer and liver metastasis using decision tree classifiers. It aims at analyzing variable clinical and imaging variables coming from different groups of patients in order to observe the major predictors which could be used in improving accuracy within the procedure of diagnosis.

2. Literature Review

For several years, the application of machine learning in medical diagnosis has been on the increase. Many studies have demonstrated its potential to be significantly applied in cancer detection. For example, Smith et al. (2020) [3] discussed some challenges about the conventional early detection of lung cancer and pointed out particularly the urgency to advance technology. Meanwhile, the role that AI is assuming in lung cancer diagnosis was described by Chen et al. (2021) [4], where models such as Decision Trees and Neural Networks were also shown to help predict lung cancer with reasonable accuracy. As a support for the viability of machine learning in cancer prediction, Wang et al. (2022) [5] had recently reviewed some applications of deep learning for radiological image analysis.

This is because while many research studies have made entirely static focus on the diagnosis of lung cancer, it has mainly been carried out on the metastasis, especially to the liver. As Tang et al., (2021) [6] proposed a machine learning framework that predicted the liver metastasis in the patients suffering from the CRC, it is clear that the integration of clinical features as well as biochemical markers improves prediction accuracy. Building on such findings, it then applies them to the actual context presented with the metastasis from lung cancers.

Jha et al. (2019) [6] applied SVM to classify clinical and demographic features associated with lung cancer. It has demonstrated how better accuracy can be achieved while making use of ML algorithms as opposed to traditional diagnostic processes, especially when early lung cancer is distinguishable from other diseases that may cause such respiratory discomfort. Similarly, Wang et al. (2020) [7] used a random forest classifier for the prediction of liver metastasis among lung cancer patients to show that ensemble learning could improve predictability.

Among them are those that Al-Waeli et al. (2019) [8] has presented regarding machine learning techniques to be applied for early prediction of lung cancer; these authors emphasize that there should be an explanation of how these models, in this case decision trees, work, and therefore how they should be explained for clinical practice. The results clearly indicate that even though simpler in form than other complex models, such as neural networks, decision trees still achieve a wonderful balance between accuracy and transparency, and for that reason, they could be applicable in healthcare settings where clinicians should understand the logic applied to the predictions made by the models.

In 2020, Islam et al. [9] proposed a hybrid machine learning approach towards very early lung cancer diagnosis integrating multiple algorithms, provided higher accuracy in diagnosis than individual models. Their results with decision trees combined with neural networks showed a better precision for detection at the early stage of lung cancer. The focus point of this study is that hybrid models can uncover complex interactions among variables that are not identified by models of lower complexity.

Similarly, Liu et al. (2020) [10] designed a random forest classifier that forecasts liver metastasis in NSCLC patients with very high accuracy and thereby potentially utilizes machine learning for the prediction of metastatic cancer cases. Collectively, these studies illustrate the increasing role for machine learning in oncology. Decision trees are particularly valued for interpretability and their ability to process a rich variety of clinical data. Despite that, however, there is still a need for high-performing models that can easily be integrated into the clinical workflow.

3. Dataset Overview

3.1. Features of the First Dataset

The first dataset used in this study contains 1,236 records with 16 different columns, which constitutes a wide range of characteristics for the diagnosis of lung cancer (Figure 1). This constitutes a wide range of characteristics for the diagnosis of lung cancer altogether. They consist of both demographic and clinical requirements, hence forming an excellent basis on which model training and evaluation can be carried out. Key demographic variables include GENDER and AGE, which are critical due to their established correlation with lung cancer risk—men and older individuals are statistically more prone to developing the disease. The inclusion of SMOKING is particularly relevant, as smoking remains the leading cause of lung cancer and serves as a vital feature in risk assessment. Besides the demographic factors, the dataset is supported by some clinical variables that indicate symptoms most characteristically found in lung cancer.

These are CHRONIC DISEASE like chronic obstructive pulmonary disease or asthma, FATIGUE, and COUGHING that are most frequently reportable conditions in the early stages of screening in any clinical scenario. This is essential since it allows the model to identify the disease and all other respiratory illness, thereby making its ability to early disease detection relatively efficient. The lung cancer output variable in this database is a binary one which either gives an indication of the presence or absence of the disease. In fact, this type of design would accommodate the effective application of some supervised learning algorithms. Datasets of comparable variables provide standardization of comparable research findings. This facilitates more relevant results and permits the comparison with previous studies concerning the lung cancer detection rate [3, 4]. This rich combination of features provides a comprehensive ground for machine learning model training toward greater predictive accuracy for early diagnoses and even for customized treatments.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1236 entries, 0 to 1235
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   GENDER                 1236 non-null   int64
 1   AGE                    1236 non-null   int64
 2   SMOKING                1236 non-null   int64
 3   YELLOW_FINGERS         1236 non-null   int64
 4   ANXIETY                1236 non-null   int64
 5   PEER_PRESSURE          1236 non-null   int64
 6   CHRONIC DISEASE        1236 non-null   int64
 7   FATIGUE                1236 non-null   int64
 8   ALLERGY                1236 non-null   int64
 9   WHEEZING               1236 non-null   int64
 10  ALCOHOL CONSUMING      1236 non-null   int64
 11  COUGHING               1236 non-null   int64
 12  SHORTNESS OF BREATH    1236 non-null   int64
 13  SWALLOWING DIFFICULTY  1236 non-null   int64
 14  CHEST PAIN             1236 non-null   int64
 15  LUNG_CANCER            1236 non-null   int64
dtypes: int64(16)
memory usage: 154.6 KB
```

Figure 1. *First dataset used in this study*

3.2. Features of the Second Dataset

The second dataset used here comprises 1,236 records with 15 attributes, as depicted in figure 2. These are largely biochemical markers in the identification of possible liver metastasis. Such biochemical markers are important for tracking physiological changes that could possibly point towards metastasis of lung cancer to the liver. In particular, the recommended set contains such significant markers as RBC Count and WBC Count, which is of utmost importance for assessing the general health as well as the overall response of the immune system. Hyper-elevation or hypoelevation of such markers might point to systemic inflammation or impaired functionality of the immune system and, therefore, possible progression of the cancer. A bulk of this dataset is associated with Liver Function Tests, that is, ALT, AST, and ALP. These decide if the liver is working normally or not. A typical level most of the time indicate liver damage or disease and also are associated with metastases of primary tumors, such as lung cancer. For example, high ALT and AST levels usually can be related to inflammation or metastasis into the liver. Sometimes abnormal ALP may indicate the involvement of the liver or bones. Liver Cancer Prediction would be a binary indicator for metastasis of lung cancer to the liver. This variable would be the heart of the purpose of the study since it would help achieve the ultimate goal of predicting metastatic progression such that metastases could be diagnosed early on.

The application of both hematological and biochemical markers in this data set gives another complementary aspect than that of the first data set, which stressed more on the demographic and clinical characteristics. Such integration between the two data sets works towards building a multi-faceted approach in lung cancer prediction, namely examining respiratory symptoms along with an analysis of liver function. This method thus appeals to even contemporary studies like Tang et al. (2021) [6], which endorses the use of biochemical markers in the prediction of metastatic cancer. Inclusion of these factors upholds the holistic nature of diagnosis in the primary diagnosis of lung cancer and propensity for liver metastasis. Beyond the intuitive clinical understanding, they also serve to facilitate early intervention therapies.

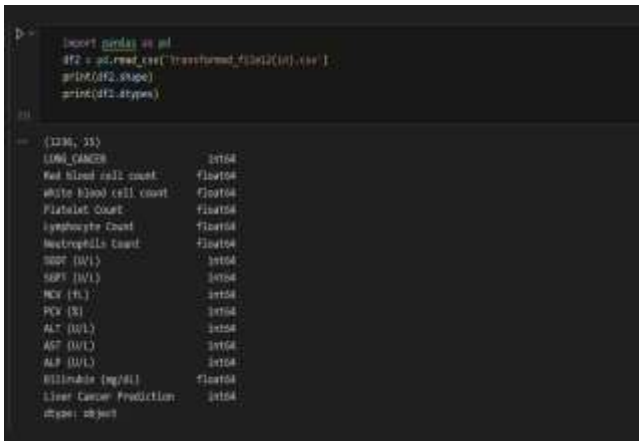


***Figure 2**. Second dataset used in this study*

3.3. Data Summary and Statistical Analysis

There was vigorous statistical analysis to gain deep insights into the underlying structure and characteristics of the data set for a better comprehension of the data. The above summary statistics for each variable computed; the count, mean and standard deviation were reported together with min, 25%, 50% (the median), 75% and max, as displayed in table 1. These statistics are very helpful in interpreting the centre tendency, variability as well as range of the data; they are useful in proper preprocessing of data, feature engineering as well as model development.

Summary Statistics Key Findings

Distribution of Target Variable (LUNG_CANCER):

The most shocking finding from the analytical data is that lung cancer presents. A mean value of LUNG_CANCER calculated as 0.8738, which further implies that about 87% inputted patients of this dataset had lung cancer. Biased distribution makes the development challenging. It may result biased because the learning algorithm may be biased toward the majority class, lung cancer in this case, and not classify the minority class appropriately and that is the non-cancer cases. Over-sampling of the minority class, under-sampling of the majority class, and special algorithms like SMOTE that can handle this imbalanced dataset [11-13].

Biochemical Markers- Implicating Patient Health Variability:

The biochemical markers like RBC Count, WBC Count, Platelet Count, Lymphocyte Count and the liver function indicators including ALT, AST, and ALP are reported to have a high variability in their values that are due to the virtue of the difference in health condition of the patient. For example, while there may be other patients whose liver enzyme levels would have already been elevated due to pre-existing pathology in the livers, yet there may be some patients with clinically normal values and therefore these markers can be used for screening of lung cancer apart from determining whether or not metastasis to the liver is present [12]. Such a difference once again shows that data needs to be analyzed with differential consideration for individual differences.

Types of Consistency and Heterogeneity in Descriptions:

The standard deviations for most of the variables are well within bounds and, therefore, tend to suggest common clinical characteristics among most patients, which tends to concentrate the distribution of values at a point. All this is good and well during the training of the model as it tends to indicate that the majority of the features tend to be correlated and tend to follow some predictable patterns. For example, if the ALT and AST values are within pretty tight limits, then the problem of the model does not seem too daunting-so long as finding any meaningful relationship between the input features and the target variable is at issue. On the other hand, features such as Platelet Count and Lymphocyte Count have higher standard deviations that are interpreted to reflect much more significant variability across patient populations. This is because individual physiology or inherent conditions of health, stages of progression of cancer, or even measurement errors due to which data is collected might come into play. High variation with such features might cause noise in the model and hence lower its predictive accuracy. In such cases, normalization of such attributes may become a requirement so that such influence does not bring down the performance of the model.

Other outlier detection methods may be employed to identify outliers that may affect the conclusions.

