ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

Early Prediction of Disease Using Machine Learning: Leveraging Medical Data for Accurate Classification

Rahul P. Mahajan[1]

[1]Research and Development Department Healthcare and Medical Device Development Industry
College of Engineering, Pune, India

Publication Date: 2025/04/15

Abstract: The accurate prediction of diseases at early stages is vital to enhance patient outcomes especially when dealing with fatal conditions such as cancer. The most prevalent cancer that may be lethal if left untreated is lung cancer. Clinical success in diagnosis and treatment depends on discovering health conditions during early stages when treatment remains effective for severe cases. There are a number of methods for predicting cancer severity that make use of deep learning and machine learning. A deep learning approach that utilizes Convolutional Autoencoders (CAEs) performs detection of lung cancer from examinations of histopathology images. The model training for assessing classification performance utilizes LC25000 dataset by adopting advanced preprocessing methods that execute data augmentation along with noise reduction and normalization. The CAE model brings superior performance than standard deep learning techniques CNN, VGG19 and ResNet-50 by attaining accuracy at 99.41% with precision at 98.52%, recall at 98.51% and F1-Score at 98.51%. However, using ROC and Precision-Recall curves, the model shows that it can differentiate between various cancer subtypes. In medical contexts, the study shows that deep learning techniques may accurately identify early lung cancer on a wide scale, leading to better clinical diagnosis.

Keywords*: Healthcare, Disease Diagnosis, Clinical Research, Lung Cancer, MRI, X-Ray, CT Scan, Medical Imaging, Machine Learning, LC25000 Data.*

How to Cite: Rahul P. Mahajan. (2025). Early Prediction of Disease Using Machine Learning: Leveraging Medical Data for Accurate Classification. *International Journal of Innovative Science and Research Technology*, 10(3), 2897-2907. https://doi.org/10.38124/ijisrt/25mar1858.

I. INTRODUCTION

Medical facilities operate as critical components by detecting illnesses to provide management and treatment for diverse health conditions. The continuous evolution of medical science, clinical research, and advanced technologies has significantly improved disease detection and patient care [1][2]. Among various health conditions, chronic and life-threatening diseases such as cancer pose a major challenge to healthcare systems worldwide [3][4]. Cancer, a complicated illness characterized by uncontrolled cell development, affects millions of individuals each year and remains one of the top causes of death [5][6]. Timely diagnosis enables more effective treatment and improved patient outcomes, making early discovery crucial for increasing survival rates.

A considerable portion of the world's cancer fatalities are attributable to lung cancer, which is among the most aggressive forms of the disease. It arises when aberrant cells in the lungs proliferate uncontrollably as a result of genetic abnormalities [7]. The primary risk factors include smoking, air pollution, occupational hazards, and genetic predisposition. The timing of a lung cancer diagnosis is crucial to the disease's prognosis. Early diagnosis is crucial, as advanced-stage lung cancer significantly reduces the chances of survival due to metastasis [8].

Traditional diagnostic techniques for lung cancer include medical imaging modalities like MRI, X-ray, and CT scans, which provide detailed anatomical views of lung tissues to detect abnormalities. Additionally, sputum cytology is used to examine mucus samples for the presence of cancerous cells, while tissue sampling through biopsy remains the gold standard for definitive diagnosis. Unfortunately, these techniques may be time-consuming and often need expert interpretation, which delays the start of therapy and diagnosis. The integration of AI and ML in medical diagnostics has revolutionized early disease prediction and classification [9]. ML models have shown to be very effective in enhancing diagnosis accuracy via pattern recognition and analysis of intricate medical data. DL, a subset of ML [10], enhances image-based cancer detection by extracting intricate features from MRI, CT scans, and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

histopathological images, reducing human error and optimizing decision-making [11]. AI-driven approaches enable faster and more reliable lung cancer detection, assisting radiologists and clinicians in early intervention. Using ML methods to forecast the occurrence of lung cancer in its early stages by combining information from imaging studies, sputum cytology, and biopsy samples.

A. Motivation and Contribution of the Paper

Healthcare expense reduction, together with better patient results, depends strongly on identifying illnesses early. However, traditional diagnostics face challenges like late detection, human error, and inefficiencies in handling large medical data. A game-changing answer is provided by ML and DL, which automate feature extraction and categorization with remarkable accuracy. This study is motivated by the need to develop an advanced ML-based framework that enhances disease prediction by leveraging medical data, including histopathological images and clinical records. The study's overarching goal is to help patients and healthcare professionals alike by increasing diagnosis accuracy and allowing for earlier intervention via the use of classification models and strong preprocessing methods. The area of early illness prediction using ML has benefited greatly from the contributions made in this paper:

- Histological images of lung and colon cancer are included in the LC25000dataset, which is used for both training and validation of the model.
- Implements advanced preprocessing techniques, including noise removal, data augmentation, and normalization to ensure high-quality input for ML models.
- Used MobileNetV2 for automated feature extraction, capturing intricate patterns in histopathological images without manual intervention.
- Various deep learning architectures, such as CNN, CAE, VGG19, and ResNet-50, are employed and compared for disease classification to identify the most effective model.
- Evaluates model performance by calculating confusion matrices and employing measures like recall, accuracy, precision, and F1-score to provide a thorough assessment of classification effectiveness.

B. Significance and Novelty

This study significantly advances lung cancer prediction by integrating ML techniques with rigorous data preprocessing to enhance diagnostic accuracy. Unlike traditional methods, which are prone to late detection and human error, the proposed approach leverages DL models trained on high-quality, preprocessed histopathological images and clinical data. The novelty lies in the comprehensive preprocessing pipeline and classification models. Furthermore, the comparative evaluation of multiple DL architectures, such as CNN, VGG19, and ResNet-50, provides insights into the most effective model for lung cancer detection. This research helps improve healthcare AI-driven diagnostic systems by enhancing preprocessing and using cutting-edge classification approaches, leading to more dependable, scalable, and efficient systems.

C. Structure of the Paper

This paper is organized in the following way: Section II explores related studies on lung cancer detection. Section III discusses the proposed approach for lung cancer detection. Section IV provides a comparative analysis of model performance with visual representations. Finally, Section V provides a brief overview of the key topics and recommendations for more research.

II. LITERATURE REVIEW

A literature study on medical data-driven ML for lung cancer prediction reveals how deep learning architectures have improved diagnostic interpretability and classification accuracy.

Zhao et al. (2025) carried out comprehensive transformer-based fusion techniques and ablation experiments on multi-scale structures to investigate the effect of characteristics acquired at various scales on the precision of lung nodule classification. Verification findings on the LUNA16 dataset demonstrated that the proposed MSTD achieved 90% in terms of F1Score (94.5%), Specificity (96.5%), and Sensitivity (91.1%), indicating that it correctly identifies both benign and malignant lung nodules.

Tisha and Ani (2024) implement a machine-learning algorithm to predict lung cancer by using the most correlated symptoms of lung cancer. By applying data visualization and feature extraction techniques, they found the most correlated symptoms of lung cancer and predicted them using LR, SVM, and KNN classifiers; among them, SVM and LR both showed higher accuracy in lung cancer detection, with 98.52 % accuracy in the training set and 98.39 % accuracy in the testing set [12].

