ENHANCING LUNG CANCER DIAGNOSTICS THROUGH ADVANCED COMPUTATIONAL METHODS

1st Mr. G. Sriram Ganesh, MTech, (Ph.D)
dept of COMPUTER SCIENCE and ENGINEERING
SRI VASAVI ENGINEERING COLLEGE
Tadepalligudem, INDIA
Sriramganesh8107@srivasaviengg.ac.in

2nd M. L. V. S. K. S. SEKHAR
dept of COMPUTER SCIENCE and ENGINEERING
SRI VASAVI ENGINEERING COLLEGE
Tadepalligudem, INDIA
20A81A0537@sves.org.in

3rd P. DURGA SRI SAI ASWIN
dept of COMPUTER SCIENCE and ENGINEERING
SRI VASAVI ENGINEERING COLLEGE
Tadepalligudem, INDIA
20A81A0544@sves.org.in

4th M. ADI LAKSHMI
dept of COMPUTER SCIENCE and ENGINEERING
SRI VASAVI ENGINEERING COLLEGE
Tadepalligudem, INDIA
20A81A0532@sves.org.in

5th K. BHUVANESWARI
dept of COMPUTER SCIENCE and ENGINEERING
SRI VASAVI ENGINEERING COLLEGE
Tadepalligudem, INDIA
20A81A0523@sves.org.in

6th PAVAN J
dept of COMPUTER SCIENCE and ENGINEERING
SRI VASAVI ENGINEERING COLLEGE
Tadepalligudem, INDIA
20A81A0541@sves.org.in

***Abstract*—In the contemporary landscape of global health, cancer remains a formidable challenge, with lung Carcinoma emerging as a particularly impactful and potentially life-threatening disease. It is quite challenging for medical professionals to detect Melanoma. It usually occurs in individuals of all genders as an outcome of unmanageable lung cell growth. This constitutes a severe respiratory problem in both the inhalation and the exhalation of the chest. The exact cause of this cancer and its definitive treatment have yet to be fully discovered. When cancer is identified in its early stages, it could be treated effectively. Current diagnostic techniques often face limitations in accurately identifying early-stage lung cancer, leading to delayed intervention and reduced treatment efficiency. To overcome this problem our web application will be the solution by providing a model for predicting and detecting lung cancer utilizing cutting-edge technologies, including ML and DL.**

***Keywords:* lung Carcinoma, diagnosis, Machine Learning(ML), Deep Learning(DL).**

I. INTRODUCTION

Cancer is a widespread disease and ranks among the top causes of death, characterized by abnormal cell growth. Cancer is a imperative issue of death globally, with a considerable scrap of these fatalities attributed to lung cancer. [5]. Lung cancer, a prevalent and lethal form of cancer, results from uncontrolled malignant cell growth in lung tissue, often leading to fatalities. With early detection, the treatment for the disease will be effective. among the various cancers, lung cancer poses a substantial global health burden, impacting millions worldwide with its severe respiratory complications and potential lethality. Despite advancements in medical science, the timely identification and its effective treatment remain challenging, leading to higher rates of illness and death.

Clinicians utilize X-rays, CT scans, PET scans, biopsies, and clinical evaluations in diagnosing lung cancer. Challenges in result interpretation and variations in descriptions can cause delays in treatment initiation post-diagnosis. The complexity of understanding test outcomes and inconsistencies in characterization can hinder timely exploit for patients interpreted with lung cancer. Addressing these challenges is vital to ensure timely and effective disease management. Enhancing the clarity of diagnostic findings and standardizing descriptions can streamline the process from diagnosis to treatment initiation, enhancing patient care and outcomes in combating lung cancer.

New and effective ways of hospitalization has upgraded using new technologies through analyzing extensive datasets, these AI techniques aim to reveal patterns and intuitions that can enhance the identification of lung cancer, leading to more effective medical treatments.

