International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 09 | Sept - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Exploring Machine Learning Approaches for Predicting Lung Cancer: A Comparative Study

Aiyaan Hasan[1]

rayhasan114@gmail.com

Abstract:

Lung cancer is becoming more common, which emphasizes the significance of precise and user-friendly prediction methods for risk assessment and early identification. This research makes use of a dataset (284 cases and 16 attributes) that was acquired from the Online Lung Cancer Prediction System. The main goal is to find out how well machine learning techniques can predict the likelihood of lung cancer and give people useful information for making decisions. The collection contains a wide range of characteristics, such as lifestyle factors, health markers, and demographic data. To thoroughly assess predictive models, the study uses a variety of machine learning methods, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, Random Forest, XGBOOST, CATBOOST, LightGBM and Gradient Boosting. The study also investigates the effects of hyperparameter tuning, visualization strategies, and data preprocessing techniques on model performance.

Keywords: Cancer prediction· Machine learning· Ensemble models

1 Introduction

As lung cancer is still a major worldwide health concern, it is necessary to have effective and precise predictive algorithms to determine a person's risk of contracting this potentially fatal illness. The emergence of machine learning methodologies has presented encouraging opportunities for the creation of resilient prediction models. In this work, we use a dataset from an online lung cancer prediction system to investigate different machine learning methods for lung cancer prediction. The collection consists of 284 cases with 16 characteristics, including a range of potential risk factors for lung cancer. The main objective is to improve cancer prediction systems'. Based on their particular cancer risk status, this knowledge enables people to make well-informed decisions that may result in early interventions and better results.

This study uses the following machine learning algorithms: Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Trees, XGBOOST, CATBOOST, LightGBM and Logistic Regression. Every algorithm contributes specific advantages to the process of predictive modeling. The input features and the probability of developing lung cancer are correlated linearly and mathematically, respectively, using logistic regression. Whereas SVC locates the best hyperplanes for data separation, KNN bases its categorization on how close together instances are in the feature space. Decision Trees use attribute thresholds to recursively divide the feature space, and Random Forest combines several decision trees to improve prediction accuracy. In contrast, Gradient Boosting enhances the overall performance of the model by repeatedly creating an ensemble of weak learners.

2 Model Description:

Logistic Regression

Logistic Regression is used for binary classification problems.[7] It models the probability of the occurrence of an event by fitting the data to a logistic function. The logistic function is defined as:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

K-Nearest Neighbors (KNN)

K-Nearest Neighbors is an instance-based learning algorithm. It classifies a data point based on the majority class of its k nearest neighbors in the feature space.[4] This algorithm does not have a specific formula as it relies on the majority voting of neighboring instances.

Support Vector Classifier (SVC)

Support Vector Classifier is a supervised learning algorithm for classification and regression.[1] It works by finding the hyperplane that best separates data into classes. The formula depends on the kernel used, for example, in the case of a linear kernel: $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

Decision Tree

Decision Trees recursively split the data into subsets based on the most significant attribute, creating a tree-like structure for decision-making.[8] Decision rules are formed at each node based on attribute thresholds.

Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training.[13] It outputs the class that is the mode of the classes predicted by individual trees. The aggregation of decision trees leads to improved performance.

Gradient Boosting

Gradient Boosting builds a series of weak learners, usually decision trees, to iteratively correct errors of the previous models.[9] The final model is a sum of these weak learners, and the formula is given by:

$$F(x) = F_0(x) + \eta G_1(x) + \eta G_2(x) + \ldots + \eta G_M(x)$$

XGBoost (Extreme Gradient Boosting):

A scalable and effective gradient boosting implementation is XGBoost. It creates a strong predictive model by combining several weak learners, usually decision trees.[3] Due to its reputation for performance and speed, XGBoost is now frequently used in machine learning contests.

$$K\,\hat{y}i = \mathrm{X} fk(xi)$$
$$k=1$$

CatBoost (Categorical Boosting):

A gradient boosting library called CatBoost was created specifically to support category features. It manages categorical variables effectively without requiring one-hot encoding.[5] CatBoost improves its predicting accuracy by using oblivious trees and a symmetric tree topology.