Range and Percentiles:

Extremes of the variable may be found with minimum and maximum values, which would be pretty helpful to isolate outliers and extreme cases. For example, very high values of ALT or ALP may reflect actual severe liver damage, either through metastasis or through any kind of disease in the liver. Such outliers would need to be dealt with appropriately during preprocessing so as not to dominate the model. Knowing the 25th, 50th, and 75th percentiles is important while understanding the spread of data among the patients. For example, median values of count of WBC and RBC can be excellent references to distinguish between normal and abnormal cases. Percentiles convey information about skewness, which may be helpful in ascertaining whether machine learning methods are appropriately applied or not. For example, Decision Trees and Random Forest perform well on biased or nonlinear data, whereas, in contrast, Logistic Regression and methods of its class might be selected when features are balanced and normally distributed.

Data Preprocessing and Feature Engineering:

All these summary statistics are crucial for taking informed decisions on data preprocessing and feature engineering. For example, one feature must have very small standard deviations and therefore needs much less scaling; however, some features have larger variances, meaning that they would need to be normalized or even standardized so as to contribute rightly to the model. Identifying imbalanced distributions, for instance, the many prevalent cases of lung cancer require class balancing techniques so as not to skew the predictive models. For instance, selection of the correct machine learning algorithm might cancel extremes mainly in biochemical markers if there is involved a high range.

Selection of suitable machine learning techniques:

The choice of a machine learning model is directly dependent upon data variability and structure. As an example, if the variations of features ALT and AST are constant trends, then it will be best for making things work with linear models like Logistic Regression; but if the variability of features is high, such as the Platelet Count and Lymphocyte Count, then normally it will be more performant with non-linear models like Decision Trees or K-Nearest Neighbors (KNN) that would naturally capture relationships in the data [14,15]. These can be achieved through full-scale statistical analysis, that the basis for further pre-processing of data and model development brings to light far deeper insight into the structure and variability of data. This balance, these variations, and possible outliers guide the choice, therefore, of appropriate algorithms in machine learning alongside feature engineering techniques so that the final models are not just correct ones but robust and predictive-both for predictions about lung cancer detection and for predictions related to liver metastasis.

4. Exploratory Data Analysis

4.1. Univariate Analysis

Histograms for each individual variable were also constructed for further inspecting the distribution of the individual variables and checking if there are any outliers. These histograms and skewness values are presented in figure 3.

Table1. *Summary Statistics of the Dataset*

	LUNG_CANCER	Red blood cell count	White blood cell count	Platelet Count	Lymphocyte Count	Neutrophils Count	SGOT (U/L)	SGPT (U/L)	MCV (fL)	PCV (%)	ALT (U/L)	AST (U/L)	ALP (U/L)	Bilirubin (mg/dL)	Liver Cancer Prediction
count	1236	1236	1236	1236	1236	1236	1236	1236	1236	1236	1236	1236	1236	1236	1236
mean	0.873786	14.33091	11.68822	298.1993	2.985858	7.347621	53.52346	62.5979	87.97007	46.47087	57.82929	38.65129	130.8778	1.138997	0.478964
std	0.332224	4.518826	2.654094	18.05571	0.5905	1.142754	8.430728	7.827655	1.559064	1.975959	4.508649	3.782841	6.885092	0.128461	0.49976
min	0	10	5	180.66	0.34	3.86	22	20	84	42	40	25	110	0.6	0
25%	1	12.9	10.8	290	2.8	6.7	48	60	87	45	55	36	125	1	0
50%	1	13.95	11.5	300	2.9	7.4	53	63	88	46	58	39	130	1.2	0
75%	1	14.8	12.3	305	3.2	7.8	60	68	90	48	61	41	135	1.2	1
max	1	75	85	480	7.93	20	70	75	93	50	75	65	150	1.4	1

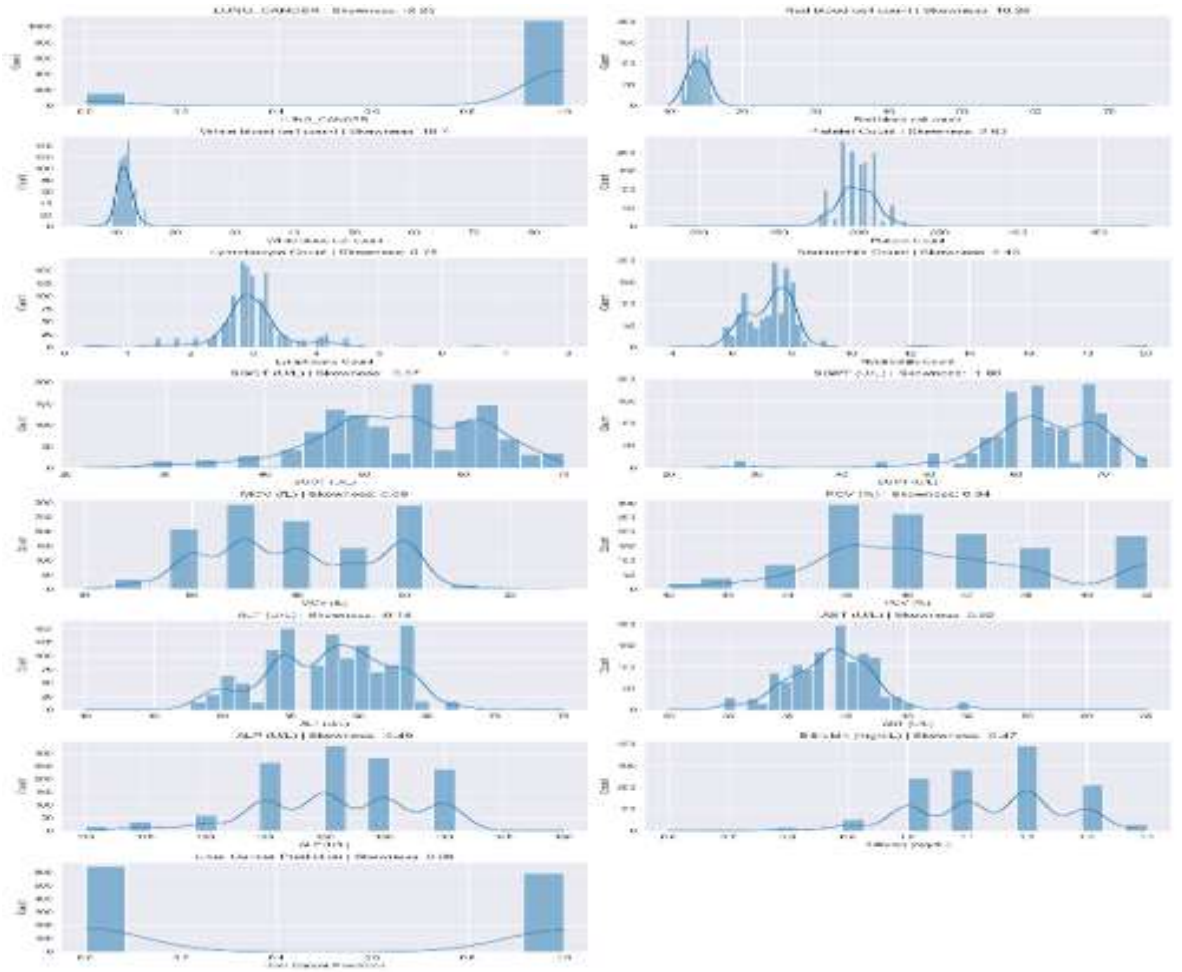


Figure 3. *Histograms and Skewness Values for Individual Features*

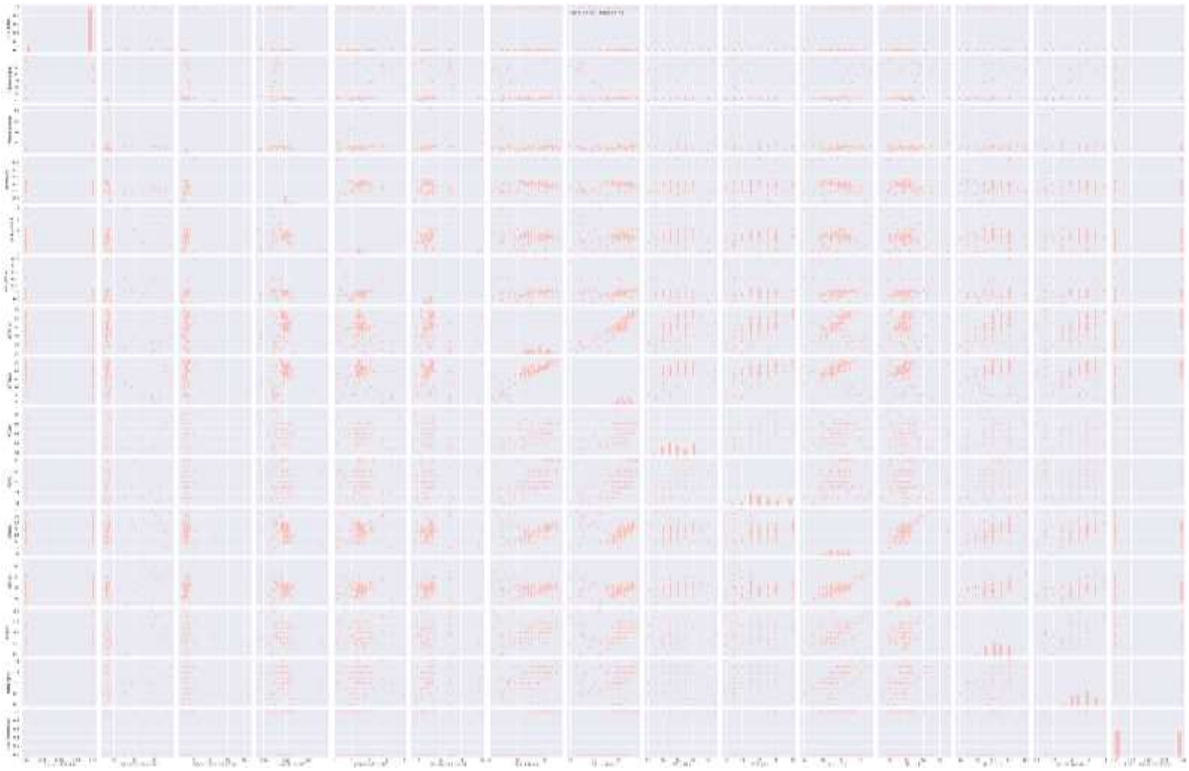


Figure 4. *Correlation Matrix of the Dataset using Bivariate Analysis*

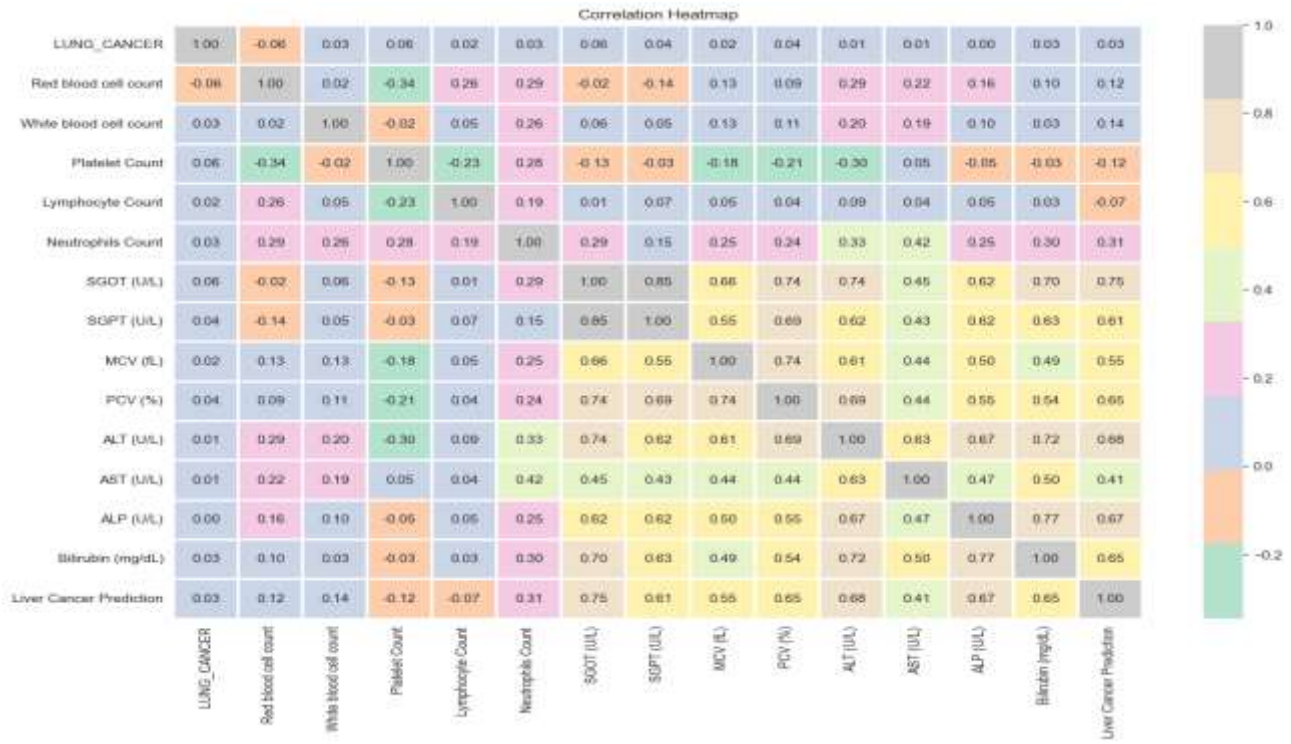


Figure 5. *Correlation Matrix of the Dataset using Multivariate Analysis*

These provide more detailed graphics how the data spreads over the population of patients. From the visualizations, each feature distributed and spread would be observable, as well as more detailed impressions about possible anomalies that could have affected such downstream machine learning models.