Enhancing Lung Cancer Prediction with Machine Learning Algorithms:

For predicting lung cancer, conventional ML algorithms similarly SVM, K-Nearest Neighbor, Logistic Regression, and Decision Tree Classifier are utilized. Among these we have choosen SVM, it owed high accuracy rate. Unlike CNNs, these algorithms work on parameters but not on image data. They excel at handling small datasets and can identify patterns effectively for predictive tasks. Smoking is undoubtedly confirmed as intricate in lung melanoma i.e. smoking was habit to patients with lung melanoma in many cases patient found to have this habit. However, several habits of those patients such as their smoking rate can be useful in early prediction [23]. Smoking is accountable for inducing 90% of it. Impregnation of tobacco smoke also causes lung cancer i.e. known as passive smoking [6]. By integrating various features such as medical history and some parameters like smoking, passive smoking, and so on, these algorithms collaboration therapists in risk stratification and helps them to treat the abnormality more efficiently.

Enhancing Lung Cancer Detection with CNNs:

In Medical pathology, CNN(Convolutional Neural Networks) takes center stage. These neural networks excel in imaging analysis tasks, making them essential for detecting anomalies indicative with lung melanoma in medical images. Despite their effectiveness, CNNs encounter challenges like data limitation as the image data in medical Pathology are not so readily available. Researchers aim to optimize CNN architectures and training strategies to enhance their performance in accurately identifying maladies of lung melanoma from medical images predominantly Convolutional Neural Networks (CNNs) excel in interpreting medical imaging data, similarly CT-Scans of lungs. These algorithms automatically identify abnormal patterns indicative and, aid in faster and more accurate diagnoses.

Our aspiration is to built-up a user-friendly web-based platform that integrates advanced technologies to boost forecast and detection of this abnormality. By utilizing sophisticated computational algorithms like SVM, known for its suitability and high accuracy in our application to overcome current diagnostic limitations and equip healthcare professionals with more enhanced tools for premature recognition and intervention.

we've accumulated insights from diverse domains like medicine, computer science, and data analysis. We are hopeful of improving the methods of tackling the lung cancer by using advanced technologies such as ML and its superset technology. Through these innovations, We are hopeful of creating a well-equipped arsenal of algorithms and procedures to formulate the world a better place by being able to fight these abnormalities.

II. LITERATURE REVIEW

A study [1] by Firdaus, Qurina, Riyanto Sigit, Tri Harsono, and Anwar Anwar, describes a lung melanoma recognition system utilizing CT-Scan images. The system incorporates multiple stages: pre-processing for image quality enhancement, segmentation to isolate the suspected cancerous region, attributes extraction based on characteristics like area and contrast, and final classification of cancer as benign or malignant. Their research reports an accurateness of 83.33 % in distinguishing between non-cancerous and cancerous cancers.

Raoof, Syed Saba, M. A. Jabbar, and Syed Aley Fathima [6] explored the efficacy of various classification algorithms and ensemble learning prototypes in identifying lung carcinoma. They evaluated several classifiers, including Multi Layered Perceptron(MLP), ANN like DNN, Decision Tree classifiers, Bayes Classifiers, , and SVM, along with ensembles like Random Forest and majority voting. Their findings highlighted that the Gradient Boosted Tree surpassed all other classifiers in accuracy.

Prof. Sangam Borkar and Nidhi S. Nadkarni[3] proposed and created an algorithm-based system to categorize lung melanoma in CT scans, that system was built-up by using diverse methodologies like image enhancement, segmenting the images, and extracting the information from the image. Median filtering was noted for its effectiveness in removing noise without blurring, and mathematical morphological operations were crucial for precise segmentation of lung and tumor areas. Radhika, P. R., Rakhi AS Nair, and G. Veena [8] Documented how ML algorithms are efficient in enhancing disease diagnostics, emphasizing lung cancer, renowned for its high mortality rate. They specifically mentioned the real time uses of algorithms such as Random Forests, SVM, K-Nearest Neighbor, and Naive Bayes in predicting lung cancer. Rahman, Md Sakif, Pintu Chandra Shill, and Zarin Homayra [9] introduced identifying and forecasting lung malignancy models with DL Model. This model focuses on the early identification and forecasting of lung cancer, starting with the preprocessing of CT images for quality improvement, followed by the separation of the left and right lungs to streamline the process, reduce complexity, and enhance accuracy in deep Learning’s neural network application. Anuradha D. Gunasinghe, Achala C. Aponso, and Harsha Thirimanna [2] created a system applying ML and DL techniques to support analyzing the lung diseases such as asthma, COPD, and lung cancer by analyzing patient data and chest X-rays. This system is designed as a decision support tool for physicians, aiming to facilitate early detection and improve patient care. M. Siddardha Kumar and Prof. Dr. K. Venkata Rao [4] suggested a unique machine learning methodology for lung melanoma identification that relies on direct interpretation of im- age features rather than numerical analysis, aiming to enhance diagnostic accuracy from imaging studies.

