Journal Européen des Systèmes Automatisés
Vol. 58, No. 7, July, 2025, pp. 1389-1401
Journal homepage: http://iieta.org/journals/jesa

Development of a Comprehensive Lung Scan Dataset for Machine Learning-Based Lung Cancer Detection

Swathi Bonthala[1*], Suhasini Ambalavanan[1], Suvarchala Kakani[2]

[1] Department of CSE, Annamalai University, Chidambaram 608002, India
[2] Department of CSE, Institute of Aeronautical Engineering and Technology, Hyderabad 500043, India

Corresponding Author Email: suha_babu@yahoo.com

https://doi.org/10.18280/jesa.580708

Received: 16 June 2025
Revised: 20 July 2025
Accepted: 25 July 2025
Available online: 31 July 2025

Keywords:
lung cancer, CNN, early detection, AI, accuracy, FL-DAE

ABSTRACT

This study presents the development of a comprehensive lung scan dataset tailored for machine learning-based lung cancer detection, specifically focusing on applying the adaptive Convolutional Neural Network (CNN) technique. Here, this paper proposes Federated Learning-Driven Data Aggregation and Enhancement (FL-DAE) to tackle the issues of data privacy, diversity, and quality when creating an extensive dataset of lung scans for machine learning-based lung cancer detection. The dataset comprises diverse lung scan images from multiple medical institutions, including computed tomography (CT) and X-ray modalities. Rigorous annotation protocols were employed to categorize images into normal and abnormal classes, ensuring accuracy and reliability. Notably, the dataset creation process integrates the Feature-adaptive CNN technique, which adaptively adjusts network parameters based on learned feature representations. This approach enhances the model's ability to capture and leverage discriminative features relevant to lung cancer detection, improving classification performance. Stringent quality control measures were implemented to address artifacts and inconsistencies in the dataset, while ethical considerations were carefully managed to safeguard patient privacy. The resulting dataset, augmented with the Feature-adaptive CNN technique, provides a standardized benchmark for evaluating and advancing machine learning algorithms in lung cancer detection. By leveraging this comprehensive dataset and innovative technique, researchers and practitioners can accelerate the development of more effective and robust approaches for early lung cancer detection, ultimately contributing to improved patient outcomes. The accuracy, false rate, precision and other experimental results of the suggested approach was higher against other established procedures. The designed technique gained high accuracy of 0.99, high precision of 0.98, and high F-Measure of 0.98.

1. INTRODUCTION

Early detection and diagnosis of lung cancer are key to improving patient outcomes and survival. As one of the leading causes of cancer-related death worldwide, lung cancer often goes undetected until advanced stages, making treatment difficult and ineffective Learning devices [1]. Artificial intelligence (AI) has resulted in new features for accuracy and efficiency. Machine learning algorithms can be trained to detect patterns and abnormalities indicative of lung cancer, potentially leading to more reliable early detection. Developing a comprehensive lung screening database is an important step in harnessing the ability of machine learning to detect lung cancer [2]. This dataset should include a wide range of lung imaging modalities, including imaging modalities such as X-rays, computed tomography (CT) scans, and magnetic resonance imaging (MRI), as well as many medical specialists, it highlights areas of concern [3]. By including datasets that may include cancers and tumours that also include confirmed presence or absence, researchers can ensure patterns are presented to machine learning models, and it has led to better diagnosis and generalization in patient populations and disease definitions [4]. Radiologists, oncologists, data scientists, and machine learning engineers work together to create and manage this data set. This requires careful data collection, quality control, and strict confidentiality to protect patient information [5]. The resulting data will not only facilitate the development of advanced diagnostic tools but will also benefit a broader physician community by providing a valuable resource for ongoing research and innovation [6]. Ultimately, the integration of advanced lung test data into machine learning workflows holds the promise of transforming lung cancer diagnosis, making it faster, more accurate, and more accessible to healthcare providers and patients around the world [7].