Sathishkumar and Parameswari (2024) developed a model that uses Radial Basis Function Neural Network algorithms and Divide and Conquer Kernel Support Vector Machine Learning techniques to forecast lung cancer in its early stages using the lung cancer patient databases from Kaggle. This is done in order to overcome these conventional methods. Accuracy, precision, and recall are the three measures used to assess these algorithms' performance. With an accuracy of 98.5, neural networks outperform algorithms like SVM. The dataset is preprocessed using techniques, including the approach known as SMOTE, to solve the problems of class imbalance [13].

Singh et al. (2023) takes a look at the performance evaluations of a number of ML ensemble methods that have been applied to the problem of lung cancer diagnosis, including SVM, XGBoost, LightGBM, AdaBoost, CatBoost, and RF. AdaBoost and XGBoost gave accuracy ratings of 96.77% and 96.76%, respectively, better than the other methods. Consequently, they deduce that ML-based methods have enormous potential to enhance lung cancer detection and lower death rates [14].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

Vishwakarma et al. (2023) choose the most effective algorithms for predicting lung cancer. For the purpose of lung cancer prediction, this study included a number of ML methods, including Naïve Bayes (90.61 percent accuracy), Decision Tree (91.52% accuracy), Random Forest (93.22% accuracy), Logistic Regression (96.61% accuracy), and Multilayer Perceptron (98.30 percent accuracy). Among these algorithms, MLP stands out as the most effective in diagnosing lung cancer [15].

S, R and B, (2022) the majority of lung cancer predictions are based on multi-stage categorization. Numerous techniques, including SVM, KNN, DT, LR, NB, and RF, are used to train the dataset; their greater accuracy has been shown. The Random Forest method has resulted in an improved performance level of 88.5% accuracy [16].

Mamun et al. (2022) examined a few studies on ML and ensemble learning approaches for lung cancer prediction models. Additionally, this study was updated to include their recently created ensemble learning methods. A dataset of 309 people with and without lung cancer was used to construct these approaches utilizing the oversampling SMOTE method of surveying. Results: Their results show that compared to other ensemble strategies, XGBoost outperformed them all with regard to accuracy (94.42%), precision (95.66%), re-call (94.46%), and AUC (98.14%) [17].

Despite significant advancements in ML for lung cancer detection, existing models face challenges such as class imbalance, overfitting, limited generalization, and insufficient interpretability in real-world clinical applications. Most studies rely on traditional classifiers or ensemble techniques without effectively leveraging deep feature extraction. To address these gaps, propose a Convolutional Autoencoder (CAE)-based model that integrates automated feature extraction with DL, ensuring higher classification accuracy and improved generalization. Table I summarizes various ML methodologies for lung cancer prediction, highlighting their datasets, performance metrics, and limitations.

Table 1: Summary of the Related Work on Lung Cancer Prediction Using Machine Learning

References	Methodology	Dataset	Performance	Limitations & Future Work
Zhao et al. (2025)	Multi-scale transformer-based fusion techniques for lung nodule classification	LUNA16 dataset	F1-Score: 94.5%, Specificity: 96.5%, Sensitivity: 91.1%	Further evaluation on diverse datasets needed
Tisha and Ani, (2024)	Logistic Regression, SVM, KNN with feature extraction and data visualization	Lung cancer symptoms dataset	Training Accuracy: 98.52%, Testing Accuracy: 98.39%	Needs real-world clinical validation
Sathish kumar & Parameswari (2024)	Radial Basis Function Neural Network and Divide & Conquer Kernel SVM	Kaggle lung cancer patient dataset	Neural Network Accuracy: 98.5%	Requires testing on larger and more diverse datasets
Singh et al. (2023)	Ensemble techniques (SVM, XGBoost, LightGBM, AdaBoost, CatBoost, Random Forest)	Various lung cancer datasets	AdaBoost Accuracy: 96.77%, XGBoost Accuracy: 96.76%	Further generalization across clinical datasets needed
Vishwakarma et al. (2023)	Various ML models (Naïve Bayes, Decision Tree, Random Forest, Logistic Regression, MLP)	WHO lung cancer statistics	MLP Accuracy: 98.30%	Needs validation on real-time patient data
S, R, and B (2022)	Multi-stage classification, segmentation (Threshold, Marker-controlled Watershed)	Various lung cancer datasets	Random Forest Accuracy: 88.5%	Improvement in segmentation techniques required
Mamun et al. (2022)	Ensemble learning (XGBoost, LightGBM, Bagging, AdaBoost) with SMOTE oversampling	Survey dataset (309 individuals)	XGBoost Accuracy: 94.42%, Precision: 95.66%, Recall: 94.46%, AUC: 98.14%	Requires expansion to larger medical datasets

III. METHODOLOGY

The suggested research makes use of ML methods to accurately classify medical data in order to make early predictions of lung cancer. Initially, medical data (LC25000) is collected from publicly available datasets containing diagnostic and clinical records. The data quality receives enhancement through image-processing methods such as noise reduction and augmentation together with normalization. The following step involves extracting features using MobileNetV2-based DL architectures to maintain essential patterns found in the data. The dataset undergoes split testing through an 80:20 ratio which preserves unbiased evaluation procedures. The next step involves implementing different ML algorithms among which are CNN and Variational Autoencoders (CAE) together with VGG19 and ResNet-50. Finally, model performance is evaluated employing metrics like accuracy, precision, recall, and F1score, with confusion matrix analysis providing deeper insights into classification effectiveness. Figure 1 lays out the steps used to improve ML-based early illness prediction by integrating data preparation, model training, and performance evaluation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