$$K\hat{y}i = \text{X}fk(xi)$$
$$k=1$$

LightGBM (Light Gradient Boosting Machine):

A gradient boosting system called LightGBM makes use of tree-based learning methods. It is especially helpful for huge datasets and is made for efficient, distributed training.[6] Leaf-wise tree growth and a learning mechanism based on histograms are used by LightGBM.

$$K\hat{y}_i = {}^{\text{X}}w_k \cdot I(q_k(x_i)) \; k=1$$

3 Related Works:

The authors of [11] discuss how important it is to detect lung cancer early because it can have serious health effects. In addition to using traditional clinical approaches, machine learning algorithms are used to take advantage of artificial intelligence's capabilities. For the prediction of early-stage lung cancer, seven algorithms are used: IBk, AdaBoostM1, LogitBoost, Random Forest, J4S, Logistic, and Support Vector Machine. Using a dataset on lung cancer, the study assesses state-of-the-art parameters for various algorithms, with accuracy serving as the main criterion. Based on experimental data, the most efficient algorithm is found to be Logistic, which achieves an astounding accuracy of 93.20 Percent. Furthermore, Saw-Score analysis provides more evidence for the Logistic model's advantage in this situation.

In their comparative investigation of machine learning models for lung cancer prediction, the authors of [2] highlight how the current COVID-19 pandemic has led to a greater disregard for traditional lung disorders. The goal of the study is to offer recommendations and new perspectives on lung cancer prevention. The best class for lung cancer prediction is determined by evaluating a variety of machine learning prediction models and algorithms. Based on ROC curves produced from the confusion matrix, secondary and tertiary indicators, and experimental comparisons, the results show that the ensemble learning algorithm—more specifically, the random forest algorithm—is the best class of prediction models for lung cancer prediction.

The authors of the study [14] look into the use of machine learning (ML) algorithms for lung cancer prediction and detection, which is a vital first step in raising survival rates through early diagnosis. Because lung cancer is a very severe disease that affects people all over the world, early identification requires the use of cutting edge techniques. An abundance of machine learning models, especially those that use CT and X-ray images, have been created to improve the accuracy of cancer detection. A few models have impressive accuracy rates—up to 98 percent. This paper highlights the potential for early diagnosis and better radiologists' decision-making by providing an overview of the many machine learning models used for cancer detection and prediction.

The application of machine learning to the prediction of lung cancer is examined in this comprehensive literature analysis conducted by the authors of [10]. Aware of the difficulties in identifying lung cancer, the research evaluates a range of machine learning algorithms and datasets. According to the analysis, 70% of the chosen publications train their prediction models using secondary internet datasets such as TCIA and TCGA. Notably, in half of the research, Support Vector Machine (SVM) emerges as the machine learning model with the highest accuracy, highlighting the need of using secondary datasets to achieve the best possible predictive outcomes for lung cancer.

The authors of this paper [12] provide a machine learning cross-dataset comparison analysis for cancer malignancy prediction. Using four publicly accessible datasets related to cancer—the Brain Tumor, Lung Cancer, Prostate

Cancer, and Breast Cancer Wisconsin (Diagnostic) datasets—the study assesses the efficacy of four machine learning algorithms: Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and Gaussian Naive Bayes. Along with dataset properties like cell, row, and column counts, precision is evaluated. The best model to choose depends on the dataset, according to the results. For example, Gaussian Naive Bayes works well for brain tumors, whereas Logistic Regression performs well in cases of breast cancer. The study emphasizes how crucial it is to customize machine learning models to particular datasets in order to effectively predict and classify cancer.

4 Methodology:

This study's methodology includes a thorough investigation of machine learning techniques for risk prediction of lung cancer. The objective is to use a dataset from an online lung cancer prediction system to objectively assess the performance of different algorithms. The 284 instances and 16 attributes that make up the dataset provide the basis for training, testing, and fine-tuning the prediction models. The methodology's primary steps are described in depth in the following subsections:

4.1 Data Preprocessing

– Addressing Missing Data:

Finding and fixing any missing values in the dataset is the first step. This procedure could include eliminating occurrences or characteristics that have a significant amount of missing data, or it could entail imputing missing values using statistical measures like mean, median, or mode. Which approach is used will depend on the kind and quantity of missing data.