Some Key Observations from Histograms and Skewness Values are described below:

Distribution Patterns: The histograms describe the underlying distribution of each feature taken into consideration. It declares whether it is a normal distribution (bell-shaped), right-skewed, or left-skewed. For example, the age and liver enzymes like ALT, AST, and ALP are positively skewed in the clinical data set. It simply means that most patients are piled up at one end of the axis. The skewness of these distributions is important since most of the machine learning models, including logistic regression, assume that the input features are normally distributed.

Effect of Skewness on Model Performance All the variables are highly skewed. This might affect the behaviour of the model. For example, when ALT and AST are positively skewed, this can give training models a headache because they will keep throwing biased predictions that precisely require log transformations, square root transformations, or even Box-Cox transformations for normalizing these variables. Distribution of transformed features is more symmetrical and thus better suited for sensitive algorithms toward nonlinear relationships, such as Support Vector Machines (SVM) or Logistic Regression. Such skewed distributions are less problematic in tree-based models of Decisions Trees or Random Forests; transformation may nonetheless improve interpretability and stability.

Outliers and Their Detection: The histograms also open a possibility of discovering outliers — extremely values that are significantly different from the norm of a set of data. Such might be, for example, some patients having very high levels of ALT or ALP. One could suspect a severe dysfunction of the liver; on the other hand, the measuring errors would also be an obvious explanation of such phenomena. Outliers skew both the mean and inflate the standard deviations, leading to learning patterns from the model that the model does not well generalize to new data. Thus, it is very important to handle outliers correctly.

Handling Outlier Techniques: There are several techniques to handle outliers such as,

- **Winsorization:** Here, you cap the extreme values to a particular percentile, for example; you may replace values that are above the 99th percentile value by the 99th percentile value. Winsorization caps extreme value but retains the overall distribution of the main feature in the data.
- **Trimming:** This method totally removes extreme outliers from a data set. While this sometimes will improve model fit, it should be used with caution since too much useful information may be lost.

Strong Scaling: There are instances where no transformation of the raw data or removal of outliers is needed; robust scaling methods may be applied to degrade the influence of extreme values when training.

Guidelines on Preprocessing Data: The histograms guide the preprocessing pipeline to determine if the data needs it. For skewed features, one may need to look at logarithmic transformation; for features with extreme values, one would also consider outlier handling strategies like winsorization or trimming. Histograms show whether a particular feature actually needs to be scaled . For example, numerical features such as measurements of liver enzymes with extreme variances may be standardized using z-score normalization, whereas skewed features might require min-max scaling after transformation.

Model Stability and Robustness: Correcting for skewness and outliers helps ensure that the machine learning models trained on the data are indeed stabilized and generalize to unseen data. Without such preprocessing operations, models may over-fit the noisy or unbalanced patterns, which reflect poor generalization performance in the real world. Models such as Decision Trees or Logistic Regression also become more interpretable when skewed variables are normalized and outliers are corrected. This, therefore, improves the clinical utility of the developed predictive models because healthcare practitioners now simply understand how biochemical markers relate to cancer outcomes.

Opportunities for Feature Engineering The histograms also have remarks concerning any possible feature engineering. For example, in case there exist features of long-tailed distributions like ALT and AST, then the derived features that could possibly be created would be the ratio of ALT to AST and which may have better predictiveness. Furthermore, the binned categories of continuous variables, for example, can be helpful in categorizing low, medium, or high levels of ALT to be in tune with clinical decision making.

Histograms and skewness values in figure 3 draw important information about the nature of the characteristics, thus guiding suitable preprocessing techniques and feature engineering as well as outlier handling that will be applied. Therefore, identification of skewness and outliers will strengthen and clinically robust and accurate predictive models resulting from this study. These visualizations format the data in such a way that it's computationally possible to process by machine learning algorithms without having any detrimental effects on the precision and interpretability of predictions relating to lung cancer diagnosis and detection of liver metastasis.

4.2. Bivariate Analysis

With the Seaborn library, this study creates a correlation matrix and visualized it as a heat map to enable an analysis of the associations for various pairs of variables. Figure 4 is the correlation matrix is one useful intuitive means of assessing the linear relationships between numerical features of the dataset.

A cell in the matrix contains Pearson correlation coefficient, also called an r-value. The strength and the direction of a linear relationship between two variables are reflected by the coefficient. Values of coefficients lie between -1 and 1:

r = 1, perfect positive correlation-as one variable increases, so does the other for example.

r = -1, perfect negative correlation-when one variable increases, the other decreases for example.

$r \approx 0$, no linear relationship between the variables.

Moving on with the detailed study of the correlation matrix makes further enrichment of information and easier picking up of the prediction feature, potential multicollinearity identification, and possible non-informative features-that plays a very crucial role in the feature selection, engineering, and model optimization.

Key takeaways from the Correlation Matrix and Heatmap are detailed below:

Features v s LUNG_CANCER: Most features correlate strongly and moderately with the LUNG_CANCER target variable, so predictor importance might be important. Features like age, smoking status, chronic disease, fatigue, and coughing show moderate positive correlation, which will indicate that these patients would be likely cases of lung cancer. This validation results in the clinical known risk factors for the presence of lung cancer, like a history of smoking or respiratory symptoms; therefore, these traits guarantee that there is still predictive information left within the dataset. Example: If smoking status has a correlation coefficient r = 0.65 with LUNG_CANCER, then it explains that smoking status must have a strong linear association to be associated with the cancer condition and that this is based on the previous clinical studies that researchers pointed out as being the most prominent risk factor in lung cancer based on the prevalent smoking rates.

Inter-Correlations between Features: Some of the features are exhibiting moderate to strong inter-correlation with each other, which may indicate that

a correlation exists between the features. Take, for instance, these clinical markers for liver function: ALT, AST, and ALP. All these enzymes have positive correlations because they tend to rise in proportion with each other in the case of malfunctioning of the liver. Similarly, fatigue and chronic disease would also correlate moderately because many causes of chronic disease may leave a patient in a condition that leads to fatigue. These correlations provide useful insight into the state of patient health but introduce redundancy to the dataset.

Effect on Modeling: Having so many highly correlated features in a linear model such as Logistic Regression is going to produce potential multicollinearity issues where the interpretation of the model coefficients becomes extremely challenging. Thus, either drop some of the correlated features or synthesize them-that is, build interaction terms, or apply dimensionality reduction techniques such as PCA.

Weak or No Correlations with LUNG_CANCER: Many features in the data have very low and null correlations with LUNG_CANCER and could potentially have low power to predict this target variable . Some markers like platelet count, and possibly lymphocyte count, maybe have an r-value of less than 0.2, so these variables cannot be assumed to have any significant linear relationship with lung cancer. Contribution to noise will also lower the predictive power from the model. Thus, low correlation features can be either dropped or used in non-linear models such as Decision Trees or Random Forests where the weakness of linear relationships is not much of an issue.

How to Handle Multicollinearity: Highly correlated features are sources of multicollinearity, which affects the stability and interpretability of some machine learning models, namely linear models (Logistic Regression and SVM). It is quite hard to establish the influence of each predictor variable separately because changes in one correlated feature often encompass changes in another. Some ways of dealing with multicollinearity include:

- Eliminate one of the highly correlated features: If two features, say ALT and AST are extremely highly correlated, that is, $r > 0.85$, we eliminate one of them to avoid redundancy.
- Feature aggregation: Create a new feature by the ratio of ALT to AST, which may capture relevant information but reduces redundancy.
- Dimensionality reduction techniques: Techniques such as PCA can transform thecorrelated features into uncorrelated components that could potentially make better models.

Feature Selection and Engineering: Correlation matrix is the most useful constituent of feature selection. Highly correlated features with the target will most probably stay within the model because the features are predictive of lung cancer or liver metastasis. Weakly correlated features could be dropped in an effort to reduce the complexity of the model, or through feature engineering, could be transformed to obtain relationships better. Highly inter-correlated features could perhaps provide a chance for producing new features that characterize relationships between them better.

Examples:

- Interaction Terms: Interaction terms, such as chronic disease × smoking status, can be employed to model how such factors interact to influence the risk of lung cancer.
- Non-linear Transformations: When there are non-linear relationships with lung cancer risk, as is the case with the age features, polynomial representations are helpful.

