Utilizing Machine Learning for Predicting Lung Cancer Outcomes

Dr. A.V.Sriharsha
Professor
Dept. of Data Science
Mohan Babu University, Tirupati

Mr.C. Syamsundar Reddy
Sub-Inspector of Police
O/o. Intelligence Department @ State Hqrs
Vijayawada, AndhraPradesh - 520008 (India)

Abstract

Lung cancer, a severe health condition often triggered by prolonged smoking habits and genetic factors, presents significant treatment challenges. Early diagnosis can significantly improve survival rates, yet many cases are only identified at advanced stages. Recent advancements in machine learning have led to promising pre-diagnosis tools that utilize these technologies. Machine learning allows computers to learn from and interpret data to perform specific tasks effectively. This review outlines various machine learning models designed for predicting the likelihood of lung cancer using datasets derived from genetic, clinical, and histological backgrounds. Each model employs distinct algorithms, demonstrating high performance in prediction accuracy. The findings indicate that these models could be highly beneficial in the early screening processes for lung cancer.

Keywords: machine learning, lung cancer, prediction.

1. Introduction

Machine learning is a powerful tool for medical diagnostics, enabling computers to analyze data and perform specific tasks. Researchers identify key features for lung cancer prediction and gather real-world data to train models. These models are intended to find relationships and patterns in data that human analysts might not notice right away. Each model's performance is assessed following training to ascertain its dependability and accuracy in predicting lung cancer. This approach aims to refine lung cancer screening, making it more accessible and accurate for those at risk.

Because of their accuracy and consistency, random forests are frequently used to categorize patients with lung cancer. For instance, a study demonstrated that a random forest model achieved an impressive accuracy of 98.507% in diagnosing lung cancer based on critical factors Murad et al. Another comparative analysis confirmed that random forests were the optimal prediction model among various machine learning algorithms Chen. Decision trees have also proven to be highly effective. In one study, they achieved a perfect accuracy rate of 100% in predicting lung cancer outcomes Mishra and Gangwar. Ensemble techniques like XGBoost have shown superior predictive performance. In comparison with other algorithms, Singh et al.'s XGBoost model produced an accuracy score of 96.76% in diagnosed lung cancer. Furthermore, Lynch et al. found that a customised ensemble technique with a root mean square error (RMSE) of 15.05 performed better than other approaches when applied to the SEER database. Early diagnosis has also benefited from deep learning advances [17][19]. Building deep neural network models has been crucial for leveraging high-throughput sequencing data, thereby enhancing the effectiveness of diagnosis Ruban et al. CNN-based models have been validated on independent datasets for predicting lung cancer using imaging data. These models showed good predictive accuracy with an AUC of 0.92, as reported by Peschl et al. Survival rates could be considerably raised by incorporating machine learning into the prediction of lung cancer. For example, detecting lung cancer at an early stage can increase the survival rate from 17.7% to 54.4% Joshua et al.. This underscores the critical role of ML in facilitating early and accurate diagnosis, which is paramount in enhancing patient outcomes. Most studies incorporate a range of clinical parameters, such as age, sex, and cancer staging, along with imaging data and genetic information. The inclusion of patient-reported outcome measurements has also shown to improve predictive performance, highlighting the importance of comprehensive data integration in ML models Sim and Yun.

2. Background

Lung cancer, with a five-year survival rate of only 18.6%, is the second most common type of cancer worldwide. Lung cancer is expected to claim the lives of 127,000 people in the United States alone in 2023, in addition to 238,000 new cases. The disease's risk is increased by a number of factors, including radiation, air pollution, smoking, and secondhand smoke exposure. Frequently, lung cancer does not present symptoms in its early stages, which delays diagnosis until it is significantly advanced. As such, effective early detection methods are crucial.

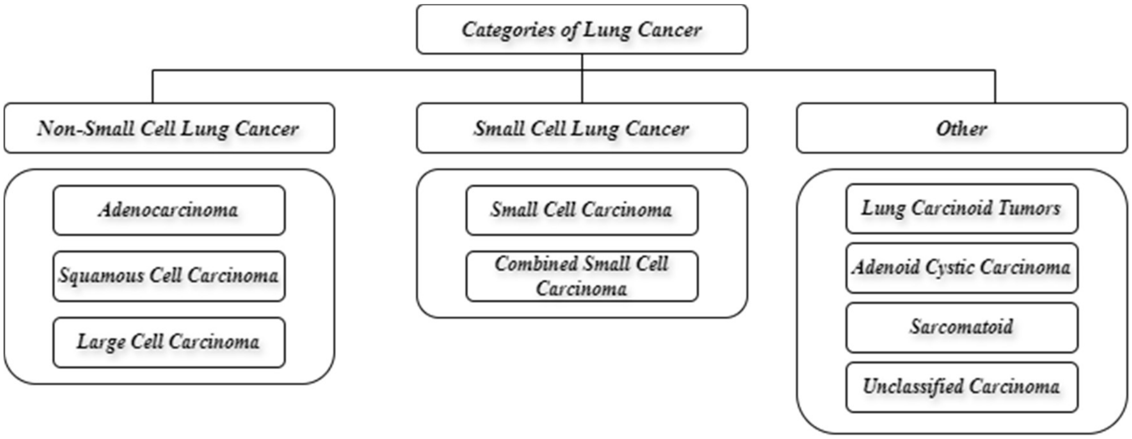


Fig.1: Lung Cancer Categories

Low Dose Computerized Tomography (LDCT) is the primary method for early detection of lung cancer. This method of medical imaging takes several X-ray pictures from various viewpoints, which a computer uses to produce fine-grained cross-sectional views of the body. Despite its widespread use, LDCT struggles with high rates of overdiagnosis and false positives.

Overdiagnosis occurs when non-threatening growths are mistakenly classified as cancerous, leading to unnecessary and potentially harmful treatments. Additionally, with only 2.6 doctors per 1,000 people in the U.S., not everyone may have access to LDCT screenings.

According to statistics from the American Lung Association, in 2021, only a small fraction (5.8%) of the eight million Americans identified as high risk for lung cancer underwent LDCT screenings. This highlights a significant need for more effective and economical methods to pre-diagnose lung cancer on a broader scale. In response, many researchers have been exploring the potential of machine learning algorithms for early diagnosis.