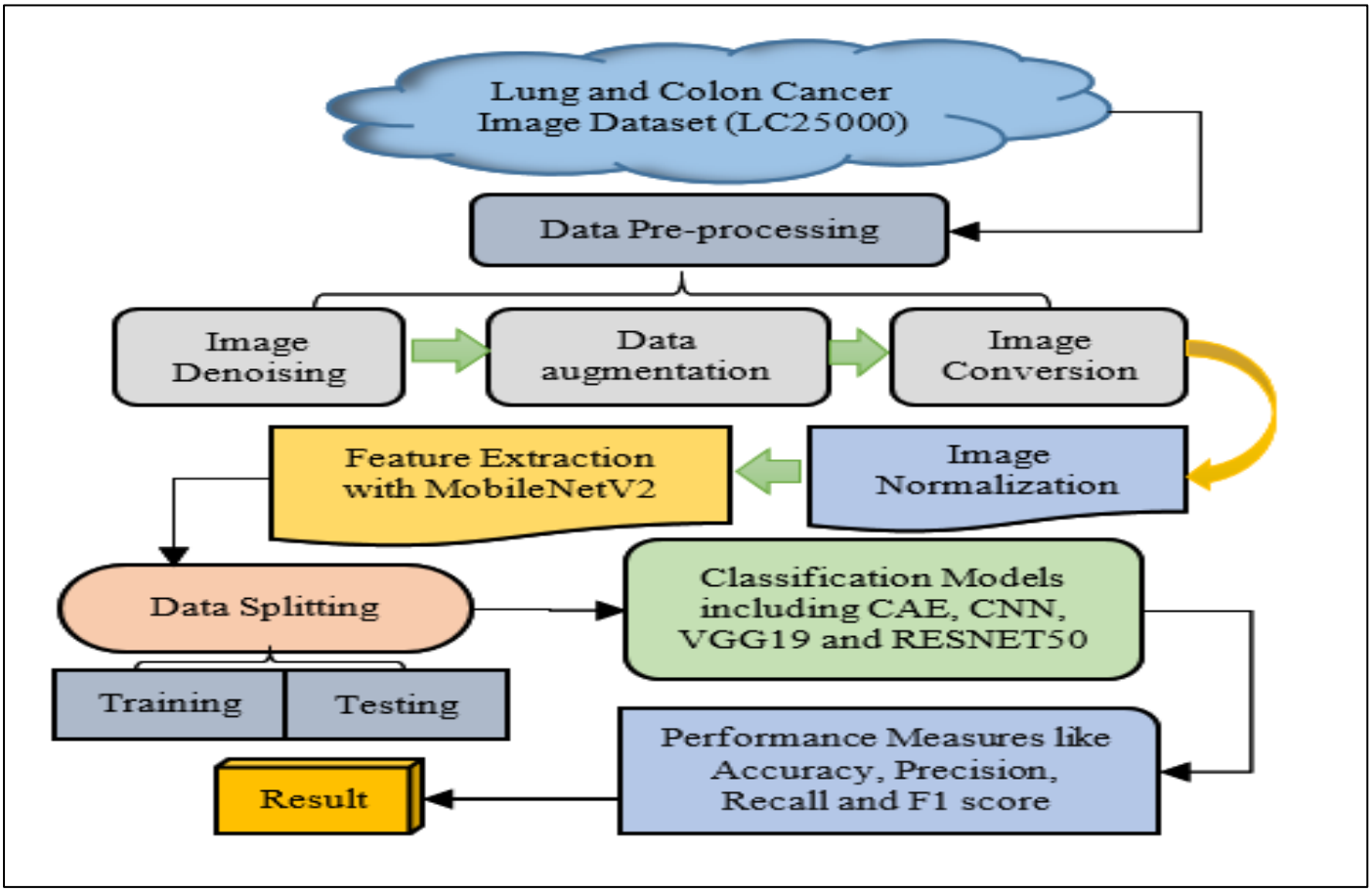

Fig 1: Flowchart for Lung Cancer Prediction

The systematic proposed approach steps are explained in below:

A. Data Collection

The publicly available LC25000 dataset contains 25,000 histopathological pictures with their main purpose being to detect lung and colon cancer. The 5 evenly distributed classes in the dataset feature 5000 pictures. The images in this dataset maintain a high level of resolution (768 × 768 pixels) through which deep learning models analyze stained tissue samples under microscope visualization. Figure 2 displays sample images for each class.

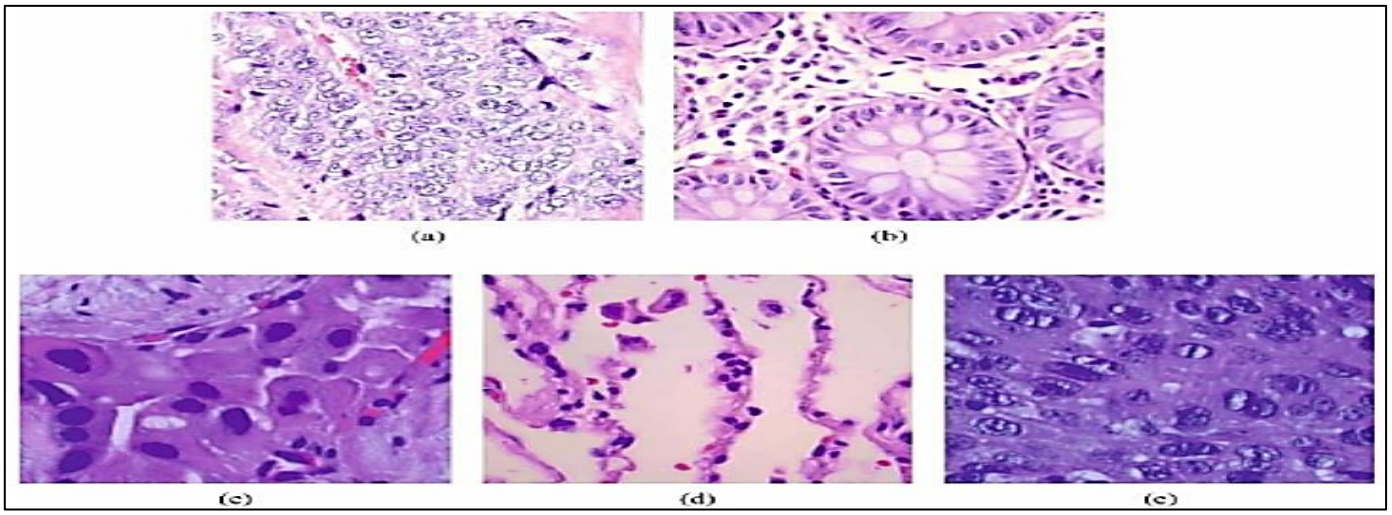

Fig 2: Sample Images of Data

Figure 2 shows sample histopathological images such as Colon Adenocarcinoma (Con-Adc), Colon Benign Tissue (Con-BeT), lung adenocarcinoma (Lug-Adc), Lung Benign Tissue (Lug-BeT), and Lung Squamous Cell Carcinoma (Lug-Scc)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

B. Data Preprocessing

Data preprocessing plays a critical role in data analysis and machine learning projects. In this study, we carried out data transformation involving handling missing or damaged data and converting data into a suitable format for machine learning algorithms. Missing values were carefully imputed to avoid bias and maintain prediction accuracy, while categorical variables were label-encoded to convert them into numerical values. Additionally, continuous numerical features (Total Charges, Monthly Charges, Tenure Months) were normalized using Min-Max Scaler to fit within a predefined range, typically 0-1.

These preprocessing steps ensure that the data is appropriately prepared for the machine learning algorithms used in this study Data preprocessing plays a critical role in data analysis and machine learning projects. In this study, we carried out data transformation involving handling missing or damaged data and converting data into a suitable format for machine learning algorithms. Missing values were carefully imputed to avoid bias and maintain prediction accuracy, while categorical variables were label-encoded to convert them into numerical values. Additionally, continuous numerical features (Total Charges, Monthly Charges, Tenure Months) were normalized using Min-Max Scaler to fit within a predefined range, typically 0-1.