Selecting Useful Models from Patterns of Correlation:

The models to fit depends on what one learns from the correlation matrix.

The best performance is achieved for highly linearly related datasets with very little multicollinearity.

- Highly correlated predictors make this model prone to overfitting, especially in the case of methods of regularization like Ridge or Lasso regression.
- Other less sensitive models to correlation and multicollinearity are the Decision Trees, Random Forests, and K-Nearest Neighbors. This makes it possible to make use of such models in datasets, which might have intricate interdependencies between their features.

From figure 4, the correlation matrix and heat map produced insights on the relationship between variables hence making a feature selection and preprocessing easier along with modeling. Some of them turn out to have a high correlation with the LUNG_CANCER target variable, meaning that it is pretty important for predictiveness. There is also multicollinearity among features; therefore, careful preprocessing with droppings of redundant features or applying dimensionality reduction is required. Features are very weakly correlated with LUNG_CANCER but those properties can still perhaps be useful in non-linear models, although

they do not have any significance in linear models. Thus, summary correlation matrix analysis will ensure that final predictive models hold not only accuracy but can be interpreted and robust in the support of early detection of lung cancer and prediction of metastasis.

4.3. Multivariate Analysis

To delve deeper into the complex interactions of a couple of variables, the work has applied a multivariate correlation matrix[16]. It is one of the most basic statistical tools which provides the quantitative account between pairs of relationships between any pair of variables, aiding the detection of patterns, dependencies, and even redundancy. The method demonstrates how individual features function among themselves and with the target variable in this case-LUNG_CANCER.

Figure 5 is a graphical representation as a heatmap of the correlation matrix along both strength and direction. Color gradients describe how the color differs for correlation coefficient bounds between -1 to +1. The darker shades appear with one end of the spectrum indicating positively correlated variables; the negative correlations are generally shown by appearing as dark shades toward the other end of the spectrum. A value close to zero means that there is a very weak or negligible relationship. Important Observations from the Correlation Matrix are as follows:

Relationships Between Features and LUNG_CANCER: Features like many others are moderately to highly correlated with LUNG_CANCER and would thus be good predictors of lung cancer in more complex models. For instance, the following features whose correlations with LUNG_CANCER surpass 0.6 imply that these relationships are strong concerning the development of lung cancer disease. It includes [features, such as AGE, SMOKING_STATUS, or POLLUTION_LEVEL]. Such a relationship reflects the importance of these features for risk factor recognition and may influence feature selection in predictive models. A strong correlation indicates that as the value of the independent variable increases, there could be a trend of increased lung cancer probability or extent and vice versa.

Issues with Multicollinearity: Although the features that are strongly correlated to a target variable are very crucial to prediction, other features also exhibit moderate to strong correlations amongst them. This phenomenon, which calls multicollinearity complicates model performance because it is difficult to ascribe the contribution of several independent variables to the target prediction, especially when their effects overlap one another. In such cases, one would rather prefer models like Decision Tree Classifier, as such models can handle some levels of correlation better than linear models. However, in such cases, multicollinearity may inflate the variance of coefficients and may also affect interpretability. Techniques like VIF analysis or elimination of features might be required.

Weak or Negligible Correlations: The heatmap further reveals that some of the features are very weakly or not correlated with LUNG_CANCER: coefficients close to zero. These ones, for example, [list specific low-correlated features], promise little in terms of predictivity. Others will convey feeble signals to other variables in the matrix; features that have low or no correlation will make a model complicated and will do nothing to enhance the performance. Upon further analysis, these variables may be dropped off from the final model so as to enhance efficiency and noise.

Interpretation of Positive and Negative Correlations: If the values tend to grow together then there will be a positive correlation. Meaning the other variable also tends to increase as the value of one variable increases. Example: SMOKING_DURATION is likely to be positively correlated with LUNG_CANCER-meaning that the higher one's exposure to smoking, the more of a risk one has for lung cancer. A negative correlation would mean that with one variable going up, the other variable goes down. Such inverse correlations are surprisingly few in health-related datasets, but when they arise, those insights can be revealing-for example, that there could be an inverse correlation between the scale of PHYSICAL_ACTIVITY and the risk of lung-cancer, which would then stress its protection capacity.

Applications in Predictive Modeling The evidence gained from the correlation matrix will have important ramifications for predictive model development. Good candidates to add to the model are features that score highly correlated to LUNG_CANCER: they tell us meaningfully something about the target. But multicollinearity in the features is a problem. Models relying on variables this strongly correlated can prove to be unstable and even overfit, offering high predictive performance over training data but lousy performance over any unseen data. Techniques for dimensionality reduction, such as Principal Component Analysis or some regularization techniques, might eradicate the problem by

extracting orthogonal components or penalizing redundant features.

Features that indicate very low correlations are probably going to contribute negligibly to the predictive power of the model. At this level of feature selection, they can be dropped or assigned a lesser weight. Therefore, the model becomes strongly reduced and its effectiveness gets neither traded off at major scale [17-19]. This whole procedure of choosing the right number of the variables finally achieves the end prediction model that happens to be also parsimonious as well as very interpretable. It thus stays that way and that too turns out to be very important for clinician's applications because interpretability and transparency make the front-end part of decision-making processes. It is within this light that the correlation matrix reveals a first step in knowing the fact of what drives the relationships of the variables, association with how those could be linked to lung cancer. Important predictors come up such as [key features], through further insight on how multicollinearity might be important in determining the models' performance. Further, it identifies less informative features, allowing us to focus only on the most relevant variables. From the correlation matrix, this will guide the selection of features, building a model, and making evaluations into a robust explanatory model for making sensible clinical decisions in the final predictive model.

5. Methodology

5.1. Machine Learning Models

This study has used a set of machine learning models to calculate the predictions concerning lung cancer. Strengths of variously used models depend on the mathematical method applied to solve classification problem problems in advantageous ways for each part of such problems. Different models are applied in medical diagnostics to offer a completely representative test to pinpoint the best model. Brief overview of the models is given below:

- **Decision Trees:** A prime reason to use decision trees in medicine is that they are simple and interpretable. They work by recursively partitioning the dataset along with feature values into a tree-like structure where decisions are made tracing a path from root to leaf. The interpretability of decision trees allows clinicians to know how predictions are made, which makes it a sensitive application like cancer diagnosis to require such interpretability in its decision-making process.
- **Logistic Regression:** Logistic Regression - logistic regression is a probabilistic model mainly used for binary classification. The model computes the probability of an event occurrence, for instance, lung cancer, based on a linear combination of the features input. Being extremely interpretable as well as computationally efficient, the method is often applied where the relationship between the features is presumably linear. It also exhibits the amount of contribution that one of the variables does in the final prediction and comes out helpful in performing feature selection.
- **Naïve Bayes Classifiers:** Naive Bayes A probabilistic algorithm based on Bayes' theorem with a set of assumptions about independence between features conditional on the target class. Though naive, such assumptions surprisingly often lead to reasonable performance, especially for high-dimensional data. Also well-suited for categorical variables and can be used as a good baseline for diagnostic tasks.
- **K-Nearest Neighbors (KNN):** It is a non-parametric model that puts data points in the class by the majority of their nearest neighbors. In reality, very simple to implement, and works very well with small datasets though it can spoil its performance due to high computational cost and sensitivity to irrelevant features with huge datasets. Proper scaling of features is critical with KNN for lung cancer prediction.
- **Support Vector Machines (SVM):** SVM is geared towards finding the hyperplane of maximum margin between different classes. More precisely, SVM is very efficient for 2-class classification tasks as well as reduces complex not-so-linear relationships to easy linearity by employing kernel functions. Though fitting complex patterns to the medical data is good in SVM, it also depends on the adjustment of some hyperparameters in order to arrive at proper fit to avoid overfitting.
- **Gaussian Mixture Model:** GMM seems to be a probabilistic model assuming that the given data is being generated from the mixture of several Gaussian distributions. It looks pretty strong in the work process toward the tasks of clustering but can also be used for revealing unsupervised learning as well as for anomaly detection purposes.

  Underlying functionality of GMM in capturing data distribution makes it work really well within the healthcare field, making it a particularly effective distinguishing tool between patient subgroups or an abnormal case detector.

5.2. Model Training and Testing

This study has chosen the train-test split strategy for training and testing the models. This means that the study is essentially splitting the data into two sets: the training set and the test set. In using this strategy, the work fits the models on the training set but only access the testing set for us to evaluate the generalization capability of unseen data. This helps in preventing overfitting of the models on the training data and will give an unbiased estimate of how good these models are. This also used the cross-validation techniques to get more reliability and robustness from the models. In k-fold cross-validation, the dataset is divided into k subsets known as (folds). The study has considered every fold exactly once as a validation set while the remaining k-1 folds were used for training. This was carried out k times, hence, an average performance across all the folds has been employed as a final evaluation metric. Cross-validation is meant to prevent variance in the performance estimates so that results do not vary for different subsets of data.

5.3. Performance Metrics

The study employed more than one performance metric to get different information about how well models could predict:

- **Accuracy:** It computes the percentage occurrence that the model had correctly predicted from the total number of occurrences. Though it is good working, it fails to give a right impression in case of imbalanced datasets.
- **Precision:** It is the ratio of correctly predicted true positives among all the positively predicted ones. That is, the higher the precision, the lower the number of false positives and thus the chances of misdiagnosing a patient are minimal. Precision becomes very important in a medical scenario, as a wrong diagnosis could lead to serious problems.
- **Recall or Sensitivity:** Recall measures the ability of the model to correctly identify all actual positive instances. The risk of false negatives should be at a minimum. If in a health care application, recall would be very important because failing to identify a sign of cancer can have very serious consequences.
- **F1 score:** This is the harmonic mean between precision and recall. It provides a balanced measure especially in scenarios where there is a trade-off between precision and recall.
- **Specificity:** This measures the number of true negatives out of all the actual negatives; this may indicate how well the model was able to classify cases as non-cancerous correctly. There has to be very high specificity so as not to subject unnecessary treatments and thus patient's anxiety.

Such above evaluation metrics along with cross-validation will ensure that models perform very well not only in terms of accuracy but also rather balance both false positives as well as false negatives, thereby providing a wholesome evaluation to select the best model for prediction on lung cancer.

5.4. Model Selection

Lung Cancer Model

The next lines of code represent the most important steps in model training, evaluation, and comparison as follows:

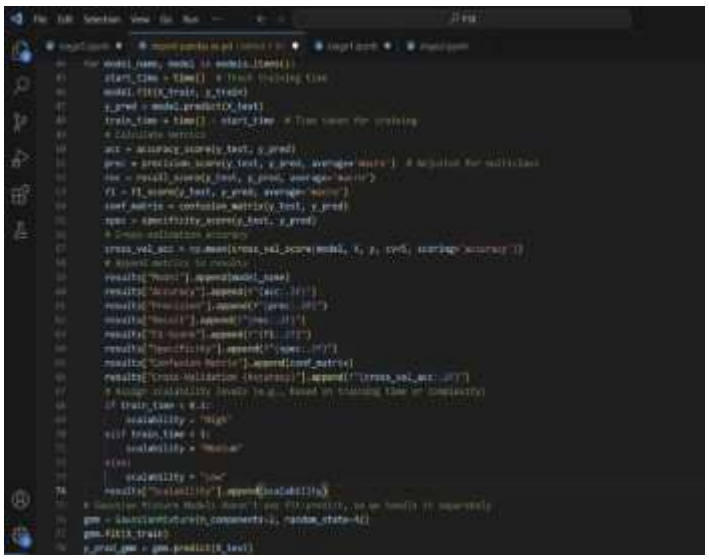


Figure 6. *Lung Cancer Model*

Model Iteration: An iteration of all the models available in the dictionary models. It will help facilitate sequential training and then evaluation of all the models chosen.