Predictive models for lung cancer outcomes are essential for improving early detection, prognosis, and treatment strategies. These models utilize various factors, including clinical data, genetic markers, and patient behaviors, to estimate the risk and progression of lung cancer. One significant study developed a multivariable lung cancer risk prediction model incorporating low-dose computed tomography (CT) screening results. This model, validated using data from the National Lung Screening Trial, demonstrated superior predictive capabilities compared to models excluding screening results Tammemägi et al. Another study focused on developing the Lung Cancer Prognostic Index (LCPI) using gene expression datasets. The LCPI successfully differentiated patients into risk groups and predicted survival probabilities up to 15 years post-surgery, highlighting its utility in long-term prognosis. Machine learning (ML) techniques have been pivotal in advancing lung cancer prediction. Logistic regression models achieved a high accuracy rate of 95% for classifying patients with lung cancer based on various clinical parameters Ojha and Maharajan. Furthermore, random forest and logistic regression models have been effectively used for risk analysis, incorporating symptomatic and behavioral features Ding et al. Integrating metabolic markers with epidemiological data can enhance predictive models. For instance, a risk prediction model developed and validated in a Chinese population incorporated such markers to address false-positive rates and overdiagnosis in low-dose CT screening, thereby improving lung cancer mortality predictions Lyu et al. Including patient-reported outcome (PRO) measurements with clinical variables has improved the predictive performance of survival prediction models. This approach enhances the prediction of 5-year disease-free lung cancer survival, underscoring the value of patient-centric data in predictive modeling Sim and Yun.

Radiomics features, which are quantitative measures extracted from medical imaging, have shown significant potential in predicting lung cancer outcomes. For example, changes in radiomics features during radiation therapy can serve as indicators of tumor response, providing valuable insights for treatment adjustments Fave et al. Clinical stage at diagnosis remains a critical factor affecting lung cancer prognosis. Earlier stages correlate with higher survival rates, emphasizing the need for early detection and accurate staging Yong. Additionally, a study highlighted that even minimal lung function damage, indicated by forced expiratory volume, significantly increases lung cancer risk Calabrò et al.

This study offers a detailed examination of various machine learning models designed for the pre-diagnosis of lung cancer. It particularly concentrates on models that are informed by datasets encompassing three critical aspects of lung cancer: genetic predispositions, clinical data, and histological findings. These elements are closely linked to the incidence of lung cancer, providing a rich source of information that significantly enhances the predictive accuracy of these models. By integrating these diverse datasets, the machine learning models can analyze patterns and make highly reliable predictions regarding the likelihood of lung cancer, thereby facilitating earlier and more effective interventions.

3. Related Work

Advances in Artificial Intelligence including ML and DL were highlighted and persistent studied in the recent past, about the use of gene expression data in lung-cancer detection. These studies emphasize the use of novel architectures, early detection, interpretability of the models, integration of multiple data types, and semi-supervised approaches.

Using gene expression data, Md. Rezaul Karim et al. [1] present a method to increase the elucidate and predict various types of cancers. From Pan-Cancer Atlas researchers demonstrates that rigorous training of CNN and VGG16 networks were employed on gene expression data, which consists of 9,074 cancer patients—representing almost 33 distinct types of cancers. The method achieved an average precision of 96.25% in predicting cancer types. Using cancer transcriptomes available in the Pan-Cancer Atlas project, a thorough model for classifying cancers is presented. Important genes are identified and visualized using GradCAM++ in conjunction with CNN and VGG16 networks. The study also identifies feature importance and biomarkers using SHAP and gradient boosted trees. OncoNetExplainer outperforms existing methods in both pre-model and post-model interpretability, enhancing predictions' reliability and transparency. Future directions include combining data from other genomics projects and integrating multimodal data.

Suli Liu and Wu Yao et al. [2] postulated novel approaches in deep learning in the early diagnoses of lung-cancer. The authors projects issues related to high-dimensionality and imbalanced nature of data amongst gene expression datasets, where their proportion is relevant to mortality rate and lung cancer also with the swift advancement of high-throughput sequencing technology. They propose a method of gene selection to identify highly associated genes with lung-cancer. The DNN model proposed employs pivotal loss to handle imbalanced data and k-fold cross-validation to ensure robustness. The proposed model, trained on RNA-seq data from the TCGA and ICGC datasets, demonstrates high accuracy with an AUC of 0.99, significantly outperforming traditional machine learning methods like SVM, LR, KNN, and RF. The study highlights the effectiveness of deep learning in capturing complex non-linear relationships in gene expression data, thus providing a powerful tool for the early and accurate prediction of lung cancer. The use of KL divergence for gene selection and the implementation of a deep neural network with focal loss contribute to the improved performance and generalization of the model, offering a promising direction for future research in cancer diagnostics.

Through the identification and curation of 26 clinically annotated gene expression datasets from 2786 cancer cases across various cancers, Borisov et al. made a significant contribution to personalized oncology. The study aimed to build machine learning (ML) models to predict chemotherapy responses, addressing the challenge of limited clinically annotated molecular profiles. The datasets were evaluated for ML applicability and robustness using leave-one-out cross-validation, with 23 datasets deemed suitable. The study underscored the importance of robust training and validation datasets for developing ML-assisted diagnostic tools and proposed data harmonization and transfer learning as potential solutions. The authors believe this collection of datasets will be valuable for ML applications in oncology and the development of next-generation cancer biomarkers.

Rukhsar et al. [4] in ther article contributes significantly to cancer classification by utilizing deep learning and RNA-Seq gene expression data. The authors analyze RNA-Seq data from five different cancer types from the Mendeley data repository and convert it into 2D images using normalization and zero padding. Relevant features are extracted and selected using various deep learning algorithms, including CNN, ResNet50, ResNet101, ResNet152, VGG16, VGG19, GoogleNet, and AlexNet. The study evaluates the performance of these models using four data splitting strategies and k-fold cross-validation. CNN achieves the highest accuracy, outperforming existing models by 97%. This study highlights the effectiveness of deep learning for automatic feature extraction and cancer classification from RNA-Seq data, highlighting the potential of these methods for improving cancer diagnostics accuracy and reliability.