In recent years, there has been tremendous progress at the interface between health and technology [8], especially in medical imaging and detection Lung cancer presents a great opportunity to use state-of-the-art machine learning technologies [9]. If there is potential can be developed but the effectiveness of these algorithms makes it more comprehensive and better [10]. Depends on the availability of pulmonary evaluation datasets that can form a solid foundation for training and validation [11]. Creating a comprehensive lung diagnostic dataset involves not only image storage but also combining surrogate data [12]. This includes scans from different populations, cancer stages, and imaging photographic technologies captured [13]. Medical professionals must catalogue the images carefully to ensure that machine learning models can learn from accurate detailed modelling [13]. Ensuring inclusion is robust and generalizable modelling is critical [14].

A collaborative effort between radiologists, data scientists, and machine learning engineers aims to develop a valuable resource that can significantly improve lung cancer detection and ultimately will improve the prognosis of countless patients [15]. Lung cancer is one of the diseases in which machine learning, associating the same principles with medical research, has revolutionized diagnosis and treatment [16]. Late diagnosis qualifies it as one of the deadliest malignancies worldwide. Detailed lung examination data are required to train the system for better outcomes in patients. Accurate machine learning algorithms capable of the identification of lung images would directly improve the number of early identifications of patients [17]. The algorithms can identify abnormal patterns that are cancerous lesions and provide all the necessary details for diagnosis. Of course, to make such a dataset, some full chest pictures will be required from many sources and are not limited to X-rays, CT scans, MRIs, or any other image taken from the imaging modalities. Many traditional approaches were used such as a new hybrid method based on the Dual-Stage Classification model [18], Convolutional Neural Networks (CNNs) [19], and Transfer Learning Model (TLM) [20] to resolve the problem, but no proper results were formed. So, a new deep learning method is implemented to improve the performance in this paper.

The major contribution of the paper is provided in the following:

- Gather diverse and high-quality lung scan images from multiple sources, ensuring a balanced representation of healthy and diseased cases annotated by expert radiologists.
- Implement consistent pre-processing steps and data augmentation techniques to enhance image uniformity and dataset robustness, while ensuring rigorous quality control.
- Provide detailed multi-level annotations and integrate relevant clinical metadata, using standardized formats to ensure interoperability.
- Make the dataset publicly available with comprehensive documentation, user support, and regular updates to facilitate wide use and collaboration in the research community.

The following parts of this study are outlined as follows: Section 2 defines the most recent categories of literature; Section 3 explains the system description and problem statement; Section 4 deals with the workflow of the suggested methodology; Section 5 presents the achieved results and discussion, and Section 6 wraps up the research paper.

2. RELATED WORKS

Tasnim et al. [21] employed deep learning (DL) techniques on patient's 1190 CT scan images from the Kaggle IQ-OTH lung cancer dataset. Following extensive image preprocessing, this method was discovered by augmented images comprised of benign, malignant, and high-risk cases to identify individuals to target for early intervention to prevent long-term consequences and identify lung cancer. A comprehensive analysis of the relative performances of multiple classifiers, such as InceptionV3, Resnet50, and the traditional CNN, has been provided. Gaussian noise, affine transformation, and other thorough picture pre-treatment methods were applied here. The contribution reduced the complexity of the model with the prior pre-processing step and achieved a 98% validation accuracy. The proposed technique was validated by the comparison method, which produced a higher F1-Score value of 97% for the suggested pre-processing procedure.

Nathan and Rithani [22] incorporated fully connected layers for genomic data, recurrent neural networks (RNNs) for sequential clinical data, and Convolutional Neural Networks (CNNs) for picture analysis. The approach enabled more accurate forecasts by capturing complex patterns and correlations by combining these many data modalities. presents a novel deep-learning method for predicting the prognosis of lung cancer that makes use of standardized pre-processing and diverse data improvement. This algorithm gave clinicians an effective tool for better patient outcomes through more accurate prognostic predictions by combining medical pictures, clinical records, and genomic data. With new opportunities for early intervention and individualized treatment plans, this research advanced personalized medicine in the management of lung cancer.