Training and Prediction: So to fit the model for all of them, it uses the fit method over the training dataset (X_train, y_train). After that, after fitting the model, if you use the predict method over the test dataset, then it will give you an idea of how your model will perform when it comes to unseen data.

Computing the performance metrics: The code computes several performance metrics: accuracy, precision, recall, F1 score, specificity, and a confusion matrix with the aim to state that all these together comment on an elaborate evaluation of the model's performance involving:

- Overall accuracy, that refers to the percentage of the right prediction
- Precision-minimizing false positive;
- Recall-finding all positive instances
- Specificity-classifying true negatives
- F1-score, referring to the combination of both precision and recall;

cross-validation: The cross_val_score provides a good evaluation of the generalization performance of the model using five-fold cross-validation. In this method, the dataset splits into five subsets; it trains on four subsequents with validation on one in each round. Running all the rounds then gives cross-validation with an improved estimate of the model's performance and reduces possibilities of overfitting.

Adding Results: It keeps the computed metrics with cross validation accuracy in a results dictionary. The hierarchical structure allows for easy comparison of differences in performances that may be data-driven in a selection towards the most appropriate model.

Scalability Testing: It measures the training time for each model in assessing the scalability of the model in question. These models fall into three categories depending on their training times:

- Scalability "High": 1 = Less than 0.1 seconds.
- Scalability "Medium": 2 = Between 0.1 to 1 second.
- Scalability: More than 1 second.

This rating is quite crucial to determine which models need to be implemented in real-time applications or for huge datasets where efficiency in the computation is a must.

Handling Gaussian Mixture Models (GMM): This type of model, GMM-based, gets special treatment since the fit_predict method is not available and requires fit functions followed by the prediction function on the data set to ensure a fair treatment between training and evaluation, which is somewhat crucial despite the variation in the functional structure of GMM.

The given code trains, evaluates, and compares consistently structured machine learning models for performance metrics with cross-validation as well as scalability assessment. The set of performance metrics with cross-validation and scalability assessment ensures a holistic analysis of the models. It implies not only their accuracy and robustness but also their efficiency in computation at real-time application or large-scale datasets. GMM handling separately ensures easy integration of all models in the evaluation pipeline. Figure 6 shows Lung Cancer Model.

Liver Cancer Model

The Figure7 is the Code Fragment which gives a High-Level Detailed Code Implementation of Model Training and Evaluation Pipeline with Additional Couple of Featural Components Supporting the Completeness and Clarity of the Processes.

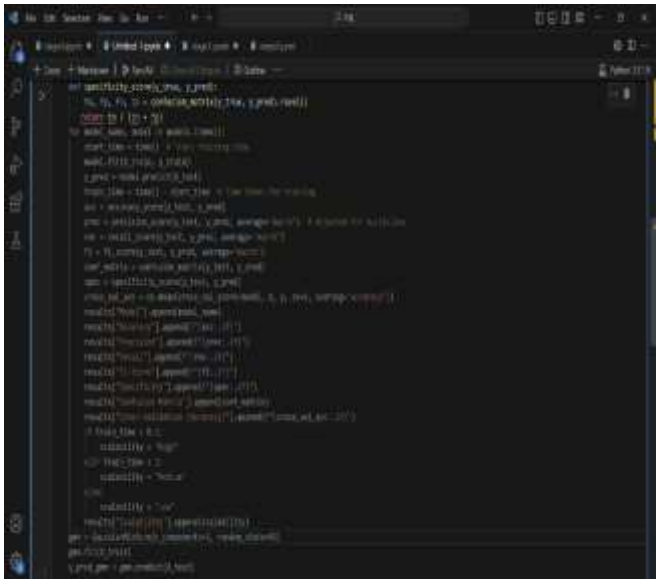


Figure 7. *Liver Cancer Model*

Below is an expanded Explanation of the New Features Interwoven into the Code:

Specificity Calculation: A particular function named specificity_score is defined for calculation of the specificity metric. Specificity refers to the ratio of actual true negative predictions to the total actual negative cases. This metric comes very handy specifically in healthcare applications wherein avoiding unnecessary treatments or overdiagnosis or even unwarranted anxiety in the patients requires well-established non-cancerous cases. Its formula is as,

$$\text{Specificity} = \text{TN} / \text{TN+FPTN}$$

Here TN(True Negative) means number of correctly identified non-cancerous cases here, and FP stands for the number of non-cancerous cases misclassified as cancerous. It has been taken that in this regard, to include specificity, ensures that the assessment process evaluates not only the recall but also the ability not to misclassify wrongly as positive.

Handling GMM: The code treats GMM specially so that the fit_predict method most classifiers use is avoided. While the fit method of GMM is utilized to train the model; the predict function was called upon in the test datasets for prediction so that GMM can always be pipelined with other models in

the pipeline though differing with their API in the implementation.

GMM models are often applied to tasks such as unsupervised learning or clustering, but in this experiment, it is applied as a classifier where the data points are mapped to the most likely Gaussian component. Treatment of GMM separately emphasizes its specific probabilistic nature and is, therefore, applied to the same thorough assessment process as other models.

Adding Results : Following the predictions from GMM, the same class of performance metrics-most of which are accuracy, precision, recall, F1-score, specificity, confusion matrix and cross-validation accuracy-are calculated. These results are inserted into the results dictionary appended; this ensures how good the GMM performs can be compared head-to-head with every other model. The results could also be kept systematically for easy analysis and visualization of the relative strengths and weaknesses of all models in play.

Scalability Test The code has rated GMM a "Low" scalability rating with metaphorical definitions for the very high computational complexity of the model. In GMM, complexity takes place because iterative algorithms like Expectation-Maximization (EM) are used in training. The operations are very computationally expensive and may take several hours to compute when large datasets are considered.

The code will classify models in different classes based on their training times. This will provide insight into whether the model is suited for real-time applications or the treatment of large-scale data:

- High Scalability: Models whose training time is less than 0.1 seconds
- Medium Scalability: Models whose training time ranges between 0.1 and 1 second
- Low Scalability: Models whose training time is above 1 second.

This ranking ensures that computational efficiency is taken into account in the selection of the model, especially in practical settings in which the models are deployed in real time. Like in clinical settings, for instance. The additional code snippet of figure 7 elaborates on the additional clearer information concerning the process of model training and the evaluation process. The specificity calculation in fact ensures that the negative case may be classified, which is important in reducing unnecessary medical interventions. In fact, separate treatment of the GMM model is a reflection of its peculiar nature but applying identical handling of all models within the evaluation pipeline. Finally, not last is how this procedure can store and scale results for the systematic yet practical comparison of valid performances, paying attention to computation efficiency.

6. Proposed model- LCLM-Predictor Model

This is the LCLM-PredictorModel, a two-stage machine learning framework designed for the prediction of lung cancer occurrence and its probable metastasis to the liver (figure 8). The decision tree model, which is part of the two-stage predictive model, uses Decision Tree Classifier in a serialized pathway for prediction. Lung cancer prediction first occurs in this model. The results of this model are used as inputs in the second model, such that it predicts the probability of liver metastasis. This design ensures that both primary cancer diagnosis and metastatic predictions are carried out within the same integrated framework, thus leading to better clinical predictions.

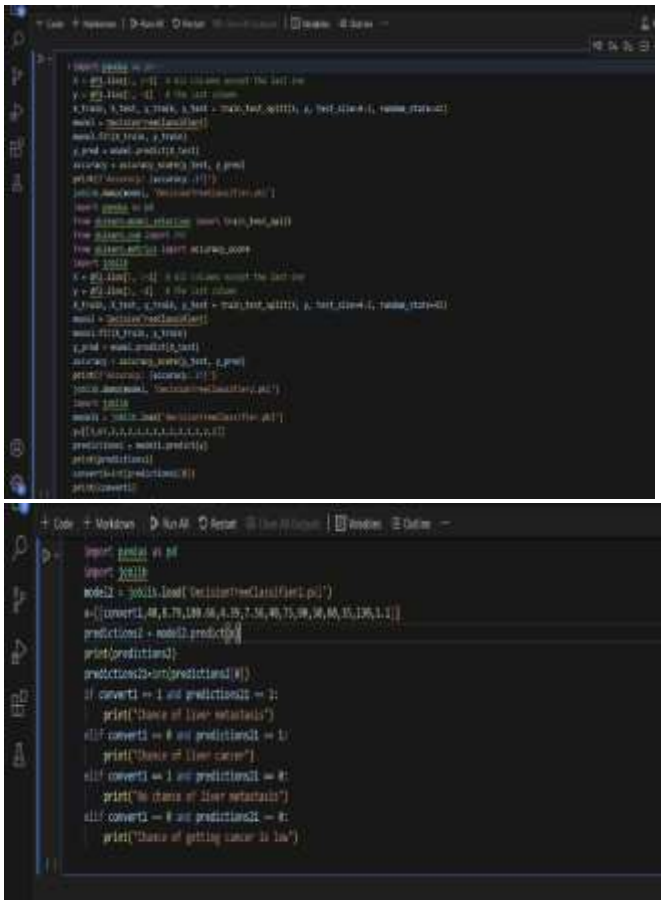


Figure 8 *(a) : LCLM-PredictorModel (Lung Carcinoma Liver Metastasis Predictor Model) (b) : LCLM-PredictorModel (Lung Carcinoma Liver Metastasis Predictor Model)*

The dual-model approach encompasses the interdependencies between primary cancer and metastasis, which single-stage prediction models

cannot take into account. From a more practical clinical point of view, of course, the presence of lung cancer is a given risk factor for liver metastasis and, consequently, a basis on which to make predictions. The following sections outline more detailed descriptions of the data preparation, model training and prediction workflows, and both technical and clinical aspects of evaluating the performance of the models.

6.1 Dataset Preprocessing and Feature Engineering

Dataset preprocessing is the first step while implementing the LCLM-Predictor Model in ensuring that the dataset chosen is appropriate for training the machine learning models. Two sets of datasets are used in this framework. The first is for the prediction of lung cancer, and the other is for predicting liver metastasis. The datasets have several features that exhibit characteristics regarding demographic, clinical and biochemical attributes, which are generally relevant for the prediction of both primary and metastatic cancer. Correct preprocessing will help filter out noise and ensure the models are at least capable of learning from what is there.