Nasser, Ibrahim M., and Samy S. Abu-Naser [10] developed an ANN model to identify lung cancer using symptoms and patient information. Validating it using ANN after training resulted in very effective accuracy.

III. PROPOSED METHODOLOGY

This methodology involves a comprehensive study of diverse ML algorithms and techniques that help to select an appropriate model using metrics like accuracy. The system architecture outlines the different stages involved in transforming datasets into training, testing and validation sets which helps the decision-making within the system. Testimony is segregated into test and training sets, a learning algorithm is employed on learning dataset, followed by various ML and DL algorithms to train perspective model.

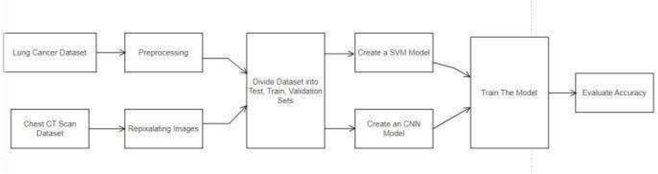


Fig. 1. Model Creation

A. Prediction steps for ML Algorithm:

1) Data Acquisition:: The dataset used, named Lung Cancer, was sourced from data.world, containing 24 features with 'Level' is very helpful in determining a patient's lung cancer phase.

2) Data Pre-processing:: Data pre-processing is decisive for preparing the initial data for utilization in ML models. It encompasses several steps:

a. Handling Missing Values: Addressing missing values is indispensable for ensuring robust model performance. Methods like imputation with mean, mode, or median values or employing separate algorithms are employed to handle missing data.

b. Encoding Categorical Data: Categorical variables are encoded into numerical values to facilitate ML model operations. Techniques like Label Encoding or mapping functions are applied to convert categorical variables.

c. Feature Selection: Feature selection involves extracting relevant features to improve model performance. correlation matrix analysis is utilized to identify attributes that play a significant role in the output, enhancing accuracy and training efficiency.

3) Splitting the Data:: The serene dataset is gash into learning and testing datasets, with the former worn to train ML models and the latter for evaluating model performance.

B. Detection Steps Using CNN:

Data Collection: CT-scan pictures for identifying lung cancer are gathered from the online source Kaggle.

1) Dataset Pre-processing:: Pre-processing of CT-scan images includes reading, resizing, noise removal, and segmentation, essential for DL model analysis in image categorization or detection tasks.

2) Validation Process:: A hold-out validation process is employed, allocating 70 % of data for training, 15 % for testing, and 15% of validation. An epoch value of 50 and the size of a batch is 32 are chosen for DL models.

C. Algorithms:

1) Support Vector Machine (SVM):: Mainly, the mentioned below algorithms are based on arithmetic. It combines or imagines the objective via nearby neighbors of a visualized point. It identifies associated items depending on the common labels of the training set's objects nearby. Predicting the necessary response for the testing set and learning an SVC classifier are two phases in the procedure.

2) K-Nearest Neighbor (KNN):: KNN uses the neighbours of a point visualized for grouping or categorising the target. It groups point mainly that relay on the majority labels of its neighboring points in the learning set. The process includes training an SVC classifier and predicting the required answer for the testing set.