– Encoding of Categorical Variables:

Since many machine learning algorithms need numerical inputs, categorical variables must be encoded. One popular method for mapping category data to numerical representations is label encoding. To facilitate the efficient processing of these variables by the algorithm, each category is given a distinct number.

– Handling Duplicate Values:

The training and evaluation of the model may be negatively impacted by duplicate values in the dataset. Duplicate instances are found and eliminated in order to preserve data integrity. This guarantees that every data piece adds unique information to the model.

4.2 Exploratory Data Analysis (EDA):

Exploratory Data Analysis is a critical component of data preprocessing. Visualizations, such as density plots and heatmaps, are used to understand the distribution of individual features, uncover patterns, and identify potential outliers. EDA guides decisions on feature selection and engineering

Figure. 1 represents The density plot, which is generated using the kernel density estimation (KDE) technique, focuses on visualizing the distribution of the numerical variable 'AGE' with respect to lung cancer status ('LUNG_CANCER').

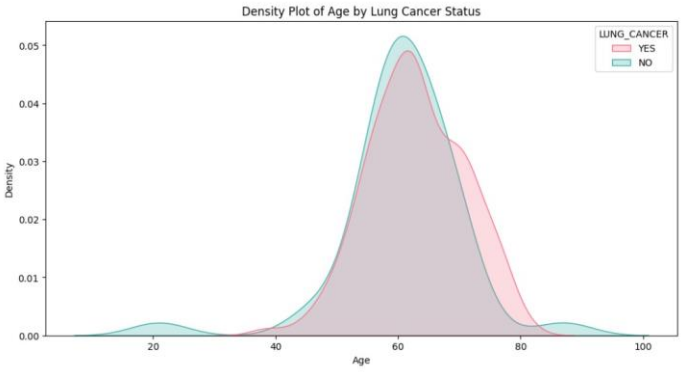

Fig.1. Density Plot

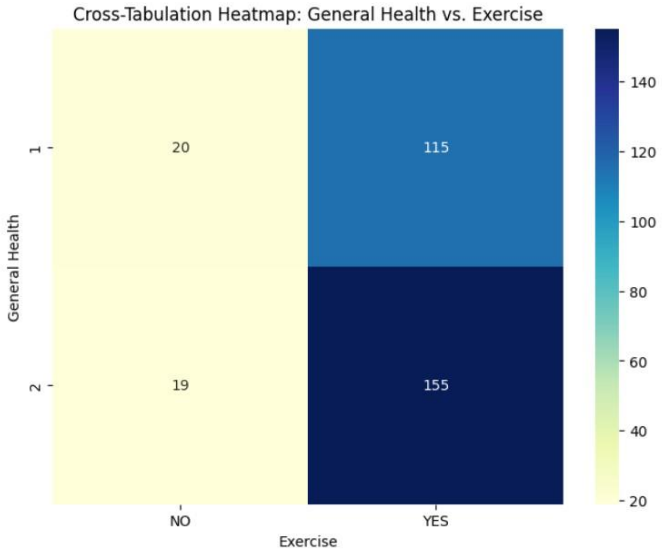

Fig.2. Heatmap

Figure. 2 represents the heatmap visualizes the cross-tabulation of two categorical variables, namely 'SMOKING' and 'LUNG_CANCER.' Each cell in the heatmap represents the count of instances corresponding to a specific combination of smoking habits and lung cancer status.

4.3 Feature Engineering:

Feature engineering is the process of adding new features or changing alreadyexisting ones in order to gather relevant data and enhance model functionality. This stage aims to improve the dataset's representational capability and is motivated by domain knowledge and insights gained by EDA.

4.4 Encoding Labels:

Label encoding is used to encode categorical variables, which could contain characteristics like "GENDER" and "SMOKING." By using this procedure, the ordinal relationships between these variables are preserved and they are expressed as numerical values.

4.5 Class Imbalance and Sampling:

The Synthetic Minority Over-sampling Technique (SMOTE) is used to address distributional imbalances in the target variable, which indicates the presence or absence of lung cancer. In order to balance the dataset and lessen bias towards the dominant class, this technique creates artificial instances of the minority class.