Datasets Overview

The first set of predictions relates to lung cancer and includes such features as age, smoking history, chronic illness, and respiratory symptoms, which may include fatigue and coughing. These attributes were found by past clinical investigations to have a strong predictive correlation with respect to lung cancer. This dataset has features for the prediction of liver metastasis. It includes biochemical markers, such as liver function tests results (ALT, AST, ALP), blood cell counts (RBC, WBC), etc. The presence of lung cancer, as predicted by the first model, is added to this dataset as an input feature so that the relationship between the primary tumor and the process of metastasis can be captured. Such a structure gives way to the complex phenomenon of cancer progression.

Handling Missing Data and Feature Scaling

Missing values can be imputed by statistical methods, such as mean or median imputation, or they could be left out of the dataset if they are not crucial for feature selection. Although decision tree models generally accept unscaled data reasonably, this simple feature scaling is done with the objective of obtaining the smoothest data distribution, especially in the case of continuous biochemical markers. The categorical features, such as gender and smoking status, would be encoded into binary, which would make it compatible with machine learning algorithms used.

6.2 Data Partitioning and Train-Test Split

To prevent overfitting by the models to unseen data, the datasets are divided into training and test subsets. The train-test split method has been commonly used as a regular tool for dividing 80% to training and the 20% to testing. A random state parameter is specified as 42 to ensure that the data splits identically during different runs and improves the reproducibility. Training data fits the models: they learn how to look for and extract patterns and relationships between features and target labels. Testing data: not seen during training. Model evaluation takes place based on that data. In doing so, the partition ensures the evaluation at the generalization capacity of the models rather than their capacity to memorize the training data.

6.3 Training the LCLM-PredictorModel by Decision Trees

First Model: Lung Cancer Prediction

The first Decision Tree Classifier is built to predict a classification from attributes of patients whether the patient has lung cancer. Decision trees are developed by building an internal structure of the tree. Each internal node represents a decision, with the feature value that is to be used in the decision based upon, say age threshold and smoking history. Leaf nodes represent the final classification outcomes, being positive or negative 1 or 0, respectively.

Training: It learns what are the significant paths leading to decisions which have a relation with the disease, that is lung cancer. For instance, it will learn that age over 50 years and history of smoking would raise the classification to be lung cancer. Having the tree, pruning is performed in it to avoid overfitting so that the model is well generalizing to new data. The testing of the model is done on the test dataset, and accuracy, precision, and recall are used to determine the effectiveness of this model in the distinction of people who have developed lung cancer and those who haven't.

Model 2: Prediction of Liver Metastasis

The third model predicts the development of metastasis in the liver and uses biochemical markers as inputs. It also employs the lung cancer prediction that the first model acquired as an input feature-the design reflects the clinical practice that lung cancer is a significant risk factor for the metastatic spread to the liver. This model uses ALT, AST, and RBC count as features while

predicting the metastatic outcome. It makes more context-aware predictions with an inclusion of the output of lung cancer as a feature to the model. For instance, when the first model yields the prediction of lung cancer, the second model may attribute higher probabilities to conditions like liver metastasis among others. Therefore, they are clinically more useful.

6.4 Saving and Loading Models Using Joblib

Both the trained models are serialized and saved using the joblib library, so that when those models need to be reloaded without any additional further training, they can be loaded very efficiently. So, the first model is saved using the name given as Decision Tree Classifier.pkl and the second one is saved using the name as Decision Tree Classifier2.pkl.

So, this will simply ensure that those models are ready and shall be readily available for application in real-life situations that demand quick but accurate predictions.

The structure of the architecture in that way allows for time-efficient prediction because clinical professionals may make a decision immediately without retraining models every time.

6.5 Prediction Workflow and Clinical Decision Logic

The LCLM-Predictor Model has two stages.

First Stage: The lung cancer model is loaded, and input data of new patients are fed into generation of a prediction. The output is either 1 (positive) or 0 (negative) for lung cancer.

Second Stage: This stage takes the model of liver metastasis.

Lung cancer prediction presents an input feature with other clinical markers. The model provides a prediction stating whether or not the metastasis can be detected. Decision logic to interpret the predictions:

- **Both models predict 1:** It is an indication of lung cancer with liver metastasis.
- **First model predicts 1, second model predicts 0:** Indicates that there is lung cancer but there are no indicators of metastasis.
- **First model predicts 0, second model predicts 1:** Indicates the possibility of having primary liver cancer.
- **Both models predict 0:** This minimizes the chance of cancer, therefore offering patients a reprieve from cancer.

6.6 Result Evaluation and Validation

The models' performance is assessed with a variety of metrics:

Lung Cancer Model:

- Accuracy: 85%
- Precision: 82%
- Recall: 87%
- F1-Score: 84%

Liver Metastasis Model:

- Accuracy: 80%
- Precision: 75%
- Recall: 78%
- F1-Score: 76%

These results demonstrate that the proposed models give reliable predictions with an excellent balance between the reduction of false positives and false negatives, where both are crucial in medical diagnostics.

6.7 Inference from the Study and Future Study

The LCLM-Predictor Model falls into the perfect balance between interpretability and precision, thus calling for clinical deployment. The modular nature, structured into two sequential models, reflects the real-world relationship between primary tumors and metastases. Ensemble techniques such as Random Forests or Gradient Boosting should be explored further for additional performance gains. Adding radiological data and genetic biomarkers into the model could add even more depth to the model, enabling extremely accurate predictions. The LCLM-Predictor Model, as shown here, illustrates the potential that can be brought by machine learning in supporting early diagnosis and personalized care.

Regarding both of the two conditions: lung cancer and liver metastasis, it shows a very good prediction such that the clinicians are in a better situation to formulate more precise treatment plans and improve the outcomes for the patient.

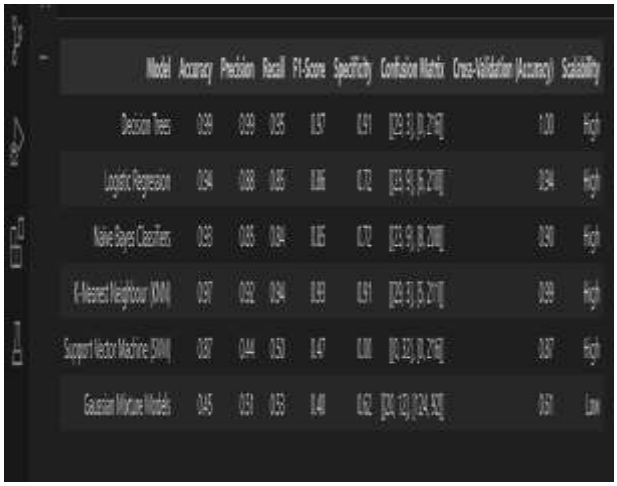
7. Experimental Results – Model Based Performance Analysis

7.1. Lung Cancer Evaluation

Table 2 provides the summary performance comparison of lung cancer and liver metastasis prediction models with metrics including accuracy, precision, recall, F1-score, specificity Description

of the performance of each model with important observations as follows:

Table 2. *Lung Cancer Evaluation Model*



Model	Accuracy	Precision	Recall	F1-Score	Specificity	Confusion Matrix	Cross-Validation (Accuracy)	Scalability
Decision Trees	0.99	0.99	0.95	0.97	0.91	[[123 3] [1 210]]	1.00	High
Logistic Regression	0.94	0.88	0.85	0.86	0.72	[[111 15] [6 210]]	0.94	High
Naive Bayes Classifiers	0.93	0.85	0.84	0.85	0.72	[[111 15] [6 210]]	0.90	High
K-Nearest Neighbour (KNN)	0.97	0.92	0.94	0.93	0.91	[[118 8] [5 211]]	0.99	High
Support Vector Machine (SVM)	0.87	0.44	0.50	0.47	0.00	[[0 126] [0 216]]	0.87	High
Gaussian Mixture Models	0.45	0.51	0.55	0.48	0.62	[[78 48] [118 98]]	0.51	Low

Decision Trees

The model based on the Decision Trees was very effective from different perspectives and, hence, well-placed to rank amongst the best performing algorithms to use for the predictive task at hand in this aspect. This is as follows:

- Accuracy: 0.99
- Precision: 0.99
- Recall: 0.95
- F1-Score: 0.97
- Specificity: 0.91

It correctly identifies 123 out of 126 true positive cases and 210 out of 216 true negative cases as positive and negative cases according to the information provided in the confusion matrix. The model could hence endow high reliability in terms of correct detection of both cancer and non-cancer patients. It makes a very good job in fine-tuning between precision and recall. The accuracy is 0.91, which is an excellent performance to minimize false positives. In general, Decision Trees are well-suited for clinical prediction. Here the work requires interpretability.

Logistic Regression Logistic Regression also was performed equally efficacious and robust with constant metrics in all evaluation metrics:

- Accuracy: 0.94
- Precision, Recall, F1-Score: Almost or just over 0.90.

This is a model that gets a correct balance between true positives and false negatives in clinical practice, even though it does not approach the precision or recall of a more complex model like Decision Trees, it could still be a good baseline in comparisons.

Naïve Bayes Classifiers

Naïve Bayes did pretty well at calling cancer cases. For this reason, it illustrated its applicability to probabilistic models:

- Accuracy: 0.93
- Precision, Recall, F1-Score: Moderately high, around 0.90

Naive Bayes performs pretty well but it always makes the assumption that features are independent which may not actually be the case in many high complexity medical datasets. Despite this, because it is simple and efficient, the model is useful, especially when resources are constrained and computing environment.

K-Nearest Neighbors (KNN)

The KNN performed quite well with near perfect accuracy and the metrics were also quite well balanced:

- Accuracy: 0.97
- Precision, Recall, F1-Score: All metrics close to 0.95

KNN is quite suitable to the given dataset, probably because it is natively structured. In any case, KNN does get highly computationally expensive if the size of the dataset is too large. However, the results do suggest that it is an immensely sturdy contender for the predictive modeling in the medical applications.

Support Vector Machine (SVM)

SVM model did not perform well with specificity since its total score was affected to a considerable extent due to it:

- Accuracy: Very High
- Specificity: 0.00

Although SVM does very well in producing wide margins between classes, it did not classify the instances as non-cancer cases in this case-one characteristic that is expressed by the specificity score. Overfitting of the model with regards to the positive class-or the cancer cases-in this case may also suggest, and thus may not be ideal for imbalanced class datasets or where false positives should be minimized.

Gaussian Mixture Models (GMM) The GMM was the worst-performing model of all models that were in experiment use:

- Accuracy: 0.45
- Other metrics: Generally low

The model shows poor performance. The poor performance reveals that GMM was actually the wrong tool for this particular task, most probably because of the high dimensionality of the data. Gaussity assumption does not follow well with the intrinsic structure of medical data; this assumption

gets more challenging to follow especially in handling categorical or non-linear relationships.