3) Decision Tree:: It combines or imagines the objective via nearby neighbors of a visualized point. It identifies associated items depending on the common labels of the training set's objects nearby. Predicting the necessary response for the testing set and learning an SVC classifier are two phases in the procedure.

4) Logistic Regression:: LR is a trendy ML algorithm used in places where the process deals with the numeric data. It provides mathematical reasoning for categorization and it is also used for regression purposes The prospect of a binary outcome is predicted using independent variables. The model parameters undergo iterative adjustments to minimize a loss function through optimization algorithms.

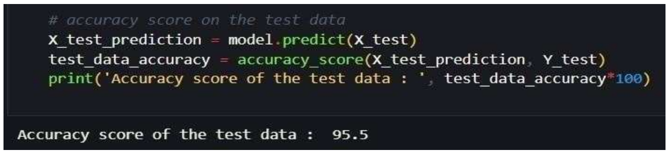
5) Convolutional Neural Network (CNN):: Deep neural networks, or CNNs, were created for image processing applications. Their effectiveness for therapeutic image scrutiny emanates from the capacity to automatically extract hierarchical properties from data. Convolutional layers are utilized by CNNs to detect patterns and attributes of pictures; pooling mantles are then worn to lower dimensionality and extract the pertinent features. Fully connected mantles are used for interpretation. The TensorFlow Keras module is make use of in it's construction, layer by layer adding layers to capture intricate correlations in CT-scan pictures to forecast of lung cancer. By carefully adjusting the network's weights and utilizing backpropagation techniques, it may identify innate correlations and subtle patterns in the data. Additionally, it includes biased input to deal with tight spots. The network's additional layers may contain varied relays on the purpose and intention of the method proposed.

IV. EXPERIMENTAL RESULTS AND EVALUATION PERFORMANCE

The dataset employed for forecasting was obtained from data.world having 1000 records and 26 features. These features and Smoking habits. Other features encompass allergies, occupational hazards, genetic predispositions, and symptoms like coughing, fatigue, and weight loss. Additionally, the dataset includes clinical indicators such as shortness of breath, wheezing, and swallowing difficulty, along with lifestyle factors like obesity, snoring, and dietary habits. The “Level” feature is likely the target variable, representing different stages or levels of lung carcinoma. In this lung malignancy prediction, it’s essential to select characteristics that hold clinical significance and are known or suspected to be linked with lung carcinoma risk or diagnosis.

For the CNN approach, the dataset originated from Kaggle, consisting of CAT scan images having both benevolent and malevolent conditions.

A. Logistic Regression: Predictive Model is about 95 % accurate.



```
# accuracy score on the test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
print('Accuracy score of the test data : ', test_data_accuracy*100)

Accuracy score of the test data :  95.5
```

Fig. 2. Accuracy (LR)

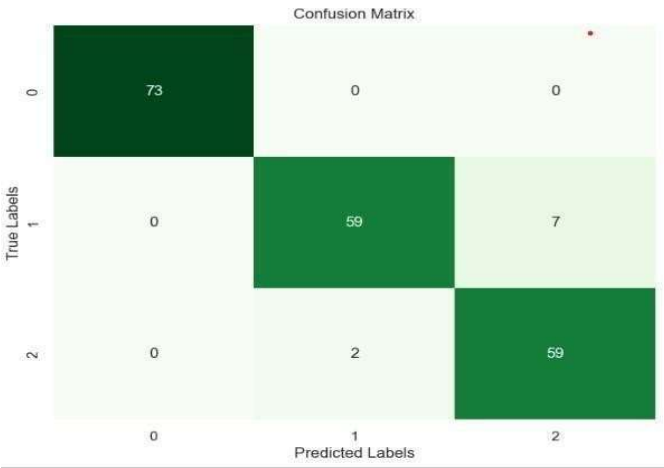
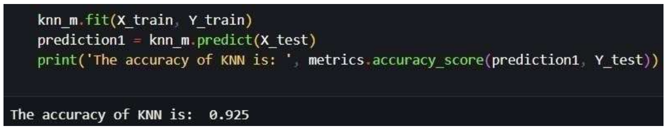


Fig. 3. Confusion matrix(LR)