Bijaya Kumar Sethi et al. [5] in their article presents a deep learning approach for prostate cancer detection and classification. The authors use the GEDAAI-PCD technique, which normalizes gene expression data and uses a hybrid Long Short-Term Memory-Deep Belief Network (LSTM-DBN) model for classification. The Enhanced Wild Horse Optimization (EWHO) system is used for hyperparameter tuning. The study demonstrates high classification accuracy and robustness against high-dimensional data and class imbalances. The experimental results show significant improvements in accuracy, precision, sensitivity, and specificity compared to traditional methods. The paper emphasizes the need for larger, diverse datasets and ethical considerations in healthcare AI applications.

In the works by Tanima Thakur et al. [6] a novel hybrid deep learning model that makes use of gene expression data to predict various types of cancers using Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) is proposed. The authors use a sandwich stacked bottleneck feature extraction technique, leveraging the strengths of VGG16 and VGG19 pre-trained models to identify crucial patterns in gene expression data. The hybrid RNN-CNN classifier classifies cancer types more accurately than existing methods like VGG16, VGG19, ResNet50, Inception V3, and MobileNet. The model achieved the highest accuracy of 0.978 for one dataset and 0.994 for another, outperforming other classifiers in terms of accuracy, mean square error, precision, recall, and F1 score.

Machine learning involves training computers to analyze data and carry out specific tasks, holding great promise for medical diagnostics [7][8]. To develop a robust machine learning model, researchers first identify the key features that are most predictive of lung cancer. They then gather relevant real-world data which serves as the training material for these models [9]. Typically, several models are tested to establish a baseline for comparison, allowing researchers to determine which model performs best.

These models are designed to discern patterns and relationships within the data that may not be immediately apparent to human analysts. After a period of training, the performance of each model is rigorously evaluated. Researchers assess the accuracy and reliability of the models in predicting lung cancer, selecting the most effective one for further study and development. This approach aims to refine lung cancer screening and make it more accessible and accurate for those at risk.

4. Methodology

This study offers a detailed examination of various machine learning models designed for the pre-diagnosis of lung cancer. It particularly concentrates on models that are informed by datasets encompassing three critical aspects of lung cancer: genetic predispositions, clinical data, and histological findings. These elements are closely linked to the incidence of lung cancer, providing a rich source of information that significantly enhances the predictive accuracy of these models. By integrating these diverse datasets, the machine learning models can analyze patterns and make highly reliable predictions regarding the likelihood of lung cancer, thereby facilitating earlier and more effective interventions.

4.1 Factors related to lung cancer and machine learning tools for prediction

This section explores the factors associated with lung cancer and the use of machine learning tools to predict it. Key factors include gene expression, which influences the likelihood of lung cancer development, and clinical data, which includes age, gender, body metrics, and blood analyses. Smoking status is a crucial piece of clinical data in lung cancer research due to the carcinogens in tobacco smoke. Hesology, the study of microscopic structures of tissues and cells, is also essential for predicting potential diseases by analyzing cellular compositions and abnormalities.

4.2 Common Machine Learning Algorithms for Prediction

The list of machine learning algorithms for lung cancer prediction includes multi-layer perceptron, random subspace, decision trees, random forest, Support Vector Machine (SVM), K-means clustering, Convolutional Neural Networks (CNNs), linear models, K-nearest Neighbors (KNN), Extreme Gradient Boosting, and Logic Learning Machine. Each algorithm has its strengths and is chosen based on the specific requirements and complexities of the data involved. Multi-layer Perceptron recognizes intricate patterns, while Random Subspace constructs classification models using random subsets of features. Decision Trees are tree-structured models that segment data into subsets, while Random Forest outputs predictions through a collective voting system. Support Vector Machine (SVM) identifies optimal hyperplanes for classifying data, particularly effective with high-dimensional data. K-means clustering efficiently handles large datasets. CNNs are primarily used in image analysis and are highly prevalent in medical imaging. Linear models establish linear relationships between features but are limited to simple, non-complex patterns. Extreme Gradient Boosting integrates multiple simpler models, managing missing data commonly found in real datasets.

4.3. Application of Models that are trained on different settings

4.3.1 Gene expression data as the training set

Many Researchers pioneered the development of machine learning models that utilized micro-array gene data to assess the likelihood of lung cancer. The models employed algorithms like Random Subspace, Sequential Minimal Optimization (SMO), and Multilayer Perceptron (MLP) for training. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) were the four categories into which the predictive results were divided. A recall metric, determined by the formula Recall = TP/(TP+FN), was used to assess the efficacy of these models. Among the algorithms, SMO demonstrated superior performance with a recall of 0.9029, indicating it correctly identified 90.29% of lung cancer cases.

Nicolas Coudray et al. [10] and his team used a Convolutional Neural Network (CNN) to predict gene mutations from image data using the NCI Genomic Data Commons dataset. The model was found to predict six out of ten common gene mutations with the values of AUC in the band of 0.733 and 0.856, aiding pathologists in more effective lung tumor assessment. The AUC value indicates the model's accuracy.

In 2017, Alicia Hulbert [11] and her team developed three Random Forest models to improve lung cancer diagnosis by detecting DNA methylation in sputum and plasma samples. The models showed moderate predictive power, but showed significant improvement in diagnostic accuracy when using sputum and plasma data. This demonstrates how machine learning can be used to enhance biosample analysis-based lung cancer detection and diagnosis.

To estimate the risk of lung cancer in 2022, Abhishek Choudhary [12] and his colleagues employed machine learning models such as K-Nearest Neighbour, Random Forest, Support Vector Machine, and Linear Model. They developed a comprehensive model that integrated all four algorithms, trained on data from five single nucleotide polymorphisms in the DNA repair gene XRCC1, and smoking status. The ensemble model, which combined all four algorithms, achieved an AUC value of 0.93, demonstrating its robustness in accurately predicting lung cancer probability. In 2019, Masih Sherafatian and Fateme Arjmand [13] used a decision tree algorithm to classify lung cancer status using miRNA data. Their model showed promising performance with an AUC value of 91.2%, demonstrating the potential of decision trees in medical diagnostics. This approach provides a viable tool for accurately classifying lung cancer, highlighting the potential of decision trees in medical diagnostics.