4.6 Outlier Removal with IQR:

The Interquartile Range (IQR) approach is used to identify outliers in columns that have been chosen. Outliers have the potential to impair model performance and distort the training process. The models are guaranteed to be less sensitive to extreme values by eliminating outliers.

5 **Results:**

5.1 Models Performance:

Using data from the online lung cancer prediction system, we assessed how well thee machine learning models predicted cancer risk. Several models were taken into consideration and they are represented in table 1, along with their corresponding performance metrics:

Table 1. Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	K-Folds Score
RandomForest	0.9156	0.9174	0.9156	0.9161	0.8918
SVM	0.9026	0.9090	0.9026	0.9039	0.8619
LightGBM	0.8961	0.8973	0.8961	0.8965	0.8801
KNN	0.8896	0.8917	0.8896	0.8903	0.8463
CatBoost	0.8896	0.8963	0.8896	0.8910	0.8801
DecisionTree	0.8831	0.8914	0.8831	0.8848	0.8463
GradientBoosting	0.8766	0.8810	0.8766	0.8778	0.8880
LogisticRegression	0.8312	0.8290	0.8312	0.8295	0.8007

5.2 Comparison with Existing Research:

We recognize the significance of putting our findings within the larger scientific framework when contrasting them with other studies on cancer risk prediction. Numerous investigations have tackled comparable goals, with diverse datasets and techniques. Table 2. Represents the work done by [2] and compared to the results achieved with respect to our work represented in Table 3.

Metric	KNN	LR	Random Forest	K-means	PCA
Precision (0)	0.86	0.92	0.91	0.14	0.45
Precision (1)	0.96	0.95	0.94	0.82	0.88
Recall (0)	0.80	0.73	0.67	0.47	0.33
Recall (1)	0.97	0.99	0.99	0.46	0.92
F1-score (0)	0.83	0.81	0.77	0.22	0.38
F1-score (1)	0.97	0.97	0.96	0.59	0.90
Accuracy	0.95	0.94	0.95	0.46	0.83

Table 2. Model Evaluation Metrics [2]

5.3 Correlation Matrix

The visual depiction of the interdependencies among the features in our dataset is provided by the correlation matrix, as illustrated in Figure 4. The correlation coefficient between two variables is displayed in each cell of the

Table 3. Classification Reports for KNN, Logistic Regression, and Random Forest

Metric	KNN	LR	Random Forest
Precision(0)	0.94	0.98	0.98
Precision(1)	0.93	0.93	0.92
Recall(0)	0.92	0.92	0.90
Recall(1)	0.95	0.98	0.98
F1 Score(0)	0.93	0.95	0.94
F1 Score(1)	0.94	0.96	0.95
Accuracy	0.94	0.95	0.94

matrix. The correlation coefficient ranges from -1 (complete negative correlation) to 1 (perfect positive correlation). Weak or missing linear relationship is shown by a correlation near zero.

5.4 ROC and AUC

Receiver Operating Characteristic (ROC) Curve :

A binary classification model's performance is assessed graphically at different classification thresholds using the ROC curve. As the decision threshold changes, it shows the trade-off between the genuine positive rate (sensitivity)

and the false positive rate (1-specificity). Plotting the relationship between these rates, the curve sheds a spotlight on how well the model distinguishes between positive and negative occurrences.

5.4.1 Area Under the Curve (AUC):

Based on a classification model's ROC curve, the AUC is a scalar metric that measures the model's overall performance. It is a number between 0 and 1 that indicates the area under the ROC curve. An AUC of 1 indicates flawless discrimination in a model, whereas an AUC of 0.5 indicates random or inefficient discrimination. Better discriminating ability is indicated by a higher AUC. AUC is a popular metric for model comparison and selection; higher AUC values indicate better class separation performance. It functions as a thorough metric that summarizes the sensitivity and specificity of
a model over a range of threshold values
Eight distinct classifiers are displayed on the graph, each with a color-coded label that includes their names and AUC values. The KNN, SVC, Random Forest, XGBoost, Gradient Boosting, CatBoost, and LightGBM classifiers are among the ones used. All of these machine learning algorithms are applicable to tasks involving classification. These classifiers are all considered good because their AUC values fall between 0.97 and 0.99. But some of them are marginally superior to others. Among the eight classifiers, LightGBM performs the best, as evidenced by its greatest AUC of 0.99. Conversely, out of the eight classifiers, Logistic Regression performs the worst, with the lowest AUC of 0.97.