These preprocessing steps ensure that the data is appropriately prepared for the machine learning algorithms used in this study Data preprocessing plays a critical role in data analysis and machine learning projects. In this study, we carried out data transformation involving handling missing or damaged data and converting data into a suitable format for machine learning algorithms. Missing values were carefully imputed to avoid bias and maintain prediction accuracy, while categorical variables were label-encoded to convert them into numerical values. Additionally, continuous numerical features (Total Charges, Monthly Charges, Tenure Months) were normalized using Min-Max Scaler to fit within a predefined range, typically 0-1. These preprocessing steps ensure that the data is appropriately prepared for the machine learning algorithms used in this study.

Data preprocessing plays a critical role in data analysis and machine learning projects. In this study, we carried out data transformation involving handling missing or damaged data and converting data into a suitable format for machine learning algorithms. Missing values were carefully imputed to avoid bias and maintain prediction accuracy, while categorical variables were label-encoded to convert them into numerical values. Additionally, continuous numerical features (Total Charges, Monthly Charges, Tenure Months) were normalized using Min-Max Scaler to fit within a predefined range, typically 0-1. These preprocessing steps ensure that the data is appropriately prepared for the machine learning algorithms used in this study.

Data preprocessing plays a critical role in data analysis and machine learning projects. In this study, we carried out data transformation involving handling missing or damaged data and converting data into a suitable format for machine learning algorithms. Missing values were carefully imputed to avoid bias and maintain prediction accuracy, while categorical variables were label-encoded to convert them into numerical values. Additionally, continuous numerical features (Total Charges, Monthly Charges, Tenure Months) were normalized using Min-Max Scaler to fit within a predefined range, typically 0-1. These preprocessing steps ensure that the data is appropriately prepared for the machine learning algorithms used in this study.

The most important step in obtaining accurate data free of unwanted distortions and highlighting the image features that will be useful for subsequent processing is preprocessing the pictures. To improve the classification results, some steps were performed are described below.

C. Image Denoising

To get rid of the noise that appears in the supplied images [18]. The salt-and-pepper noise, which causes random pixel conversions to white or black, is one kind of noise that this method excels at eliminating.

D. Data Augmentation

Data augmentation was applied to expand the dataset by generating additional images through various transformation techniques. Each subtype of cancer, initially represented by images, was augmented using left and right rotations, as well as horizontal and vertical flips, effectively increasing the sample size of images per subtype.

E. Image Conversion

Prior to processing the images to accommodate the image intensity values, the resized image was transformed to bgr2rgb. In order to make the photos square, their original dimensions of 1024 × 768 pixels were shrunk to 768 × 768 pixels.

F. Image Normalization

The range of possible values from 0 to 1 can be altered by normalization [19]. Images are normalized to a common scale after training using data that has varying intensity levels and scales [20]. The range of pixel intensities can be reduced by image normalization. In Equation (1), it can see the standard form of normalization.

$$F_{norm} = \frac{(F - F_{min})}{(F_{max} - F_{min})} \quad (1)$$

where, F is the normalization value, Fmin is the lowest pixel value, Fmax is the highest pixel intensity value relative to an image.

G. Feature Extraction

These preprocessed photos have had the MobileNetv2 architecture applied to them in order to generate a set of features [21][22]. Improving efficiency while maintaining competitive accuracy on separate tasks like object identification and image classification, it is an evolution of the original MobileNet concept. The MobileNetV2 network

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

uses the ReLU6 activation function with the Linear activation function.

H. Data Splitting

Data splitting partitions a dataset into training, and test sets for machine learning evaluation. In an 80:20 split, 80% is used for training to learn patterns, while 20% is reserved for testing to assess model performance objectively.

I. Proposed Convolutional Autoencoder (CAE) Model

A CNN [23][24]based autoencoder is called Convolutional Autoencoder (CAE). To encode the data image to its packed area representation, it employs Convolutional and Down Sampling (Pooling) layers [25]. The primary applications of autoencoders are dimensionality reduction of data and picture denoising, along with learning latent representations for the generation of new data samples [26]. As far as CAE hyperparameters go, the most important ones are the convolutional kernel size, the total number of convolutional layers, and the filter size of each layer. Their model's encoder is made up of three convolution blocks, and in between each convolution layer, there is a batch normalization layer with a 3×3 kernel size [27]. The output features of the convolutions are down-sampled using a max-pooling layer with a 2×2 kernel size, which is applied after the first and second convolution blocks [28]. Figure 3 shows the structure.

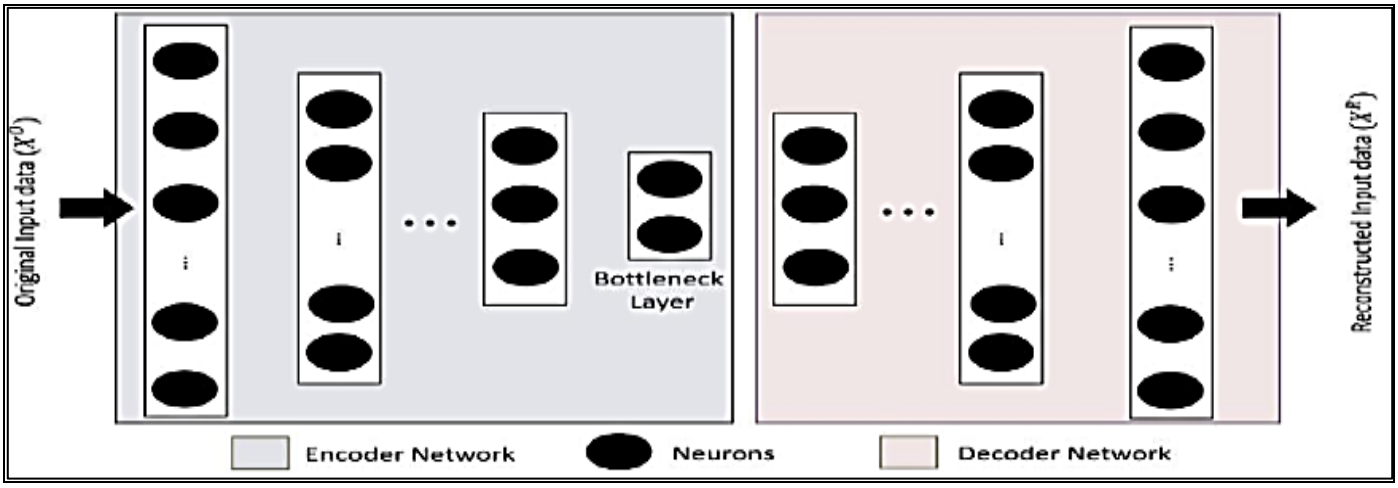

Fig 3: Structure of the Convolutional Autoencoder

A function for activation is included after the convolution-deconvolution layer, as seen in Equation (2):

$$h^k = \sigma(\sum_{l \in L} x^l \otimes w^k + b^k) \quad (2)$$

Where

- h^k= this layer's hidden representation of its kth feature map
- σ = an activation function
- x^l= the feature map that was derived from the preceding layer, which is the lth one in the set L.
- ⊗ = a 2D convolution operation
- w^k = variables that represent the current layer's k-th feature map b^k= the present layer's bias in its k-th feature map.