4.3.2 Clinical data as the training set

In 2023, Guan [14] and his team embarked on developing an advanced predictive tool using the extreme gradient boosting (XGBoost) model to forecast the likelihood of lung cancer. This research initiative, conducted under the auspices of Kaiser Permanente Southern California, involved a comprehensive dataset comprising over 200,000 data points with 834 distinct features, including body mass index (BMI), smoking history, and spirometric evaluations. After rigorous training and validation, the XGBoost model achieved an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.856. This impressive AUC value underscores the model's effectiveness in facilitating the early pre-diagnosis of lung cancer.

In 2022, Ruiyuan Yang [16] and his team leveraged machine learning to predict epidermal growth factor receptor (EGFR) mutations, a key biomarker in lung cancer prognosis. This study utilized an array of clinical data, including various blood markers, at West China Hospital to train the predictive models. The random forest [14][15] and XGBoost models stood out among the algorithms tested, each achieving AUC values of 0.825 and 0.826, respectively. These results demonstrate the potent capability of machine learning tools in accurately predicting EGFR mutations, thereby enhancing the diagnosis and treatment strategies for lung cancer.

5. Results and Discussions

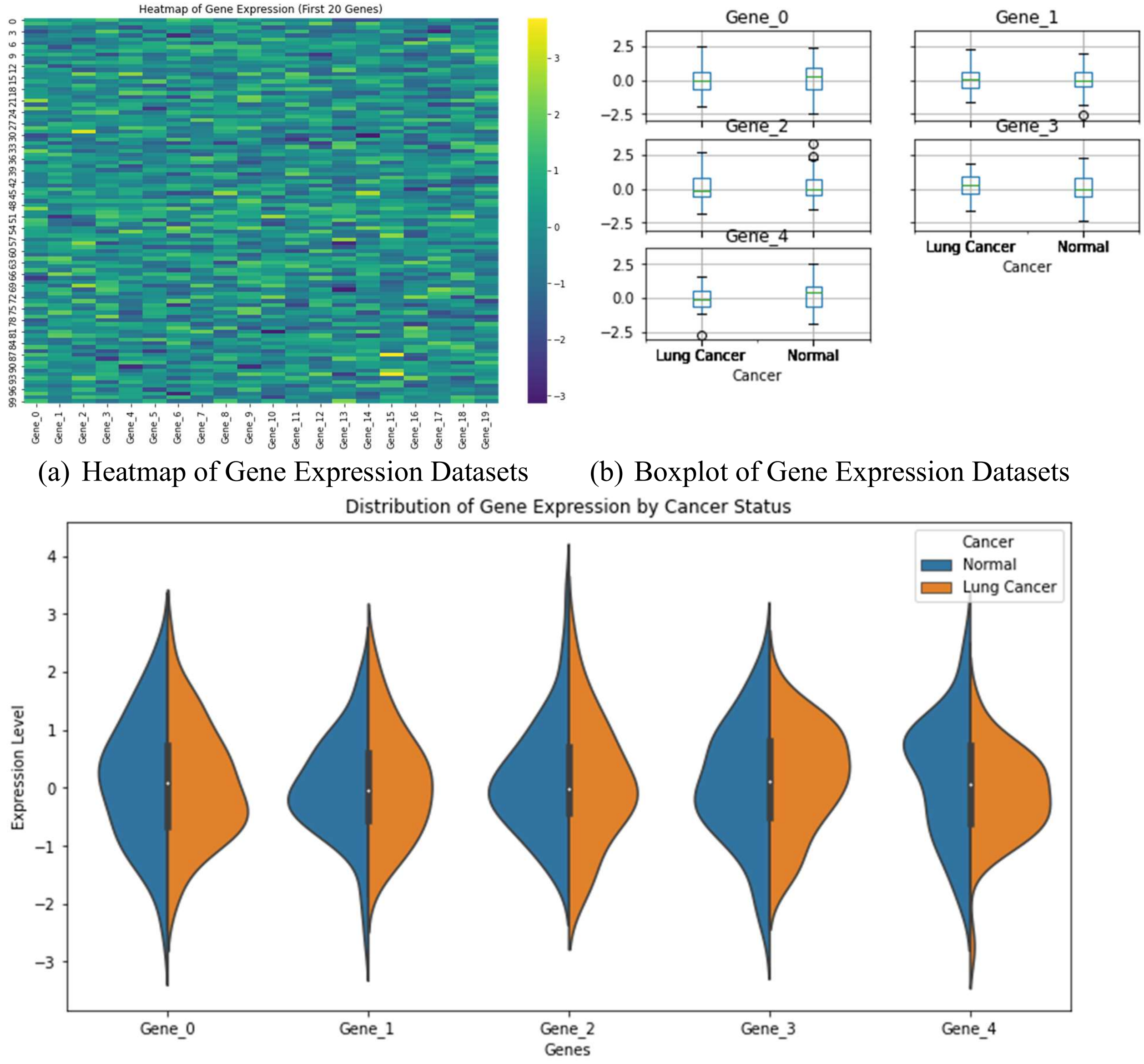
Gene expression datasets are essential for developing and testing deep learning and machine learning models for lung cancer detection. Commonly used datasets include the Gene Expression Omnibus (GEO), ArrayExpress, cBioPortal Cancer Genome Atlas (TCGA),, Lung Cancer Explorer, COSMIC (Catalogue of Somatic Mutations in Cancer), International Cancer Genome Consortium (ICGC), GTEx (Genotype-Tissue Expression) Portal, Cancer Cell Line Encyclopedia (CCLE), and Human Protein Atlas. These datasets provide a comprehensive collection of cancer genomic data, including RNA-seq data for various lung cancer subtypes. They are also useful for in vitro studies and model development. The International Cancer Genome Consortium (ICGC) coordinates large-scale cancer genome studies worldwide, including lung cancer datasets. GTEx provides gene expression data from various human tissues, including non-cancerous lung tissue, useful for comparative studies.

Table 1. Datasets - different objectives of possessing Gene Expression data for Cancer outcome prediction.