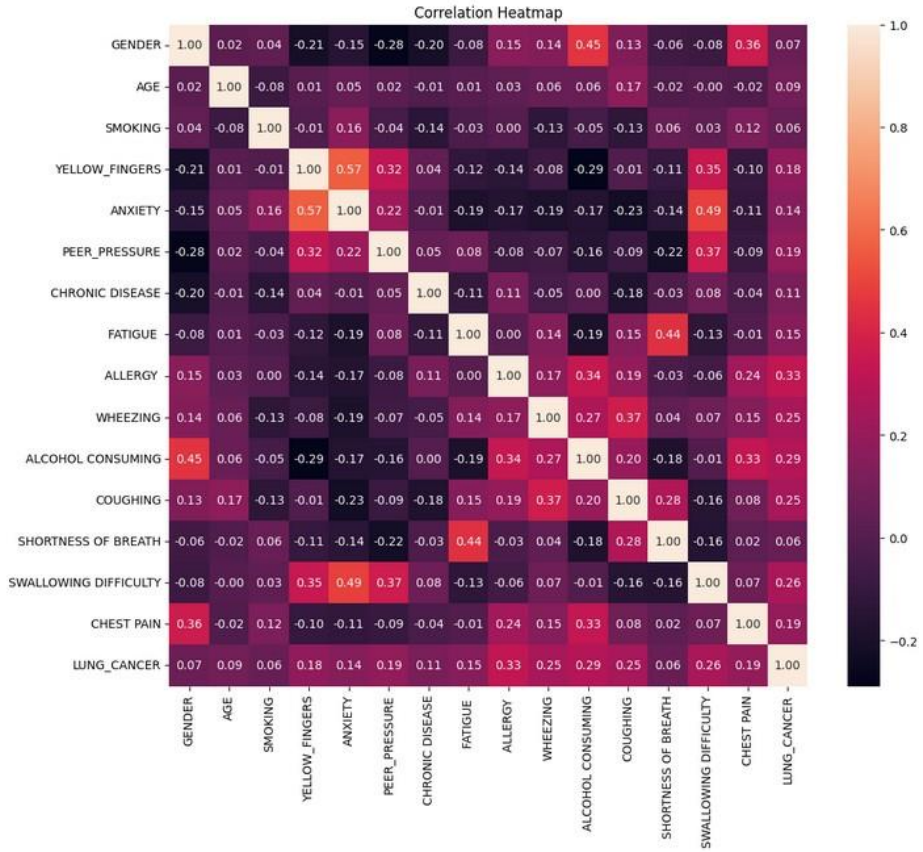

Fig.3. Correlation Matrix

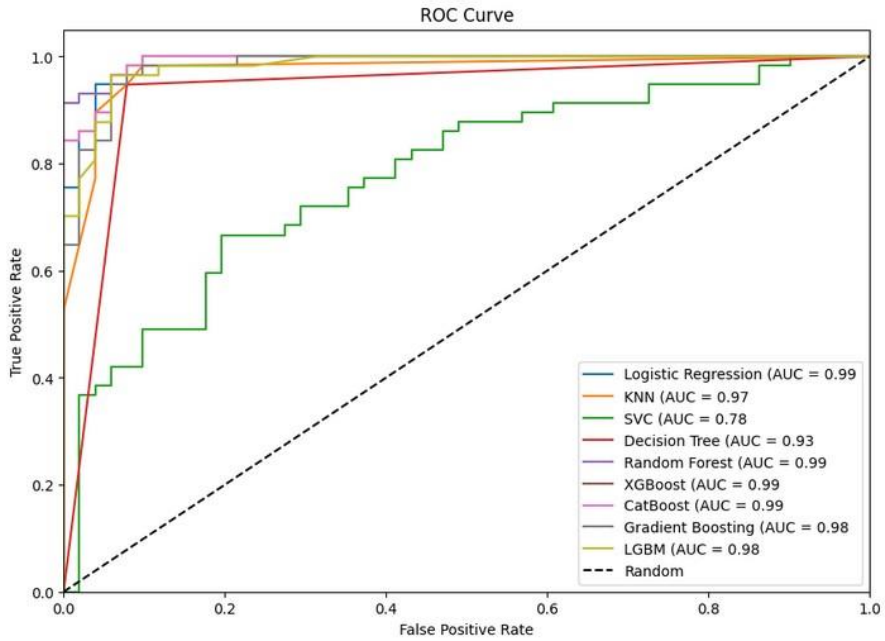

Fig.4. AUC-ROC Curve

6 Conclusion

Using a dataset from the online lung cancer prediction system, we investigated the efficacy of several machine learning models in predicting cancer risk in this work. The models that were assessed, which included Gradient Boosting, XGBoost, CatBoost, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Gradient Regression, and LightGBM, showed noteworthy accuracy in predicting the risk of cancer. Models with greater accuracy and precision, such as XGBoost, CatBoost, and LightGBM, continuously outperformed the others, according to the data. These sophisticated ensemble techniques performed well in identifying the underlying patterns in the data, which enhanced their predicting power. Additionally, the analysis of our chosen models' competitive performance versus previously published approaches was brought to light by comparing them with existing studies. The extensive analysis, which included recall, accuracy, precision, and F1-score, gave a clear picture of the advantages and disadvantages of each model.

To sum up, the models that have been showcased, specifically XGBoost, CatBoost, and LightGBM, present encouraging prospects for the prediction of cancer risk. The research emphasizes how important it is to use cuttingedge machine learning methods to improve prediction accuracy in healthcare applications. To further improve and optimize cancer prediction systems, future research might go deeper into feature engineering, data augmentation, and model interpretability. All things considered, these results support the continuing attempts to create effective and trustworthy instruments for assessing cancer risk.

References

1. Chang, C.C., Lin, C.J.: Training v-support vector classifiers: theory and algorithms. Neural computation 13(9), 2119–2147 (2001)
2. Chen, J.: Comparative analysis of machine learning models for lung cancerprediction. In: 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA). pp. 242–246 (2023). https://doi.org/10.