This optimizes the model for the training process by setting the learning rate to 0.0001, batch size to 4, and number of epochs to 25. At the end of each epoch, validation was carried out to evaluate the learning error following iterations. The models are trained using the momentum-based stochastic gradient descent optimizer, with all other hyperparameters kept at their default values.

J. Performance Measures

The proposed system's performance on the LC25000 dataset's classifications is evaluated using a confusion matrix-like statistic, among other metrics. The confusion matrix may be used to evaluate the recall, accuracy, precision, f1-score, and classification model. For each class, it shows the model's correct and wrong predictions. There are four parts to the confusion matrix:

- **True Negative (TN):** TN is the prediction for the patients without disease that was found to have no disease,
- **False Negative (FN):** FN is the prediction for the patients without disease that were found to have disease [29],
- **True Positive (TP):** TP is the prediction for the patients with a disease that were found to have a disease, and
- **False Positive (FP):** FP is the prediction for the patients with a disease that were found to have no disease.

➢ *Accuracy*

The total number of forecasts divided by the number of right predictions (TP and TN) yields the model's overall prediction accuracy [30]. Equation (3) provides the formula:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (3)$$

➢ *Precision*

Precision quantifies the proportion of positive predictions that are correct, the formula for calculating the precision is given below in Equation (4):

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

➢ *Recall*

The recall, which is sometimes called sensitivity, measures how well the model can detect positive examples. The recall formula is provided below in Equation (5):

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

➢ *F1-Score*

When dealing with balanced datasets, the F1-Score is beneficial since it strikes a compromise among Precision and Recall. It is the harmonic mean of Precision and Recall. Equation (6) presents the formula for F1-score:

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

➢ *ROC Curve*

The ROC curve evaluates a model's ability to distinguish between TP and FP for each class, with the AUC serving as a key performance indicator, where TPR is the True Positive Rate or Recall in Equation (7):

$$AUC = \int_0^1 TPR(x)dx \quad (7)$$

➢ *Loss*

Machine learning models achieve their loss metric by matching predicted results to their designated target values [31]. The training aims to decrease the loss function value as the main objective.

IV. RESULT & DISCUSSION

This section shows how successful the techniques are by analyzing the performance of the suggested deep learning models. The experiments are conducted on a system equipped with an Intel Core i5-8600K processor, GeForce GTX 1050Ti 4GB GPU, 16GB RAM, 250GB SSD, and 1TB HDD, ensuring efficient computational performance. The implementation is carried out using Python 3.8.5 in a Jupyter Notebook environment to enhance code organization and reusability. The performance evaluation is based on the LC25000 dataset, which contains histological photos for the classification of lung and colon cancer. Table II presents the CAE model performance across accuracy, precision, recall, and f1score.

Table 2: Results of Convolutional Autoencoder Model for Lung Cancer Disease Detection

Performance Metrics	Convolutional Autoencoder (CAE)
Accuracy	99.41
Precision	98.52
Recall	98.51
F1-score	98.51

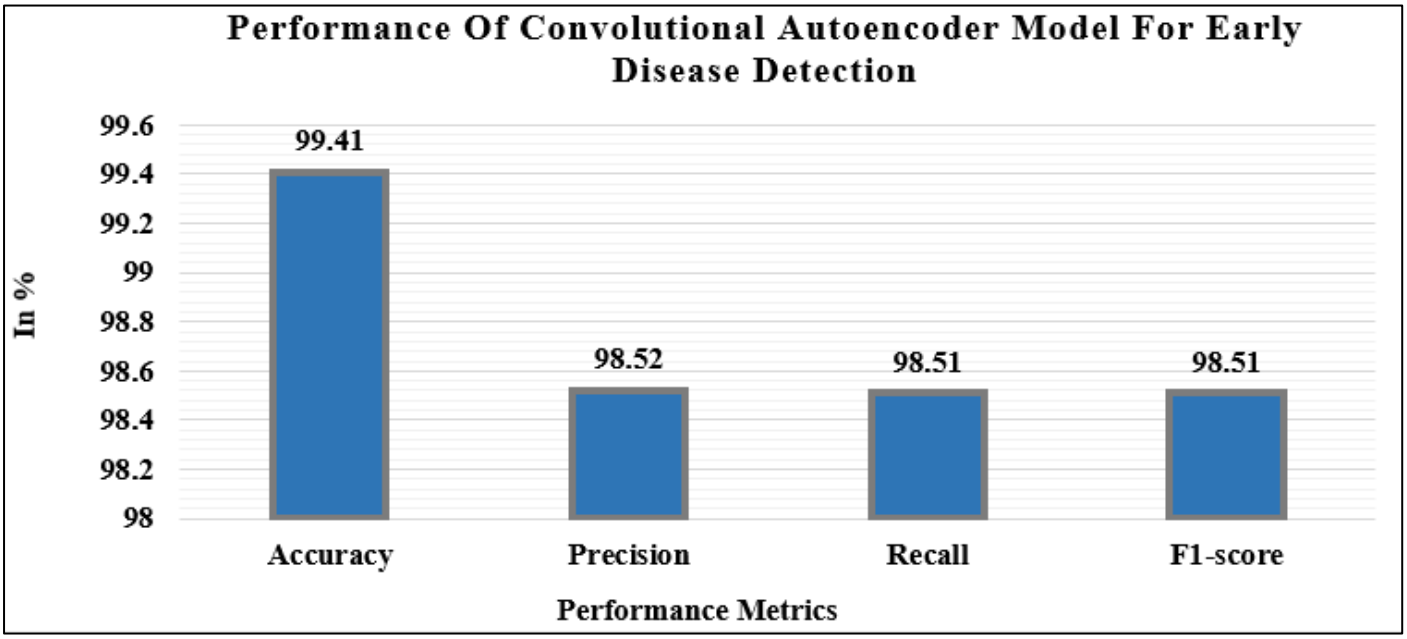

Fig 4: Bar Graph for CAE Model

Figure 4 together with Table II demonstrates how the CAE model works as evidenced in its F1score and re-call and accuracy and precision metrics. The model reaches 99.41% accuracy in its ability to properly identify and classify multiple cases. The model maintains solid performance metrics in different assessment standards which include precision reaching 98.52% and recall alongside F1-score marking 98.51%. The model demonstrates practical potential when used clinically because it correctly identifies positive cases at a suitable low level of false positive detection.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