Dataset	Nature	Web Link
TCGA	Comprehensive Collection of Cancer Genomic data	https://portal.gdc.cancer.gov/
GEO	High-throughput Gene Expression data	https://www.ncbi.nlm.nih.gov/geo/
ArrayExpress	Functional Genomics data	https://www.ebi.ac.uk/arrayexpress/
cBioPortal	Heterogenous Cancer Genomic Data	https://www.cbioportal.org/
Lung Cancer Explorer	Specialized Lung Cancer Genomic dat	http://lce.biohpc.swmed.edu/lungcancer/
COSMIC	Datasets of Lung Specific-Somatic Mutations on Cancer	https://cancer.sanger.ac.uk/cosmic
ICGC	Large Scale Cancer Genomic studies	https://dcc.icgc.org/
GTEx	Cancer Gene Expression data from various human tissues	https://gtexportal.org/
CCLE	Gene Expression data from cancer cell lines	https://portals.broadinstitute.org/ccle
HPA	Protein level Cancer Gene Expression data	https://www.proteinatlas.org/

The Kaplan test, also known as the Kaplan-Meier estimator or Kaplan-Meier curve, is a statistical method used in survival analysis, particularly in cancer research. It estimates survival function from lifetime data and is used to analyze the time until an event occurs, such as death, recurrence, or progression of the disease. The Kaplan-Meier estimator calculates the probability of survival past certain time points and can handle censored data. It is used in comparisons between treatment groups, estimating median survival time, assessing the effectiveness of new cancer therapies, and analyzing progression-free survival in clinical trials. Results are typically presented as a step function graph known as a Kaplan-Meier curve, with the y-axis representing the estimated survival probability and the x-axis representing time. The Kaplan-Meier curve is often paired with statistical tests like the log-rank test to determine if differences between survival curves are statistically significant. It can also be used in cancer detection to compare survival rates between early and late detection groups, analyze the impact of screening programs on survival outcomes, and evaluate the effectiveness of different detection methods in terms of long-term survival.

Various types of visual plots are used for gene expression analysis, viz., heatmaps, box plots, violin plots, PCA plots, correlation plots, volcano plots, interactive scatter plots, and clustering heatmaps. Heatmaps are essential in Principal Component Analysis (PCA) for visualizing the correlation matrix and loading scores of principal components. They help identify patterns and potential multicollinearity among variables by highlighting pairwise correlation coefficients. These plots help visualize gene expression levels across samples, compare cancer vs. normal samples, and display the probability density of the data. To use these plots with real lung cancer gene expression data, replace the sample data generation with the actual dataset loading, adjust column names and data structures, and preprocess the data. Consider the biological significance of the genes you're focusing on, particularly for boxplots and violin plots. These visualizations can help identify patterns in gene expression data, detect differentially expressed genes between cancer and normal samples, visualize the overall structure of the dataset, and identify potential biomarkers for lung cancer. Interpret these plots in the context of your research question and consult with domain experts in lung cancer biology.



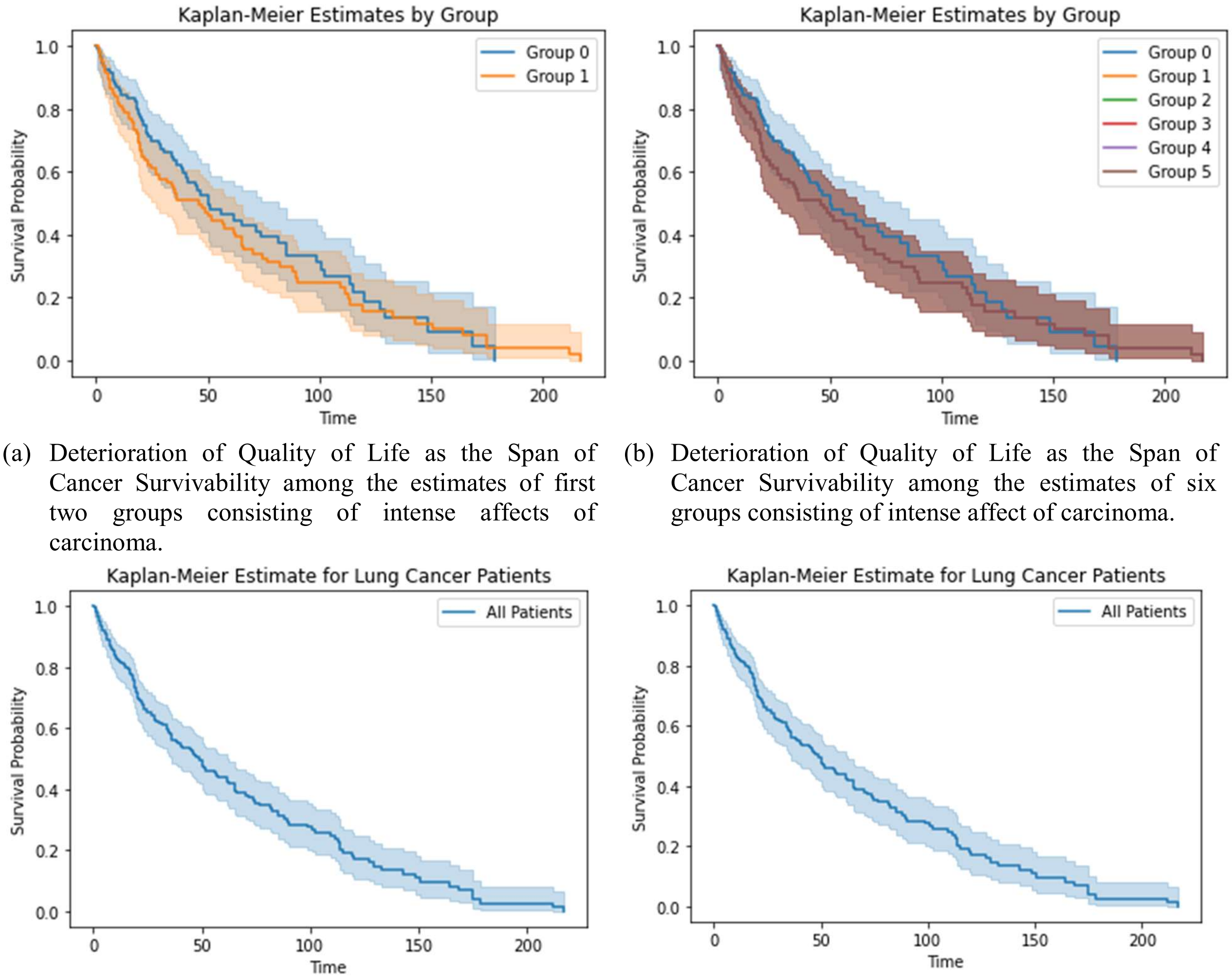
(a) Heatmap of Gene Expression Datasets

(b) Boxplot of Gene Expression Datasets

(c) Violin-plot of Gene Expression Datasets

Fig. 2 – Characteristics of Lung Cancer Gene Expression Datasets

The survival probability has been assessed among 6 groups of samples, where the 6 groups of samples contain the combination of 2 subgroups each, totally containing 12 subsamples. An estimate of survival probability is plotted for each graph pair of groups.