B. K-Nearest Neighbor (KNN): Predictive Model is about 95 %

accurate.



```
knn_m.fit(X_train, Y_train)
prediction1 = knn_m.predict(X_test)
print('The accuracy of KNN is: ', metrics.accuracy_score(prediction1, Y_test))

The accuracy of KNN is:  0.925
```

Fig. 4..Accuracy (KNN)

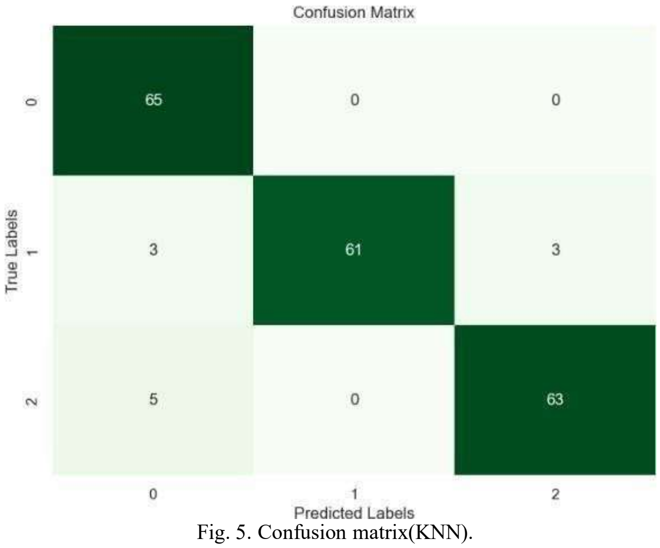
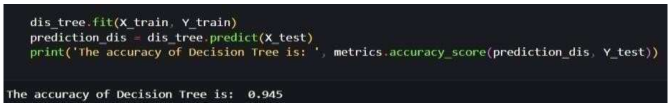


Fig. 5. Confusion matrix(KNN).

C. Decision Tree:

The final result is of 95 % acuurate .



```
dis_tree.fit(X_train, Y_train)
prediction_dis = dis_tree.predict(X_test)
print('The accuracy of Decision Tree is: ', metrics.accuracy_score(prediction_dis, Y_test))

The accuracy of Decision Tree is:  0.945
```

Fig. 6. Accuracy (DT)

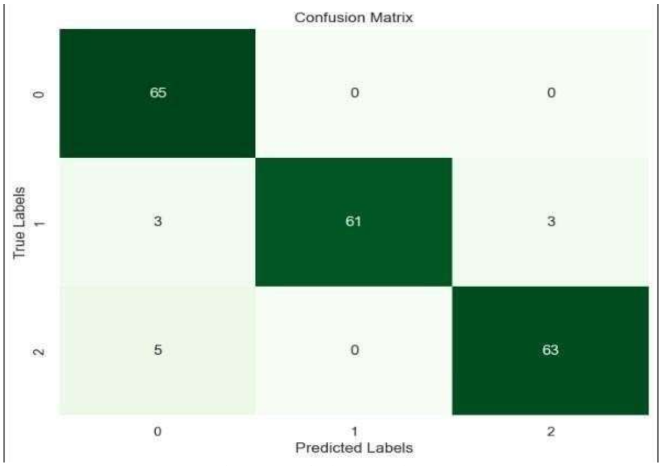
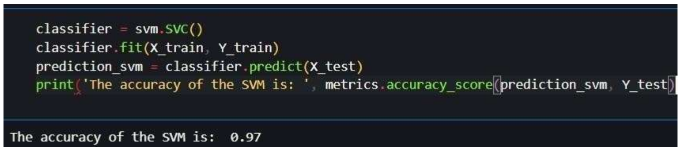


Fig. 7. Confusion matrix (DT)

D. Support Vector Machine (SVM):

The final result indicates a 97% accuracy rate, achieved through training all algorithms with 80% of the info and testing them with the remaining 20%.



```
    classifier = svm.SVC()
    classifier.fit(X_train, Y_train)
    prediction_svm = classifier.predict(X_test)
    print('The accuracy of the SVM is: ', metrics.accuracy_score(prediction_svm, Y_test)

The accuracy of the SVM is:  0.97
```