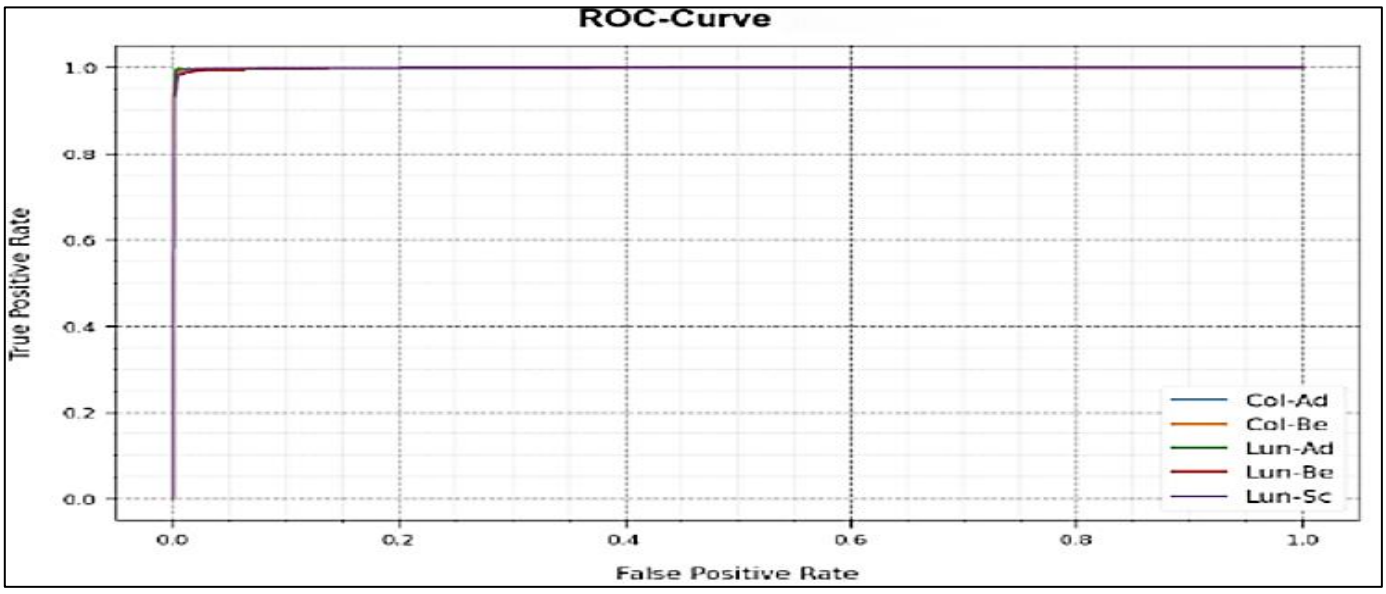

Fig 5: ROC Curve for CAE Model

The ROC curve in Figure 5 showcases the True Positive Rate against the False Positive Rate for five different disease categories: Col-Ad, Col-Be, Lun-Ad, Lun-Be, and Lun-Sc. The model delivered outstanding performance because all categories reported AUC scores near the perfect value of 1.0. The distinguished detection between positive and negative outcomes leads to solid validation that machine learning with medical data enables precise early-stage disease prediction. This method demonstrates consistent diagnosis potential which extends to multiple disease types for diverse medical applications.

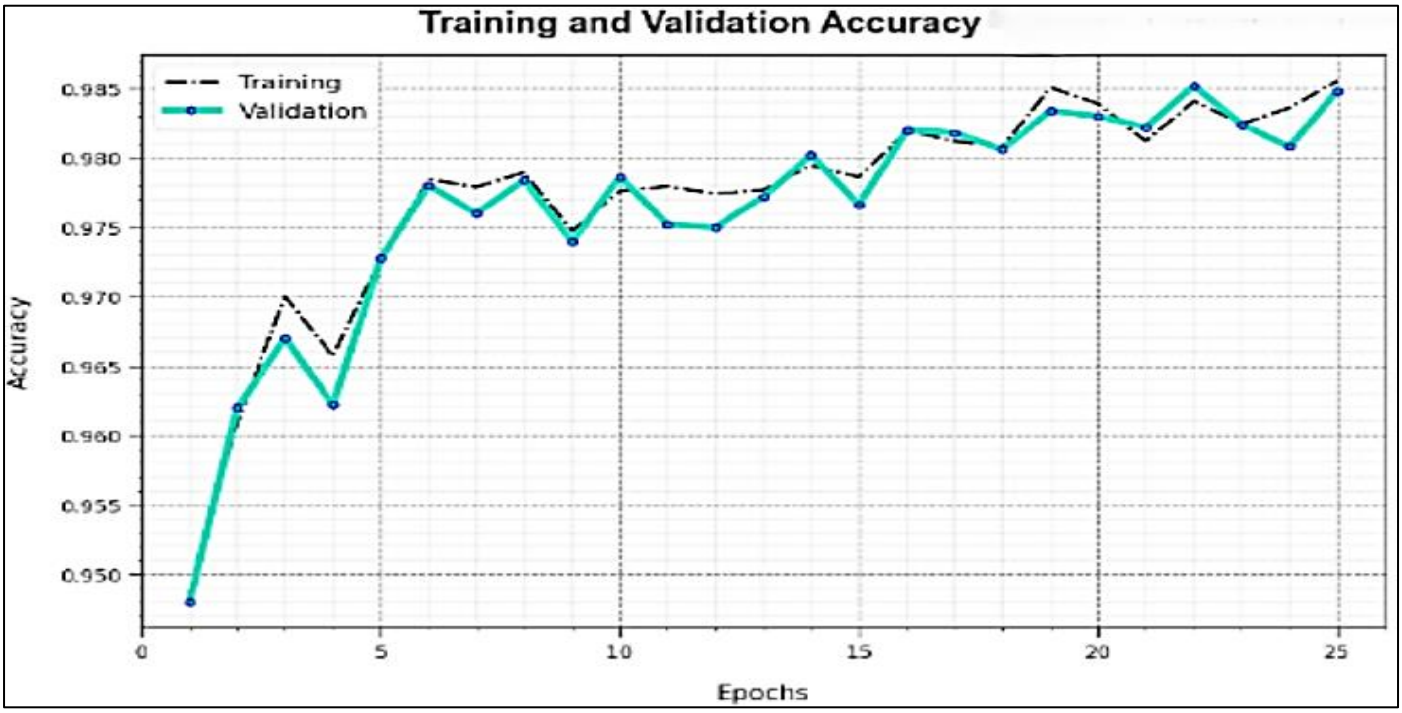

Fig 6: Training and Validation Accuracy for CAE Model

The illustration in Figure 6 shows how a model learns early disease prediction through its training and validation accuracy evolution during 25 epochs. Training accuracy (dashed line) shows a continuous rise while following every point of the validation accuracy (solid line) which indicates proper learning without major overfitting. The model reaches high performance since both accuracy values plateau at 0.98 after roughly 15 epochs. Medical data paired with ML has the potential to provide accurate disease prediction and early classification, since the model continues to exhibit consistent and dependable performance even when fresh data is included.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

Fig 7: Training and Validation Loss for CAE Model

The evaluation of early disease prediction requires examination of model performance through training and validation loss across 25 epochs, as displayed in Figure 7. Both training loss indicated by dashed line and validation loss shown by solid line decrease during the process demonstrating that learning progresses normally. The validation loss maintains a close relationship with training loss indicating minimal model overfitting. The variations in validation loss demonstrate how sensitive the model becomes to changes in medical data. A model that effectively generalizes unknown data is vital for precise early illness predictions with medical data, while a final loss plateau indicates convergence at 0.06.