(a) Deterioration of Quality of Life as the Span of Cancer Survivability among the estimates of first two groups consisting of intense affects of carcinoma.

(b) Deterioration of Quality of Life as the Span of Cancer Survivability among the estimates of six groups consisting of intense affect of carcinoma.

(c) Deterioration of Quality of Life as the Span of Cancer Survivability among the estimates of lung cancer datasets in first two groups consisting of intense affects of carcinoma.

(d) Deterioration of Quality of Life as the Span of Cancer Survivability among the estimates of lung cancer datasets in six groups consisting of intense affects of carcinoma.

Fig. 3 – Survival Rate Analyses of Subjects affected by Lung Cancer based on Gene Expression Datasets

A sample dataset with time, event, and group indicators are synthesized to fit the Kaplan-Meier survival model. The plot accommodates the overall survival curve for all patients and compares for first two different groups, performs a log-rank test to statistically compare the curves, and calculates and prints the median survival time for each group. The data set resembles the complete load of real lung cancer data, actual dataset with time, event, and grouping variables. The analysis explores overall survival trends and visualizes, compare survival between different groups, statistically tested for significant differences, median survival times are estimated. The survival curves show the probability of survival over time, and the log-rank test ***p***-value indicates if the difference between groups is statistically significant. The median survival time gives an estimate of when 50% of patients in each group are expected to experience the event.

Table 2: Average Median Time of Critical Groups of Lung Cancer in Gene Expression Data

Groups	***Average Median Time***
Group 0	50.19
Group 1	45.65
Group 2	53.35
Group 3	48.65
Group 4	50.25
Group 5	49.35
Average	**49.57**

A basic framework of Kaplan-Meier analysis is experimented, further can be adjusted and expanded based on specific lung cancer prediction needs. The average median survival time for Group 0 and Group 1 is 50.19 and 45.65 respectively. As the suspension of cure of the carcinoma or identification of carcinoma is delayed, the subject will experience risky survival. The Kaplan-Meier test estimates the probable survival rate of the subjects based on the susceptibility. The estimate in the curve is provided with survival probability vs. time that indicates the downward trend, where the higher survival probability is persistent for small period of time and lower survival probability is persistent for a big period of time.

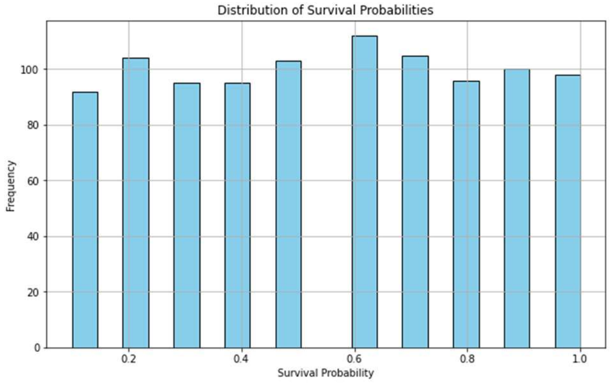


Fig. 3 – Distribution of Survival Probabilities based on Stroma Severity

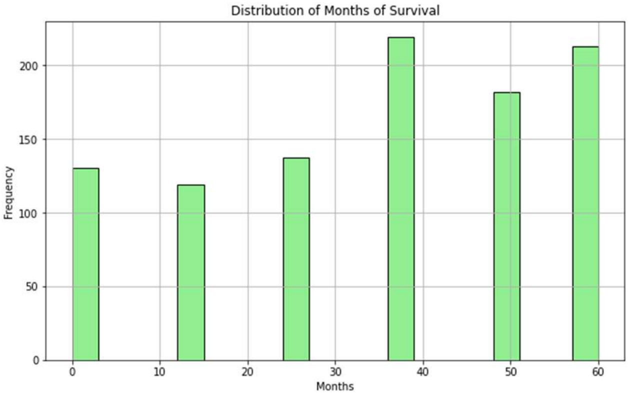


Fig. 4 –Survival Probabilities in Months

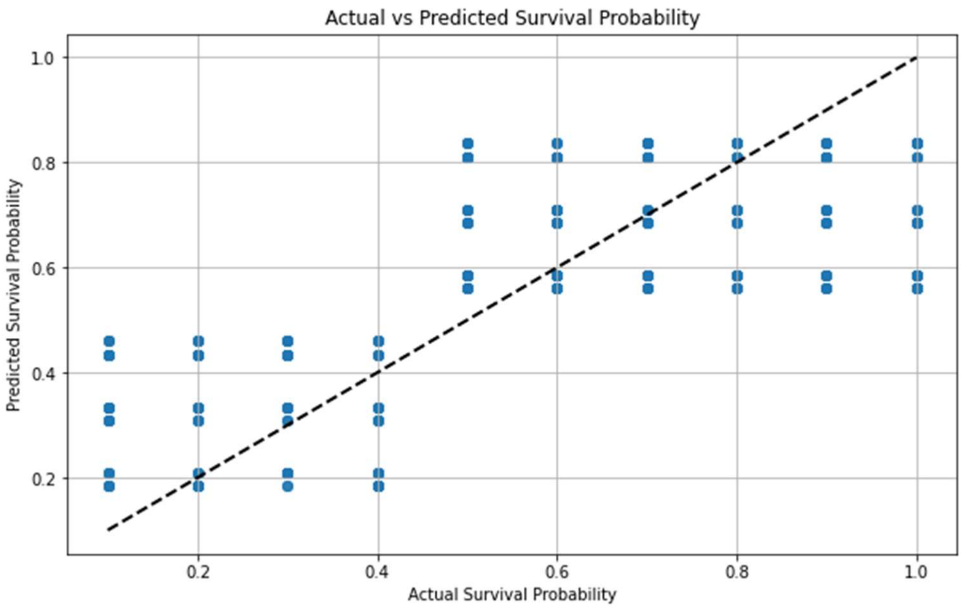


Fig.4 To demonstrate actual and predicted survivability - Ordinary Least Squares method of Linear Regression on Survival Variable