Fig. 8. Accuracy(SVM)

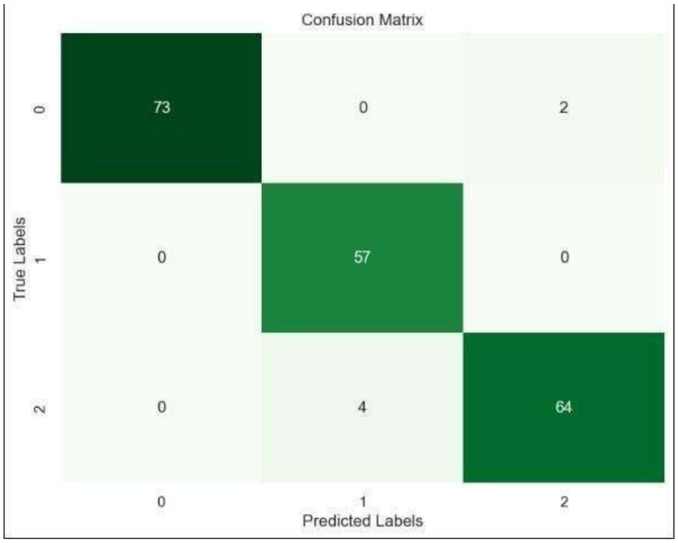
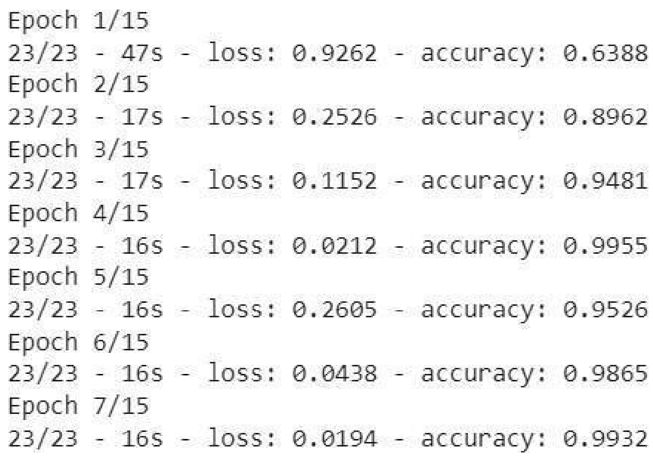


Fig. 9. Confusion matrix (SVM)

E. Convolutional Neural Network (CNN):

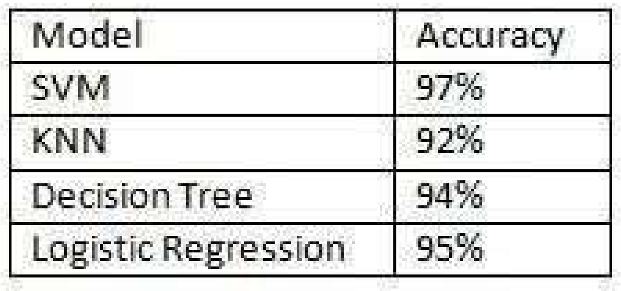
For CNN algorithm, it attains 98% Accuracy. Here are some epochs, that contain loss and accuracy



```
Epoch 1/15
23/23 - 47s - loss: 0.9262 - accuracy: 0.6388
Epoch 2/15
23/23 - 17s - loss: 0.2526 - accuracy: 0.8962
Epoch 3/15
23/23 - 17s - loss: 0.1152 - accuracy: 0.9481
Epoch 4/15
23/23 - 16s - loss: 0.0212 - accuracy: 0.9955
Epoch 5/15
23/23 - 16s - loss: 0.2605 - accuracy: 0.9526
Epoch 6/15
23/23 - 16s - loss: 0.0438 - accuracy: 0.9865
Epoch 7/15
23/23 - 16s - loss: 0.0194 - accuracy: 0.9932
```

Fig. 10. epochs for CNN

Upon conducting an scrutiny of accuracy and confusion matrix across multiple machine learning al- algorithms—such as Logistic Regression, SVM, K-Nearest Neighbours, and Decision Tree—it becomes evident that the SVM algorithm stands out as the preferred choice, primarily by reason of its superior accuracy rate.