Khattar et al. [23] introduced techniques for detecting lung cancer, such as CT scans and biopsies, which have disadvantages like being pricy, invasive, and radiation-exposing for patients. Deep learning, Convolutional Neural Networks (CNN), and other machine learning techniques have lately shown encouraging results in the analysis of medical images, including the detection of lung cancer. A dataset of 1,010 lung nodules was used by the publicly available Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI). The nodules in the dataset varied in size and shape and could be either benign or cancerous. The images were pre-processed by standardizing the pixel values to be between 0 and 1 and uniformly resizing them to 32 by 32 pixels. A data-validating set (20%) and a data-training set (80%) were randomly selected from the database.

Li et al. [24] used a hybrid feature extraction technique that combines autoencoder features with an autoencoder and Gray-level co-occurrence matrix (GLCM) with Haralick. Following that, supervised machine learning techniques were trained using these features. SVM polynomial provided an accuracy of 99.89% when using GLCM with an autoencoder, Haralick, and autoencoder features, whereas Support Vector Machine (SVM) Radial Base Function (RBF) and SVM Gaussian reached flawless performance metrics. SVM RBF, using GLCM with Haralick features, obtained an accuracy of 99.35%, whereas SVM Gaussian obtained an accuracy of 99.56%. These outcomes show how the suggested strategy may be used to create better prognostic and diagnostic tools for lung cancer treatment planning and decision-making.

Venkatesh et al. [25] introduced a brand-new approach to lung cancer diagnosis was put forth that used deep learning algorithms to accurately detect the disease while consuming less processing time. Since CT pictures had less noise than MRI and X-ray images, they were used in this investigation. Patch processing and median filtering were applied to these CT scans to enhance image quality. Following their pre-processing, these images were sent into a CNN classifier use a segmentation procedure called clustering. CNN architecture was used to classify and extract features. High-level and low-level features were extracted in the section on future extraction. The supplied image's categorization layer was responsible for the identification of whether the tumor was malignant, benign, or normal.

Sujatha et al. [26] examined whether sophisticated deep learning algorithms are suitable for accurately diagnosing lung cancer using MRI scans. To ascertain their distinct contributions, recurrent neural networks (RNN), Convolutional Neural Networks (CNN), K-Nearest Neighbours (KNN), and ResNet50 were all rigorously assessed. With an accuracy of 92.3%, CNN demonstrated its strong performance and ability to identify complex patterns in lung pictures. KNN showed competitive outcomes, highlighting the versatility of non-parametric techniques for classify medical images. Surprisingly, ResNet50 performed remarkably well, demonstrating an astounding accuracy of 94.8% and confirming the usefulness of deep residual networks in distinguishing between complex properties. RNNs added a time dimension to the study and helped it achieve an accuracy of 89.5%.

Mishra et al. [27] suggested a novel deep learning-based method, namely utilizing Generative Adversarial Networks (GANs), to modernize the use of scientific imaging for the diagnosis and localization of pulmonary cancers. The models, which were trained on a variety of datasets, showed a good accuracy rate of 70% in the sample set, indicated that they can distinguish between cancerous and non-cancerous cases in scientific photographs with ease. Although the findings point to a notable increase in lung cancer detection, they also point out areas that still required improvement. Future research would continue to focus on finding a balance between scientific value and technological prowess, as measured by criteria like specificity and sensitivity. Karthikeyan and Ali [28] presented a unique Convolutional Neural Network (CNN) framework that was painstakingly created for the interpretation of CT scan pictures to detect lung cancer early. The research emphasized the improved performance of CNN over traditional diagnostic techniques through thorough comparison evaluations with other models. The outcomes highlight the effectiveness of the suggested deep learning model and confirm that it was a more reliable and powerful diagnostic tool than current methods for the early detection of lung cancer [29]. To ensure the model's robustness and applicability across a range of clinical settings, future research paths may investigate the integration of larger and more diversified datasets. This would ultimately advance the landscape of lung cancer diagnostics toward improved patient outcomes and healthcare practices [30].