Results

Metrics for training and validation are essential for understanding performance. Lower values indicate better generalization, while higher values indicate good generalization. Precision, recall, F1-score, and AUC-ROC provide a detailed view of performance, especially with imbalanced datasets. Learning rate needs careful tuning to avoid suboptimal solutions or stuck models [17][18][19]. The number of epochs and batch size can affect training time and performance [18].

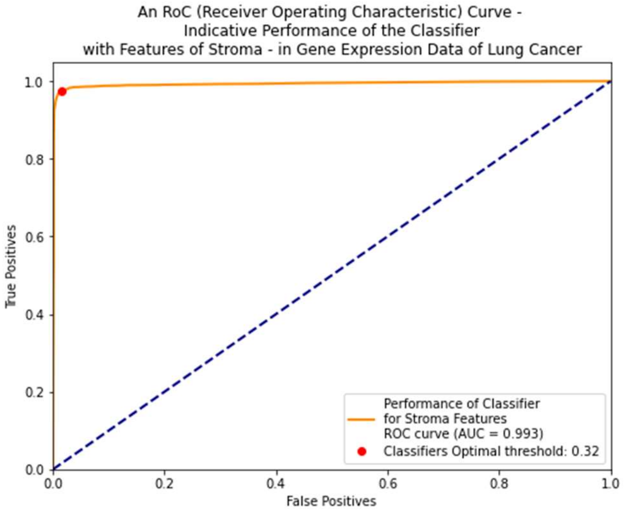


Fig.5 Performance of the Classifier for Stroma Features to determine positives in Gene Expression data of Lung Cancer

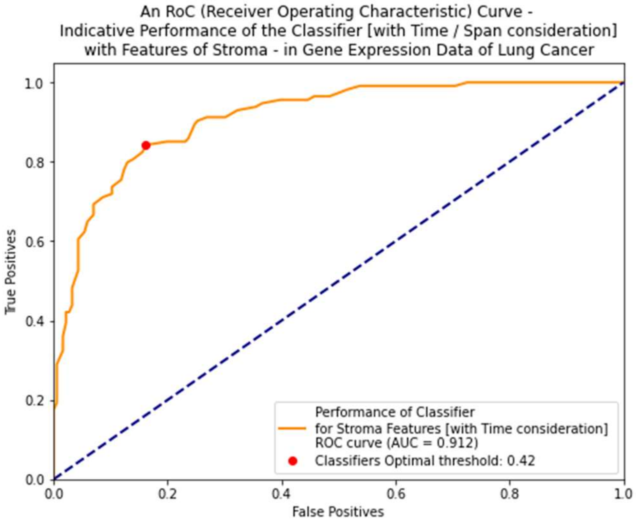


Fig.6 Performance of the Classifier for Stroma Features [with Time / Span consideration] to determine positives in Gene Expression data of Lung Cancer

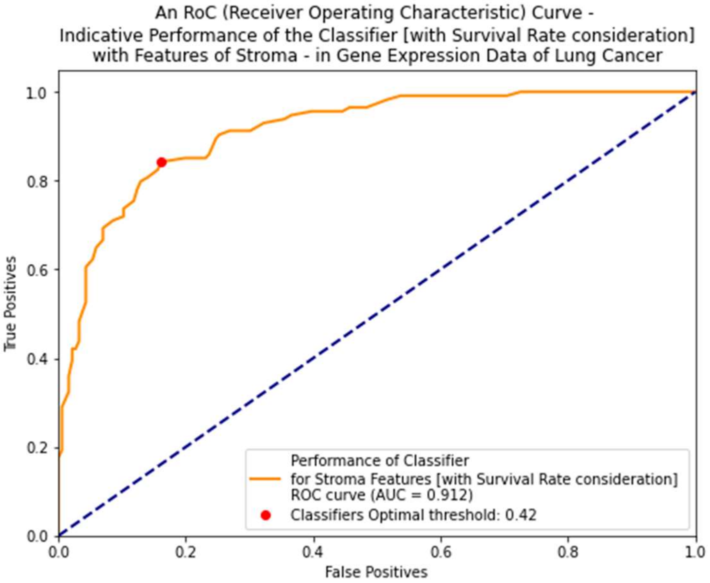


Fig.7 Performance of the Classifier for Stroma Features [with Survival Rate consideration] to determine positives in Gene Expression data of Lung Cancer

Monitoring these metrics helps in understanding the model's learning progress and making informed decisions about adjusting hyperparameters and improving the model. AUC-ROC [18][19]is a widely used method to assess the performance of a binary classification model like a Convolutional Neural Network, providing insight into the model's ability to differentiate between different classes. The AUC-ROC is a measure of model performance, ranging from 0 to 1, with a close AUC indicating a good model and a 0.5 AUC indicating a model with no discriminative power. The ROC curve follows the left-hand border and top border, with a closer curve indicating better accuracy. AUC-ROC offers threshold independence and is useful for comparing the performance of different models. The observations of the experiment referred in the above figure obtained metrics at optimal threshold Accuracy: 0.983, Precision: 0.968, Recall: 0.976, F1-score: 0.972.

6. Conclusion

This review paper lists several machine learning models trained by different backgrounds in recent years. It is incredibly challenging to treat lung cancer, but early detection will significantly improve the chances of survival. The machine learning technique provides a promising new pre-diagnosis method [19][20][21]. To predict the probability of lung cancer via machine learning models, doctors need to collect specific biological data from the patient. On average, the time and energy doctors spend on patients will be significantly shortened and saved while the accuracy of diagnosis remains the same or even higher. Machine learning in healthcare continuously evolves, meaning more factors related to lung cancer prediction will be discovered and applied. Also, multiple models can be used as cross-validation to provide a more accurate prediction.