Model	Accuracy
SVM	97%
KNN	92%
Decision Tree	94%
Logistic Regression	95%

Fig. 11. Algorithms with Accuracy

V. CONCLUSION

Within paper, we addressed the pressing need for improved lung cancer prediction and detection by clouting the capabilities of ML and DL techniques. Lung cancer continues to pose a significant challenge to global health, where early diagnosis is key to successful treatment and enhancing the prognosis for patients. Our research underscores the significance of leveraging ML and DL algorithms to augment diagnostic accuracy and facilitate timely intervention. Through the advancement of our exertion integrating ML and DL models, we provided a novel approach for predicting and detecting lung cancer based on relevant clinical and demographic features. By employing ML supervised algorithms like LR, SVM, KNN, and Decision Tree Classifier, alongside DL methodlogies like CNN, we achieved promising results, with CNN framework demonstrating an accuracy exceeding 95 % in predictive modeling using CT-scan images. Our breakthrough emphasize the prospective of ML and DL methodologies in revolutionizing lung cancer diagnostics, offering clinicians advanced tools to categorize and manage the disease more effectively. Notably, we erected that the SVM algorithm outperformed others regarding precision and suitability for our application. We achieved 97 % accurateness regarding lung melanoma prediction model using the SVM Algorithm. By overcoming the precincts of current diagnostic techniques, our web application acts as an important asset for healthcare professionals, facilitating early detection, personalized treatment strategies, and ultimately, improving patient outcomes. It comes in handy for a medical practitioner in analyzing the situations easily and more efficiently which reduces the time taken for the diagnoses and the Advanced developments in the disciplines of ML and DL to lung melanoma revealing hold promise for continued progress for melanoma concern and management.

VI. REFERENCES

[1] Firdaus, Qurina, Riyanto Sigit, Tri Harsono, and Anwar Anwar. “Lung cancer detection based on CT-scan images with detection features using gray level co-occurrence matrix (GLCM) and support vector machine (SVM) methods.” In 2020 International Electronics Symposium (IES), pp. 643-648. IEEE.

[2] Gunasinghe, Anuradha D., Achala C. Aponso, and Harsha Thirimanna. ”Early prediction of lung diseases.” In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1-4. IEEE, 2019.

[3] Nadkarni, Nidhi S., and Sangam Borkar. ”Detection of lung cancer in CT images using image processing.” In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 863-866. IEEE, 2019.

[4] Kumar, M. Siddardha, and K. Venkata Rao. ”Prediction of Lung Cancer Using Machine Learning Technique: A Survey.” In 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-5. IEEE, 2021.

[5] Mhaske, Diksha, Kannan Rajeswari, and Ruchita Tekade. ”Deep learning algorithm for classification and prediction of lung cancer using CT scan images.” In 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), pp. 1-5. IEEE, 2019.

[6] Raoof, Syed Saba, M. A. Jabbar, and Syed Aley Fathima. ”Lung Cancer prediction using machine learn- ing: A comprehensive approach.” In 2020 2nd Interna- tional Conference on Innovative Mechanisms for Indus- try Applications (ICIMIA), pp. 108-115. IEEE, 2020.

[7] Radhika, P. R., Rakhi AS Nair, and G. Veena. ”A comparative study of lung cancer detection using machine learning algorithms.” In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-4. IEEE, 2019.

[8] Rahman, Md Sakif, Pintu Chandra Shill, and Zarin Homayra. ”A new method for lung nodule detection using deep neural networks for CT images.” In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-6. IEEE, 2019.

[9] Nasser, Ibrahim M., and Samy S. Abu-Naser. ”Lung cancer detection using artificial neural network.” International Journal of Engineering and Information Systems (IJEAIS) 3, no. 3 (2019): 17-23.