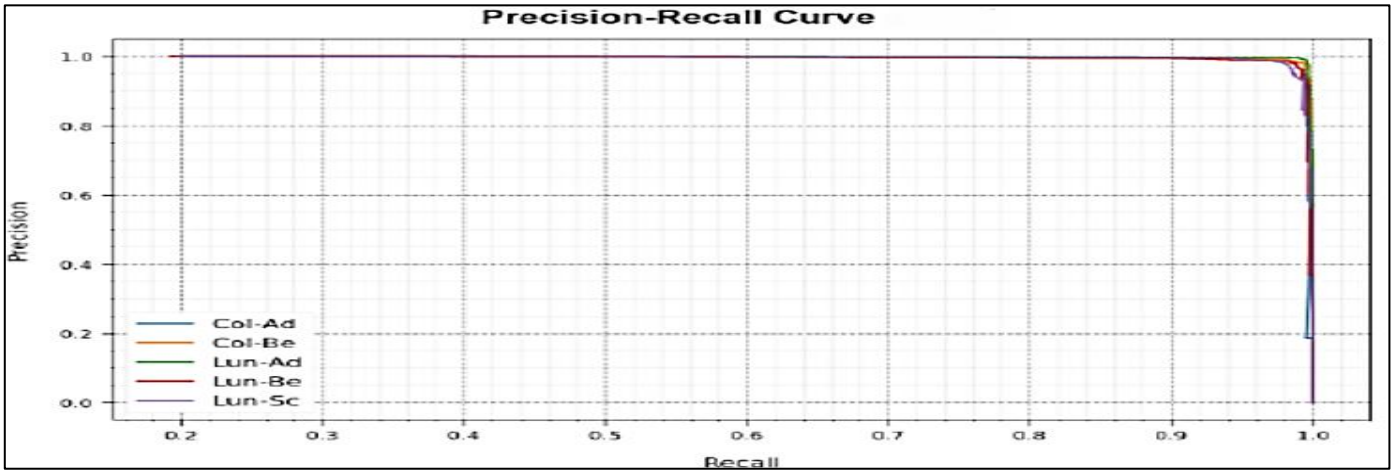

Fig 8: Precision-Recall Curve for CAE Model

Figure 8 demonstrates the performance of a model in lung cancer prediction. The graph displays precision against recall for five disease categories: Col-Ad, Col-Be, Lun-Ad, Lun-Be, and Lun-Sc. High recall and accuracy are evident across the board as the curves for all categories cling tightly to the top-right corner. This points to the model's excellent recall and precision, which indicate its good ability to detect positive cases reliably. The consistent performance across different disease types underscores the potential of this model for reliable and accurate early disease prediction leveraging medical data.

A. Comparative Analysis and Discussion

Evaluate the suggested and existing models on the LC25000 dataset. As stated in Table III, the re-call, accuracy, precision, and f1score were among the performance measures utilized to evaluate the aforementioned model (CNN, VGG19, ResNet-50, and CAE).

Table 3: Comparative Analysis between Proposed and Existing Models Performance for Lung Cancer Detection

Model	Accuracy	Precision	Recall	F1-score
CAE	99.41	98.52	98.51	98.51
CNN[32]	97	97	97	97
VGG19[33]	97.73	97	97	97
ResNET-50[34]	93.91	95.74	96.77	96.26

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1858

Table III presents a comparison between CAE, CNN, VGG19, and ResNet-50 on the LC25000 dataset. In this comparison, Convolutional Autoencoder model attains the best accuracy 99.41%, precision 98.52%, recall 98.51%, and F1-score 98.51%, indicating its exceptional classification proficiency. The CNN model exhibits a commendable accuracy of 97%, while sustaining equilibrium in precision, recall, and F1-score. VGG19 marginally surpasses CNN with an accuracy of 97.73%, while preserving precision, recall, and F1-score at 97%. ResNet-50, albeit exhibiting high precision 95.74% and recall 96.77%, attain the lowest accuracy at 93.91%, suggesting marginally reduced classification dependability. CAE outperforms state-of-the-art models, demonstrating significant efficacy in early illness prediction from histopathology pictures.

The proposed CAE model offers several advantages in lung cancer detection, including automated feature extraction, high accuracy of 99.41%, and robust generalization across cancer types. Compared to CNN, VGG19, and ResNet-50, it demonstrates superior classification performance with minimal overfitting. The model effectively reduces human error and enhances diagnostic reliability by leveraging DL for precise medical image analysis. Faster processing is guaranteed by its computational efficiency, making it ideal for real-time applications. The model achieves both precision and recall levels at 98.52% and 98.51% which indicates its potential for early disease identification that better supports clinical diagnosis and patient results.

V. CONCLUSION & FUTURE WORK

Predictability is necessary to lower the death rate from lung cancer, one of the most lethal diseases. Lung cancer stands as the world's second most fatal cancer despite having no clear indicators during its initial development phase. Accurate and speedy detection of lung cancer plays a fundamental role in improving patient medical results. The research explores deep learning analysis through the Convolutional Autoencoder (CAE) to detect lung cancer in histopathology images. Based on the evaluation of the LC25000 data model achieved 99.41% accuracy with 98.52% precision and re-call at 98.51% and an F1-score reaching 98.51%. Tests using ResNet-50, CNN and VGG19 establish that CAE holds the best capabilities for early disease identification. The implementation of this model depends on using a solitary dataset while the computational requirements remain high because training data exhibits restricted variability. The model needs additional testing to demonstrate its capability for generalizing between different imaging systems and clinical field applications. Future work will focus on enhancing model robustness by integrating transfer learning, expanding datasets with multi-source medical images, and employing explainable AI techniques to improve interpretability.

REFERENCES

[1]. K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, and T. Suryawanshi, “Human Disease Prediction using Machine Learning Techniques and Real-life Parameters,” *Int. J. Eng.*, vol. 36, pp. 1092–1098, 2023, doi: 10.5829/IJE.2023.36.06C.07.

[2]. V. Kolluri, “An Innovative Study Exploring Revolutionizing Healthcare with AI: Personalized Medicine: Predictive Diagnostic Techniques and Individualized Treatment,” *J. Emerg. Technol. Innov. Res. (*, vol. 3, no. 11, 2016.

[3]. Suhag Pandya, “Integrating Smart IoT and AI-Enhanced Systems for Predictive Diagnostics Disease in Healthcare,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 6, pp. 2093–2105, Dec. 2023, doi: 10.32628/CSEIT2410612406.

[4]. M. H. A. S. Ashish Shiwlani, Sooraj Kumar, Samesh Kumar, Syed Umer Hasan, “Transforming Healthcare Economics: Machine Learning Impact on Cost Effectiveness and Value-Based Care,” *Pakistan J. Life Soc. Sci.*, 2024.

[5]. M. T. Arora, Rajeev and Kumar, Shantanu and Jain, Nitin and Nafis, “Revolutionizing Healthcare with Cloud Computing: Superior Patient Care and Enhanced Service Efficiency,” *SSRN*, 2022, doi: http://dx.doi.org/10.2139/ssrn.4957197.

[6]. S. Pandya, “Predictive Modeling for Cancer Detection Based on Machine Learning Algorithms and AI in the Healthcare Sector,” *TIJER – Int. Res. J.*, vol. 11, no. 12, 2024.