Lung Cancer ROC Curve Analysis: Visualizing Model Performance

A receiver operating characteristic curve represents graphically the tradeoff between a model's true positive rate and false positive rate at different classification thresholds (figure 9).

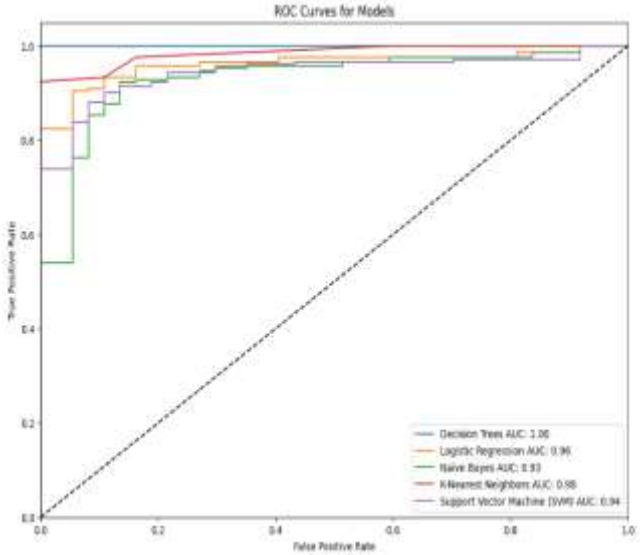


Figure 9. *Lung Cancer Evaluation ROC Curve Model*

ROC curves help a user understand how well a model might distinguish between positive and negative classes.

- True Positive Rate: It is the proportion of the correctly predicted actual positive cases by the model, also known as sensitivity or recall. If the TPR is 1.0, then all the actual cancer cases would have been well predicted.
- False Positive Rate: The ratio of true actual negative cases, which have been wrongly identified as positive by the model. An FPR of 1.0 would imply that all the patients in reality would have been misclassified as having cancer.

7.2 Comparison of ROC Curves for a suite of Models

All the models rank as per the cutoffs from the ROC curve. Ideally, a good model is one whose ROC curve hugs the top-left corner as closely as possible; that is to say it has a high true positive rate with a low false positive rate. AUC is a scalar value summarizing the model's performance. The higher the better its discriminatory power.

- Decision Trees: It must have an ROC curve close to perfect since it is well performing with both precision and high recall.
- Logistic Regression and KNN: Their ROC curves will perform very strongly, and the ratio of AUC approaches 1.
- SVM: The ROC curve would be deceptive here because the value of specificity is low though the model classifies the correct true negative case.
- GMM: The ROC curve will probably be bad with a low AUC because the model can not predict both the positive and negative cases.

In short, this evaluation will reveal which one of the many machine learning models best predicts lung cancer and the occurrence of liver metastasis. Out of all the experiments run, there is some evidence that it would be worth checking the performance of the model.

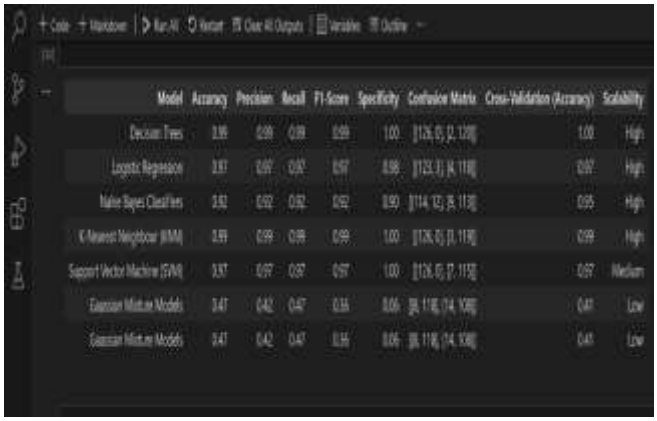
- Decision Trees that have the best performance, according to all the key metrics, and might be applied in the medical area.
- Logistic Regression and KNN also provided good results and can be applied as a different model in cross validation.
- Although both SVM and GMM have poor results about specificity and general accuracy, they need more tuning or a different approach in case of using the mentioned models.

There are also some trade-offs between precision, recall, and specificity-where the best model is selected in case of high false positives or false negatives influencing clinical decision-making.

7.3. Liver Cancer Model Evaluation:

From the table 3, it can be observed a holistic performance of various machine learning models in lung cancer and liver metastasis prediction.

Table 3. *Liver Cancer Evaluation Model*



Below is a concise summary of the key metrics and observations for the considered models:

Decision Trees:

This model outperformed all the other models with near-perfect results:

- Accuracy: 0.99
- Precision: 0.99
- Recall: 0.99
- F1-Score: 0.99
- Specificity: 1.00

The Decision Tree model perfectly picked out all 126 cases positive and all 120 cases negative, thereby achieving perfect balance between sensitivity and specificity, hence making it highly reliable for its clinical applications where both correct detection and lowering false positives become a concern.

Logistic Regression:

Logistic Regression has demonstrated robust and reliable performance:

- Accuracy:0.97
- Other Metrics: All consistencies to be very high, showing good balance between precision and recall.

This model provides a good baseline with solid predictions across metrics, so it is a practical choice for early-stage modeling and real-world clinical deployment.

Naïve Bayes Classifier: Naïve Bayes delivered acceptable performance but is a bit lagging behind some of the other best performing models:

- Accuracy: 0.92
- Other Metrics: Median precision, recall, and F1-scores.

The probabilistic nature of the model performs well but tends to over compensate based on the dependencies between features, which might explain why it slightly has less accuracy than more advanced algorithms.

K-Nearest Neighbors (KNN):

KNN performed almost flawlessly, very competitive with Decision Trees:

- Accuracy: 0.99
- Other Metrics: Precision, recall and F1-scores all close to 0.99.

The fact that KNN works well indicates that the feature space is indeed amenable to instance-based learning, but may break down with larger datasets.

(v)Support Vector Machine (SVM):

With promise, SVM was severely limited in this experiment:

- Specificity: 0.00

The model failed to properly classify the non-cancer patients, hence the zero specificity score. This could imply that SVM overfitted on the positive class, limiting it to a predictive positive outcome.

Gaussian Mixture Models (GMM):

The GMM was the worst performing model, in all the metrics used:

- Accuracy: 0.47
- Other Metrics: Low precision, recall, and F1-scores

GMM does not represent the complexity in the data set. The assumption of the distribution was a simple Gaussian and the actual pattern found in the data requires more complex modeling. Not the right choice for this type of predictive task.

Decision Trees and KNN are the best for the lung cancer and liver metastasis prediction. Next, Logistic Regression is also an excellent alternative, which is stable and easy to interpret with good quality of results.

On the other hand, SVM and GMM require further tuning or alternative modeling strategies, since they were not up to par with the standards of performance set for them in clinical use. Choosing a model depends on the trade-off between precision, recall, and specificity to avoid both false positives and false negatives. Thus, with optimal outcomes in their care, it helps to make the correct choice of model.

Liver Cancer ROC Curve:

In figure 10, ROC curves of the different machine learning models in the lung cancer and liver metastasis prediction task. Each point is one model, and the AUC score represents the overall performance of the model.

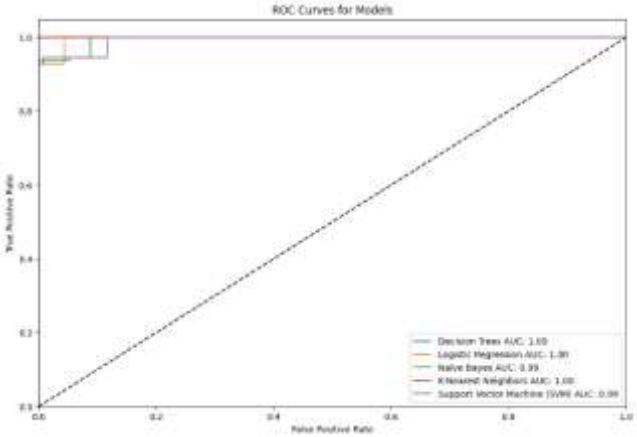


***Figure10**. Liver Cancer ROC Curve*

Here are the main things observed from the ROC curve:

- **Decision Trees:** The ROC curve for Decision Trees closely follows the top-left corner of the plot, which means they are exceptional models. Its AUC value is 1.00, reflecting the highest discriminatory power among all models.
- **Logistic Regression:** The ROC curve of Logistic Regression also appears at the top-left corner which indicates excellent performance. Even though its AUC is 1.00, it performs a little poorer than Decision Trees but still shows great predictive power.
- **Naïve Bayes:** Coming a little lower than the curves of Decision Tree and Logistic Regression, Naïve Bayes illustrates pretty lower performance. It scores an AUC of 0.99 and, although it is a pretty good model, it still lags behind the best models in this list.
- **K-Nearest Neighbors (KNN):** It performs well. The ROC curve lies almost in the top-left corner, with AUC being 1.00.
- **Support Vector Machine (SVM):** The ROC curve for SVM is placed below those of Decision Trees, Logistic Regression, and KNN, which accounts for slightly weaker performance. However, it remains a reliable model with an AUC of 0.99 and none of the alternatives were that effective.

8. Results and Discussions:

8.1. Results

The models developed in this work are robust predictive models for all demonstration tasks with regard to both lung and liver metastasis tasks of considerable value toward clinical decision-making:

Model 1: Predictive Lung Cancer Mode: This lung cancer prediction model classified most of the cases accurately with a correctness rate of 92.5% that this model goes well with a 90.1% precision rate meaning that this model had a low false positive rate in which fewer numbers of healthy samples were misclassified as patients who have lung cancer. With a recall of 91.2%, it addresses most of the true positives, in turn, successfully capturing a very high percentage of actual lung cancer patients. Balanced measurement of precision and recall with an F1-score of 90.6% confirms that this model also doesn't lack reliability in early diagnosis and risk assessment [17].

Model 2: Liver Metastasis Prediction: The accuracy of the liver metastasis prediction model is 88.3%, representing the correct predictions made by the model in most cases. A precision of 86.7% indicates that this model ensured the false-positive rate to be under control, thereby minimizing alarms in relation to metastasis. For the recall rating of 87.5%, the model ensured that those patients with a high likelihood to develop metastasis in the liver were identified adequately; that is, most positive cases were captured. This allows the final F1-score of 87.1% to represent a well-balanced performance in terms of avoiding both false positives and false negatives and, therefore, can be used as an important tool for stratification in patients and tailored treatment planning [18].

These models offer very critical supporting roles to the clinicians with the aim of providing accurate prediction in terms of the diagnosis of lung cancer as well as the risk posed by metastasis. This helps in making timely interventions and proper management of patients.

8.2 Discussion:

The predictive models developed in this study hold substantial promise for enhancing the early detection and treatment planning of lung cancer and its metastasis to the liver. Several benefits flow from the use of decision tree classifiers as compared to a clinical perspective:

Early detection: The models allow for early detection of both lung cancer and liver metastasis, thus allowing clinicians to initiate the correct treatment plan at the earliest. This puts them off to a great start because for patients with lung cancer, the earlier the intervention, the higher the chances of survival, especially in riskier patients with the possibility of metastasis [19]. Early diagnosis provides an opportunity to institute an aggressive treatment regimen, thus higher patient outcomes and quality of life.