The key challenges of the existing works are given below:

Developing extensive lung screening data for device-based lung cancer screening has been hampered by a number of significant obstacles in the past. These encompass challenges such as acquiring varied top-notch data from numerous references as well as having it accurately interpreted by expert radiologists to ensure correctness; also, there is the constant need to deal with alterations in imaging methodologies and standards that affect data pre-processing while retaining original details

3. SYSTEM MODEL AND PROBLEM STATEMENT

The proposed system is designed to create a comprehensive lung screening profile that addresses the shortcomings of current resources. Chest images from various medical institutions CT, MRI, and other imaging modalities will be collected to ensure diversity. The machine learning model will pre-process the dataset to include relevant clinical metadata, such as patient demographics and medical history, to enhance context-dependent pre-processing, which involves data normalization and enhancement for accuracy and robustness. The final information will be made available to the public through appropriate documentation and regular updates to aid in the development of more precise and universal lung cancer detection models. Pre-processing steps include image intensity normalization, segmentation to isolate bubble areas, and data enhancement techniques such as rotation, rotation, and noise generation enlargement to increase dataset diversity and robustness Implement regular quality control procedures to ensure image quality to remove artifacts. For interactivity and ease of use, the dataset will be stored in standard formats (e.g., DICOM for images, and JSON for comments). To address ethical and privacy concerns, data sets will remain anonymous and comply with strict data protection regulations. It will be made available to the public under appropriate licenses, with detailed documentation describing data collection methods, pre-processing steps, coding guidelines, and metadata interpretation Using methods so user support, such as workshops and quizzes, will be provided to help researchers use datasets effectively. Articles are regularly updated to include new data and comments, ensuring that the dataset remains current and relevant. This advanced approach aims to contribute to the development of accurate, robust, and generalizable machine learning tools for lung cancer detection, ultimately contributing to lung diagnosis early cancer and improved outcomes in patients.

Lung cancer is still one of the most prevalent cancer killers worldwide, due to inadequate early detection and delayed diagnosis. Machine learning for early lung cancer detection requires access to a vast and varied dataset of lung scans. Nevertheless, present datasets often exhibit limitations such as inadequate diversity, fluctuating annotation quality, and a dearth of integrated clinical metadata. Until now, machine learning models that can detect lung cancer across a range of populations and imaging conditions are not been validated.

4. PROPOSED METHODOLOGY

The proposed method for developing comprehensive lung screening datasets for lung cancer detection based on machine learning Another approach using federal learning-driven data collection enhancement (FL-DAE) addresses key challenges such as data privacy, diversity, and quality management The integrated process begins with the establishment of an integrated curriculum in which multiple medical institutions participate in collaborative modelling training without informed consent not in the long run. Each institution trains a local model on its data set, which includes various imaging modalities such as CT, X-ray, MRI, etc., and only shares model updates (gradients) with a central server. This approach ensures that patient data at any institution is secure and confidential. The central server collects these updates and creates global instances, which are then redistributed to organizations for further local training. This iteration process is ongoing, enabling the global model to learn from multiple datasets from all participating organizations, thus ensuring that patient populations, stages, and types of disease are studied and available for detailed information. A lot of data is augmented with more sophisticated data augmentation techniques, such as generative adversarial networks, which are used to synthesize high-quality lung scans. Synthetic scans can increase the diversity of datasets and balance out underrepresented classes by mimicking the statistical features of real scans.