The high accuracy allows the machine learning models to give a persuasive risk prediction of lung cancer, and the low cost brings a chance of screening to many people. In this way, models can be applied as a tool for large-scope screening for lung cancer in the future, and the patient can decide whether to do an LDCT for double screening. On the other hand, it takes a massive amount of data to train a high-quality machine-learning model. Still, people might not feel comfortable contributing their personal medication data due to the awareness of privacy protection. However, people's attitudes towards machine learning models will change as more and more people are helped by the models. In closing, machine learning technology is constantly improving, so its use in the medical industry will expand and become more effective in the future.

References

[1] Karim, Md Rezaul, Michael Cochez, Oya Beyan, Stefan Decker, and Christoph Lange. "*OncoNetExplainer: explainable predictions of cancer types based on gene expression data.*" In 2019 IEEE 19th International conference on bioinformatics and bioengineering (BIBE), pp. 415-422. IEEE, 2019.

[2] Liu, Suli, and Wu Yao. "*Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection.*" BMC bioinformatics 23, no. 1 (2022): 175.

[3] Borisov, Nicolas, Maxim Sorokin, Victor Tkachev, Andrew Garazha, and Anton Buzdin. "*Cancer gene expression profiles associated with clinical outcomes to chemotherapy treatments.*" BMC medical genomics 13 (2020): 1-9.

[4] Rukhsar, Laiqa, Waqas Haider Bangyal, Muhammad Sadiq Ali Khan, Ag Asri Ag Ibrahim, Kashif Nisar, and Danda B. Rawat. "*Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification.*" Applied Sciences 12, no. 4 (2022): 1850.

[5] Sethi, Bijaya Kumar, Debabrata Singh, Saroja Kumar Rout, and Sandeep Kumar Panda. "*Long Short-Term Memory-Deep Belief Network based Gene Expression Data Analysis for Prostate Cancer Detection and Classification.*" IEEE Access (2023). Digital Object Identifier 10.1109/ACCESS.2023.3346925

[6] Thakur, Tanima, Isha Batra, Arun Malik, Deepak Ghimire, Seong-Heum Kim, and ASM Sanwar Hosen. "*RNN-CNN Based Cancer Prediction Model for Gene Expression.*" IEEE Access 11 (2023): 131024-131044.

[7] Jia, Hao Ran, Wen Chao Li, and Lin Wu. "*The prognostic value of immune escape-related genes in lung adenocarcinoma.*" Translational Cancer Research (2024).

[8] Liu, Hangfeng, Jia Yao, Yulan Liu, Liping Wu, Zhiwei Tan, Jie Hu, Shigao Chen, Xiaolin Zhang, and Shuanghua Cheng. "*Diagnostic value of immune-related biomarker FAM83A in differentiating malignant from benign pleural effusion in lung adenocarcinoma.*" Discover Oncology 15, no. 1 (2024): 242.

[9] Jiang, Yupeng, Bacha Hammad, Hong Huang, Chenzi Zhang, Bing Xiao, Linxia Liu, Qimi Liu, Hengxing Liang, Zhenyu Zhao, and Yawen Gao. "*Bioinformatics analysis of an immunotherapy responsiveness-related gene signature in predicting lung adenocarcinoma prognosis.*" Translational Lung Cancer Research, https://dx.doi.org/10.21037/tlcr-24-309.

[10] Claudio Quiros, Adalberto, Nicolas Coudray, Anna Yeaton, Xinyu Yang, Bojing Liu, Hortense Le, Luis Chiriboga et al. "*Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides.*" Nature Communications 15, no. 1 (2024): 4596.

[11] Roy, Amrita, Martin Davis, Samuel Weinberg, Apurva Mallisetty, Alicia Hulbert, and Frank D. Weinberg. "*Stearic acid induces pro-inflammatory macrophage response important for lung cancer development.*" Cancer Research 84, no. 6_Supplement (2024): 176-176.

[12] Rath, Prangya, Abhishek Chauhan, Anuj Ranjan, Diwakar Aggarwal, Isha Rani, Renuka Choudhary, Moyad Shahwan et al. "*Luteolin: A Promising Modulator of Apoptosis and Survival Signaling in Liver Cancer.*" Pathology-Research and Practice (2024): 155430.

[13] Sherafatian, M., Arjmand, F. "*Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data*". Oncology Letters 18, no. 2 (2019): 2125-2131. https://doi.org/10.3892/ol.2019.10462

[14] Guan, X., Du, Y., Ma, R. et al. "*Construction of the XGBoost model for early lung cancer prediction based on metabolic indices*". BMC Medical Informatics Decision Making 2023, 107 (2023). https://doi.org/10.1186/s12911-023-02171-x.

[15] Huang, Tianzhi & Le, Dejin & Yuan, Lili & Xu, Shoujia & Peng, Xiulan. "*Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit*". PLOS ONE. 18. e0280606. 10.1371/journal.pone.0280606, 2023.

[16] Yang R, Xiong X, Wang H, Li W. "*Explainable Machine Learning Model to Prediction EGFR Mutation in Lung Cancer*". Frontiers Oncolology. 2022 Jun 23; 12:924144. doi:10.3389/fonc.2022.924144.

[17] A. V. Sriharsha, "*Detection of Holes on Indian Roads Using Information and Communication Technologies*," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ACCAI58221.2023.10199826.

[18] A V Sriharsha, Galety, M.G., Natarajan, A.K., "*Advanced Applications of Python Data Structures and Algorithms*", 2023, IGI Global Book, pp. 1–298 doi:10.4018/978-1-6684-7100-5

[19] A.V.Sriharsha et al. "*Electroencephalography image classification using convolutional neural networks.*" In The International Conference on Innovations in Computing Research, pp. 42-52. Cham: Springer International Publishing, 2022.

[20] A.V.Sriharsha, et al. (2023). "*Adaptable Fog Computing Framework for Healthcare 4.0.*" In Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022). SoCPaR 2022. Lecture Notes in Networks and Systems, vol 648. Springer.

[21] A.V.Sriharsha, "*Identification and Classification using Deep Learning Methods for Diagnosis of Mastocytosis: A Systematic Review.*" Telematique (2022): 434-433.