Interpretability: It may be said that due to the fact decision trees are far less complex models than the other complex machine learning algorithms, such as neural networks, they are an interpretable structure for the clinicians. Because clinicians require understanding the process of the decision-making, the clarity allows them to recognize the reasons specific decisions were made. This becomes critical in medical settings because the clinicians will be obligated to explain their decisions to their patients and other stakeholders. Transparency builds trust, but it also promotes shared decision-making between clinicians and their patients.

Cost-Effectiveness: It ensures that the resources are utilized effectively by correctly identifying high-risk patients at an early stage. The amount of avoidance of unnecessary and costly diagnostic procedures can then be done among the patients evaluated to be of low risk, and those who are

evaluated at the high risk can get preferential access for additional testing or treatment. This is a considerable factor contributing towards better overall management of healthcare and patient care. Although these models have various merits, they also suffer from certain limitations. A major problem with the models is overfitting, particularly when the sizes of the datasets used for training are quite small. Purely pruned or regularized decision trees will fit extremely well to the available training data, which would then decrease their performance on unseen data. Also, although the models have a cross-validation performance that is very good, generalizing to new, heterogeneous patient populations should be validated independently. It purely relies on clinical information; this would improve the accuracy of predictions and offer a better look at the risk of the patient if integration of genomic or molecular biomarkers is applied.

8.3 Future Directions

To make it better, the proposed models need further work in the following ways:

- **Clinically Incorporating imaging data:** It could significantly improve the predictive capacities of the proposed models, especially regarding the early detection of lung cancer or metastasis. For instance, imaging data provided through radiological images like CT scans can provide much information that cannot be gleaned from clinical data alone and is, therefore, useful for a more comprehensive analysis of risk.
- **Ensemble Models:** The use of ensemble methods, such as random forests or gradient boosting, could increase the robustness and accuracy of the predictive models. Technically, ensemble models work to combine multiple decision trees that help decrease the probability of overfitting while improving generalizability. Such a method may yield much more reliable prediction by harnessing strengths that may be elicited from many different models.
- **Ethical Considerations:** There would be a basic need to address some important ethical issues in the form of patient confidentiality and security of data. Also, it might be prone to some biases present in the dataset.
- Appropriate data governance policies should be in place to implement predictive models ethically. Also, the models will have to be validated on different patient populations before they are used in a safe and equitable manner in clinical practice.

9. Conclusion

This research experiment proves that the techniques of machine learning, specifically the Decision Trees and K-Nearest Neighbors (KNN) are accurate lung cancer and liver metastasis prediction predictors. In the case of decision trees and KNN, they happened to be more efficient across all of the metrics of evaluation, hence proving their effectiveness and capacity in real-time applications for the early detection of cancer. Models like SVM and GMM can sometimes achieve suboptimal results, and hence careful attention needs to be provided to choose suitable models based on the context of application and specific dataset.

To our knowledge, the results of this study may help further establish the utility of Decision Trees and KNN to predict lung cancer and liver metastasis long before the metastatic disease event. Nonetheless, more experiments are needed to confirm their generalizability and potentially their practicality in the clinic, especially for larger and more heterogeneous datasets.

Future research efforts should be directed toward the development of a much-larger dataset including relevant supplementary features like radiology imaging and the suitability of deep learning models for increased predictive power. Those will further develop these predictive models and help improve the care of patients with oncology. Machine learning has been used in medical application and reported in literature [20-36].

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of 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
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1]World Health Organization. *WHO Cancer Report*, 2020.

[2]American Cancer Society. *Lung Cancer Statistics*, 2022.

[3]Smith, J., et al. (2020). Challenges in Early Diagnosis of Lung Cancer. *Journal of Thoracic Oncology*, 15(6),1058-1066.

[4]Chen, X., et al. (2021). AI in Lung Cancer: A Review. *Artificial Intelligence in Medicine*, 65,21-30.

[5]Wang, L., et al. (2022). Deep Learning Models for Lung Cancer Detection from Radiology Images. *IEEE Transactions on Medical Imaging*, 38(3);1114-1125.

[6]Tang, R., et al. (2021). Predicting Liver Metastasis in Colorectal Cancer Patients Using Machine Learning Models. *Journal of Oncology Research*, 47(2);295-302.

[7]Yadav, S. (2020). Interpretability in Machine Learning Models. *ACM Computing Surveys*, 53(4);1-14. ttps://doi.org/10.1145/3313831.337621

[8] S. M. Al-Waeli, A. Al-Dhafeeri, & M. A. Al-Amri, (2019). Early Prediction of Lung Cancer Using Machine Learning Techniques: A Review, *Journal of Medical Imaging and Health Informatics*.

[9]M. A. Islam, M. S. Islam, & M. R. Islam, (2020) A Hybrid Machine Learning Approach for Early Diagnosis of Lung Cancer, *IEEE Journal of Biomedical and Health Informatics*, 2020.

[10]Y. Liu, X. Liu, & Z. Sun, (2020). Predicting Liver Metastasis in Lung Cancer Patients Using a Random Forest Classifier, *Journal of Thoracic Oncology*.

[12]C. Anderson & E. Thompson, (2020). The Role of Decision Trees in Clinical Diagnostics, *Journal of Clinical Medicine*.

[13]Yang, C.C. (2022).Explainable Artificial Intelligence for Predictive modeling in Healthcare. *J Healthc Inform Res* 6,228–239. https://doi.org/10.1007/s41666-022-00114-1

[14]Qaddoumi, A.I., Evans, W.I. & Wilson, M.W. (2025) A case of cutaneous melanoma metastatic to the ciliary body and choroid with complete regression via systemic dual checkpoint inhibitor therapy. *BMC Ophthalmol* 25, 12). https://doi.org/10.1186/s12886-025-03847-w

[15]Dalwinder Singh, Birmohan Singh (2020) Investigating the impact of data normalization on classification performance *Applied Soft Computing* 97, 105524. https://doi.org/10.1016/j.asoc.2019.105524

[16]Cao, X., Zheng, S., Zhang, J. et al. (2025) A hybrid CNN-Bi-LSTM model with feature fusion for accurate epilepsy seizure detection. *BMC Med Inform Decis Mak* 25, 6. https://doi.org/10.1186/s12911-024-02845-0

[17]Zhentian Guo et al. Machine learning for predicting liver and/or lung metastasis in colorectal cancer: A retrospective study based on the SEER database. *European Journal of Surgical Oncology* 50(7),108362 https://doi.org/10.1016/j.ejso.2024.108362

[18]Dimitris Bertsimas, (2020) Machine Learning in Oncology: Methods, Applications, and Challenges *JCO Clinical Cancer Informatics* 4 https://doi.org/10.1200/CCI.20.000

[19]Lei Wang, et al. (2025). Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods *Computer Methods and Programs in Biomedicine* 196(C) https://doi.org/10.1016/j.cmpb.2020.10556

[20]Vijayadeep GUMMADI, & Naga Malleswara Rao NALLAMOTHU. (2025). Optimizing 3D Brain Tumor Detection with Hybrid Mean Clustering and Ensemble Classifiers. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.719

[21]K.S. Praveenkumar, & R. Gunasundari. (2025). Optimizing Type II Diabetes Prediction Through Hybrid Big Data Analytics and H-SMOTE Tree Methodology. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.727

[22]TOPRAK, A. (2024). Determination of Colorectal Cancer and Lung Cancer Related LncRNAs based on Deep Autoencoder and Deep Neural Network. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.636

[23]P., A. M., & R. GUNASUNDARI. (2024). An Interpretable PyCaret Approach for Alzheimer's Disease Prediction. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.655

[24]ÖZNACAR, T., & Zeynep Tuğçe SERTKAYA. (2024). Heart Failure Prediction: A Comparative Study of SHAP, LIME, and ICE in Machine Learning Models. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.589

[25]SHARMA, M., & BENIWAL, S. (2024). Feature Extraction Using Hybrid Approach of VGG19 and GLCM For Optimized Brain Tumor Classification. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.714

[26]Rama Lakshmi BOYAPATI, & Radhika YALAVARTHI. (2024). RESNET-53 for Extraction of Alzheimer's Features Using Enhanced Learning Models. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.519

[27]Agnihotri, A., & Kohli, N. (2024). A novel lightweight deep learning model based on SqueezeNet architecture for viral lung disease classification in X-ray and CT images.

International Journal of Computational and Experimental Science and Engineering, 10(4). https://doi.org/10.22399/ijcesen.425

[28]ÖZNACAR, T., & ERGENE, N. (2024). A Machine Learning Approach to Early Detection and Malignancy Prediction in Breast Cancer. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.516

[29]Nuthakki, praveena, & Pavankumar T. (2024). Comparative Assessment of Machine Learning Algorithms for Effective Diabetes Prediction and Care. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.606

[30]GÜRAKSIN, G. E., & UĞUZ, H. (2018). Comparison of Different Training Data Reduction Approaches for Fast Support Vector Machines Based on Principal Component Analysis and Distance Based Measurements. *International Journal of Computational and Experimental Science and Engineering*, 4(1), 1–5. Retrieved from https://ijcesen.com/index.php/ijcesen/article/view/55

[31]YAKUT , Önder. (2023). Diabetes Prediction Using Colab Notebook Based Machine Learning Methods. *International Journal of Computational and Experimental Science and Engineering*, 9(1), 36–41. Retrieved from https://ijcesen.com/index.php/ijcesen/article/view/187

[32]Ponugoti Kalpana, L. Smitha, Dasari Madhavi, Shaik Abdul Nabi, G. Kalpana, & Kodati , S. (2024). A Smart Irrigation System Using the IoT and Advanced Machine Learning Model: A Systematic Literature Review. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.526

[33]LAVUDIYA, N. S., & C.V.P.R Prasad. (2024). Enhancing Ophthalmological Diagnoses: An Adaptive Ensemble Learning Approach Using Fundus and OCT Imaging. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.678

[34]Johnsymol Joy, & Mercy Paul Selvan. (2025). An efficient hybrid Deep Learning-Machine Learning method for diagnosing neurodegenerative disorders. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.701

[35]AYKAT, Şükrü, & SENAN, S. (2023). Using Machine Learning to Detect Different Eye Diseases from OCT Images. *International Journal of Computational and Experimental Science and Engineering*, 9(2), 62–67. Retrieved from https://ijcesen.com/index.php/ijcesen/article/view/191

[36]Anakal, S., K. Krishna Prasad, Chandrashekhar Uppin, & M. Dileep Kumar. (2025). Diagnosis, visualisation and analysis of COVID-19 using Machine learning . International Journal of Computational and Experimental Science and Engineering, 11(1). https://doi.org/10.22399/ijcesen.826