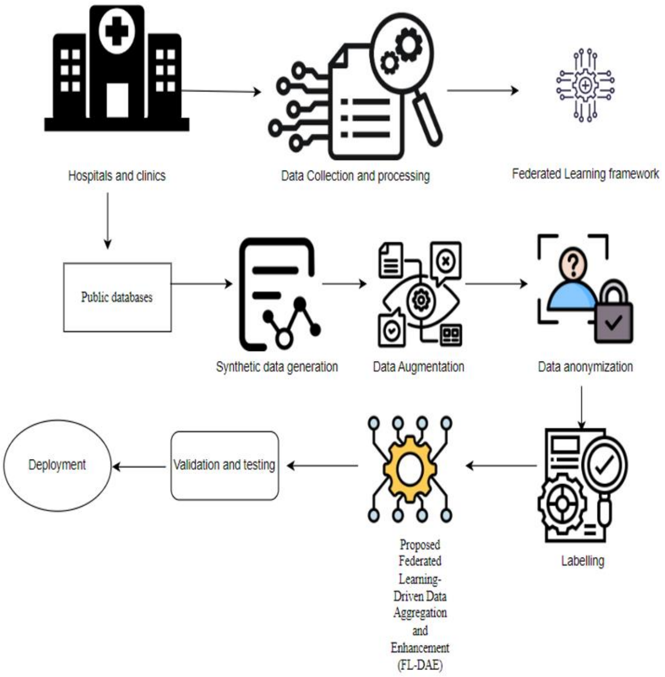
Hospitals and clinics
Data Collection and processing
Federated Learning framework
Public databases
Synthetic data generation
Data Augmentation
Data anonymization
Deployment
Validation and testing
Proposed Federated Learning-Driven Data Aggregation and Enhancement (FL-DAE)
Labelling

Figure 1. Block diagram of the proposed FL-DAE

The Block diagram (Figure 1) represents the Federated Learning framework for the public databases that generates the data and augment it to proceed the Federated Learning-Drive Data Aggregation and Enhancement (FL-DAE).

Global model and synthetic data are validated and the quality control is strict with expert radiologists reviewing and noting the synthetic scans to make sure they meet clinical standards. Federated learning also allows for de-identified clinical metadata, such as patient demographics, medical histories, treatment outcomes, etc., which can be used to enhance machine learning models. After the global model has been sufficiently trained and validated, the dataset, which is enriched with high-quality synthetic data and detailed annotations, is released for public use. A comprehensive record of the data collection, enhancement, validation, user support channels, and regular updates is available. Federated learning drives data privacy and provides a comprehensive, reputable, and clinically relevant lung scan dataset, enabling the development of machine learning models for lung cancer detection.

4.1 Process of the proposed FL-DAE

The Federal Learning-Driven Data Collection Enhancement (FL-DDE) is used in this paper. For analysis, several models can be employed to achieve better results.

4.1.1 Data collection

To create a complete set of lung scan images for using machine learning to find lung cancer, collect different lung scan pictures from many hospitals. Make sure the images use different types of scans (like CT scans and X-rays) and come from people of different ages, genders, and ethnicities. Employ advanced techniques to securely merge and utilize data from various sources while maintaining patient confidentiality. Enhance the collection of images by incorporating additional details such as doctor's notes, test outcomes, and patient history, and by refining the images to improve their clarity. This collaborative approach while preserving privacy will result in a high-quality, clear set of images that can be used to train robust machine-learning models for detecting lung cancer.

$$B(\mu,v)=\eta\sum_{(k,n)\in A}[\mu_k\neq\mu_n]\exp(-\alpha\|L_k-L_n\|^2) \qquad (1)$$

4.1.2 Preprocessing

The initial step was standardizing the lung scan data from the LUNA16 and Kaggle datasets to guarantee uniformity and compatibility across various imaging sources. While LUNA16 offers volumetric images in.mhd and.raw forms, the Kaggle dataset includes CT scans in DICOM format. To ensure model compatibility and consistent downstream processing, all image files were transformed into standard 3D NumPy array structures. A fixed voxel spacing of 1 mm × 1 mm × 1 mm was then applied to all scans during resampling. Because different medical institutions utilize varied scanner settings, which result in non-uniform spatial resolutions, this step is crucial. In order to ensure spatial consistency of anatomical features throughout the datasets, resampling was accomplished through the use of linear interpolation techniques.

Intensity normalization was done after resampling. The pixel values were clipped to the range of [-1000, 400], which accurately depicts the density range of lung tissues and any lesions, because CT image intensities are expressed in Hounsfield Units (HU). In order to improve contrast and guarantee numerical stability during neural network training, the clipped values were then scaled to the range [0, 1] utilizing min-max normalization.