[7]. A. Hage Chehade, N. Abdallah, J.-M. Marion, M. Oueidat, and P. Chauvet, “Lung and colon cancer classification using medical imaging: A feature engineering approach,” *Phys. Eng. Sci. Med.*, vol. 45, no. 3, pp. 729–746, 2022.

[8]. M. Masud, N. Sikder, A. Al Nahid, A. K. Bairagi, and M. A. Alzain, “A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework,” *Sensors (Switzerland)*, vol. 21, no. 3, pp. 1–21, 2021, doi: 10.3390/s21030748.

[9]. M. M. R. Said *et al.*, “Innovative Deep Learning Architecture for the Classification of Lung and Colon Cancer From Histopathology Images,” *Appl. Comput. Intell. Soft Comput.*, vol. 2024, no. 1, p. 5562890, 2024.

[10]. V. Kolluri, “The Impact of Machine Learning on Patient Diagnosis Accuracy:Investigating the Accuracy and Efficiency of Machine Learning Models in Diagnosing Diseases,” *J. Emerg. Technol. Innov. Res.*, vol. 11, no. 1, 2024.

[11]. V. Kolluri, “AI for Personalized Medicine: Analyzing How AI Contributes to Tailoring Medical Treatment to the Individual Characteristics ff Each Patient,” *IJRAR - Int. J. Res. Anal. Rev. (IJRAR), E-ISSN 2349-5138*, 2023.

[12]. S. N. Tisha and S. A. Ani, “Predictive Insights: Empowering Early Detection of Lung Cancer Using Machine Learning Excellence,” in *2024 IEEE Region*

10 Symposium (TENSYMP), 2024, pp. 1–6. doi: 10.1109/TENSYMP61132.2024.10752136.

[13]. S. Sathishkumar and P. Parameswari, "Optimizing Early Lung Cancer Detection with Divide and Conquer Kernel SVM and Radial Basis Function Neural Network enchanced by SMOTE," in *2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES)*, IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/IC3TES62412.2024.10877434.

[14]. D. Singh, A. Khandelwal, P. Bhandari, S. Barve, and D. Chikmurge, "Predicting Lung Cancer using XGBoost and other Ensemble Learning Models," in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, 2023. doi: 10.1109/ICCCNT56998.2023.10308301.

[15]. A. Vishwakarma, A. Saini, K. Guleria, and S. Sharma, "An Early Prognosis of Lung Cancer using Machine Intelligence," in *2023 International Conference on Artificial Intelligence and Applications, ICAIA 2023 and Alliance Technology Conference, ATCON-1 2023 - Proceeding*, 2023. doi: 10.1109/ICAIA57370.2023.10169432.

[16]. B. S, P. R, and A. B, "Lung Cancer Detection using Machine Learning," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, May 2022, pp. 539–543. doi: 10.1109/ICAAIC53929.2022.9793061.

[17]. M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *2022 IEEE World AI IoT Congress, AIIoT 2022*, 2022. doi: 10.1109/AIIoT54504.2022.9817326.

[18]. M. C. Keerthana and B. Azhagusundari, "Local binary fitting Median Filter for noise reduction in lung image datasets and classification".

[19]. B. Boddu, "Scaling Data Processing with Amazon Redshift Dba Best Practices for Heavy Loads," *Int. J. Core Eng. Manag.*, vol. 7, no. 7, 2023.

[20]. H. Zhao, C. Yang, W. Guo, L. Zhang, and D. Zhang, "Automatic Estimation of Crop Disease Severity Levels Based on Vegetation Index Normalization," *Remote Sens.*, vol. 12, no. 12, 2020, doi: 10.3390/rs12121930.

[21]. L. Yong, L. Ma, D. Sun, and L. Du, "Application of MobileNetV2 to waste classification," *PLoS One*, vol. 18, no. 3, p. e0282336, 2023.

[22]. Suhag Pandya, "A Machine and Deep Learning Framework for Robust Health Insurance Fraud Detection and Prevention," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 1332–1342, Jul. 2023, doi: 10.48175/IJARSCT-14000U.

[23]. M. Mohan Tito Ayyalasomayajula, A. Tiwari, R. Kumar Arora, and S. Khan, "Implementing Convolutional Neural Networks for Automated Disease Diagnosis in Telemedicine," in *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 2024, pp. 1–6. doi: 10.1109/ICDCECE60827.2024.10548327.

[24]. R. Tandon, "Face mask detection model based on deep CNN techniques using AWS," *Int. J. Eng. Res. Appl.*, vol. 13, no. 5, pp. 12–19, 2023.

[25]. J. Q. Gandhi Krishna, "Implementation Problems Facing Network Function Virtualization and Solutions," *IARIA*, pp. 70–76, 2018.

[26]. E. Yagis, A. G. S. De Herrera, and L. Citi, "Convolutional autoencoder based deep learning approach for Alzheimer's disease diagnosis using brain MRI," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 486–491.

[27]. D. Thakur, S. Biswas, E. S. L. Ho, and S. Chattopadhyay, "Convae-lstm: Convolutional autoencoder long short-term memory network for smartphone-based human activity recognition," *IEEE Access*, vol. 10, pp. 4137–4156, 2022.

[28]. K. Ullah *et al.*, "Short-Term Load Forecasting: A Comprehensive Review and Simulation Study With CNN-LSTM Hybrids Approach," *IEEE Access*, vol. 12, no. July, pp. 111858–111881, 2024, doi: 10.1109/ACCESS.2024.3440631.

[29]. K. Uyar and A. İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Comput. Sci.*, vol. 120, pp. 588–593, 2017, doi: https://doi.org/10.1016/j.procs.2017.11.283.

[30]. R. Ochoa-Ornelas, A. Gudiño-Ochoa, J. A. Garcia-Rodriguez, and S. Uribe Toscano, "Lung and colon cancer detection with InceptionResNetV2: a transfer learning approach," *J. Res. Dev.*, vol. 10, 2024, doi: 10.35429/JRD.2024.10.25.1.13.

[31]. W. Cui *et al.*, "BMNet: A new region-based metric learning method for early Alzheimer's Disease identification with FDG-PET images," *Front. Neurosci.*, vol. 16, p. 831533, 2022.

[32]. M. Al-Mamun Provath, K. Deb, and K.-H. Jo, "Classification of lung and colon cancer using deep learning method," in *International Workshop on Frontiers of Computer Vision*, 2023, pp. 56–70.

[33]. S. Wadekar and D. K. Singh, "A modified convolutional neural network framework for categorizing lung cell histopathological image based on residual network," *Healthc. Anal.*, vol. 4, p. 100224, 2023, doi: https://doi.org/10.1016/j.health.2023.100224.

[34]. S. U. K. Bukhari, A. Syed, S. K. A. Bokhari, S. S. Hussain, S. U. Armaghan, and S. S. H. Shah, "The histological diagnosis of colonic adenocarcinoma by applying partial self supervised learning," *MedRxiv*, pp. 2008–2020, 2020.