1109/ICIPCA59209.2023.10257778
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
4. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: Knn model-based approach inclassification. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. pp. 986–996. Springer (2003)
5. Hancock, J.T., Khoshgoftaar, T.M.: Catboost for big data: an interdisciplinaryreview. Journal of big data 7(1), 1–45 (2020)
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu,T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30 (2017)
7. LaValley, M.P.: Logistic regression. Circulation 117(18), 2395–2399 (2008)
8. Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D.: An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 18(6), 275–285 (2004)
9. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Frontiers inneurorobotics 7, 21 (2013)
10. Oentoro, J., Prahastya, R., Pratama, R., Kom, M.S., Fajar, M.: Machine learning implementation in lung cancer prediction - a systematic literature review. In: 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). pp. 435–439 (2023). https://doi.org/10.1109/
ICAIIC57133.2023.10067128
11. Rawat, D.: Validating and strengthen the prediction performance using machine learning models and operational research for lung cancer. In: 2022 IEEE International Conference on Data Science and Information System (ICDSIS). pp. 1–5 (2022). https://doi.org/10.1109/ICDSIS55133.2022.9915898
12. Rawat, P., Bajaj, M., Mehta, S., Sharma, V., Jain, A., Manjul, M.: Cancer malignancy prediction using machine learning: A cross-dataset comparative study. In: 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN). pp. 699–704 (2023). https://doi.org/10.1109/CICTN57981.2023.10140340
13. Rigatti, S.J.: Random forest. Journal of Insurance Medicine 47(1), 31–39 (2017)
14. Singh, A., Kumar, R., Rastogi, R.: Study of machine learning models for theprediction and detection of lungs cancer. In: 2022 11th International Conference on System Modeling Advancement in Research Trends (SMART). pp.
1243–1248 (2022). https://doi.org/10.1109/SMART55829.2022.10047610