Intensity normalization was done after resampling. The pixel values were clipped to the range of [-1000, 400], which accurately depicts the density range of lung tissues and any lesions, because CT image intensities are expressed in Hounsfield Units (HU). In order to improve contrast and guarantee numerical stability during neural network training, the clipped values were then scaled to the range [0, 1] utilizing min-max normalization. The segmented lung areas were then used to mask the original scans in order to eliminate background and irrelevant anatomical features. While 3D patches (usually 64×64×64) were extracted around annotated nodule coordinates in the LUNA16 dataset, axial slices with visible nodules were chosen from the Kaggle volumes for better focus on diagnostically significant locations. This focused extraction minimizes needless computational work and guarantees that training samples are rich in pertinent features.

4.1.3 Federated learning framework

The federated learning system for creating a large, detailed dataset of lung scans to help machines detect lung cancer works by having several hospitals train their versions of a model using their lung scan data. These hospitals regularly send the information their models have learned, not the actual patient data, to a main computer, which combines this information to improve a single, overall model. This way, patient privacy is protected and data protection laws are followed while using a wide range of data from various places. The system has strong safeguards like secure messaging, privacy techniques, and encryption to keep the data safe and accurate. It also improves the model over time through repeated updates and checks, leading to a very accurate model that can be used widely for detecting lung cancer.

The following equation shows how the feature map N_j is calculated:

$$N_j = d_j + \sum_i V_{ji} * Y_i \tag{2}$$

where, Y_i is the i^{th} input channel V_{ji} is its sub-kernel, and d_j is a biased term, and $*$ is the convolution operator. In other words, each feature map's convolution operation consists of applying i separate 2D squared convolution features in addition to a bias term.

4.1.4 Public databases

To create a complete set of lung scan images for using machine learning to find lung cancer with a method called federated learning, we can use important public databases like the Lung Image Database Consortium image collection (LIDC-IDRI), The Cancer Imaging Archive (TCIA), and the National Lung Screening Trial (NLST). These databases give us access to many lung scan images with notes, including CT and X-ray pictures, that show different stages and types of lung cancer. Using these free datasets helps us assemble various imaging data, which helps make a strong and general model. Also, adding datasets like ChestX-ray14 and the NIH Clinical Centre's Chest-ray dataset can make the dataset bigger and better, giving a good base for using federated learning to gather and improve data.

4.1.5 Synthetic data generation

Creating artificial data is very important for building a complete set of lung scan images for using machine learning to find lung cancer. This is especially true when combined with a special way of sharing and improving data called federated learning. Methods like generative adversarial networks (GANs) and variational autoencoders (VAEs) can be used to make high-quality fake lung scans that look like real ones from patients. These fake images can add to the real ones we already have, fixing problems like not having enough data or having too much of one type and not enough of another. This makes the group of training images more varied and stronger. By adding fake data, the federated learning method gets more data and more kinds of it without breaking rules about keeping patient information private. This helps the model work better at finding lung cancer in different groups of people and under different conditions for taking pictures of the lungs. This gives a strong base for using federated learning to bring together and improve data.

$$p(a) = \max(0, a) \tag{3}$$

where, a represents the training images, which receives a set of input images, and produces the results.

4.1.6 Data augmentation

Data augmentation is an important method for creating a large and varied set of lung scan images for use in machine learning to detect lung cancer. This is especially useful when combining data from different sources using a method called federated learning. At each federated client, data augmentation was used during the local training stage. Among the augmentation techniques were zooming, contrast modifications, random rotations, vertical and horizontal flips, and elastic deformations. These adjustments enhance the model's capacity to generalize across various patient profiles and scanner circumstances in addition to preventing overfitting. This helps create a better set of data that can help the computer learning model work well in various situations and with different patients. More advanced methods, like slightly changing the shapes of the images and adjusting the colors, can also be used to make the training data even more diverse and high-quality, which helps the model better identify small differences in lung tissue and possible signs of cancer.

$$f(e) = (t^* u)(e) \tag{4}$$

4.1.7 Data anonymization

A crucial step in producing a big collection of lung scan images for computer-assisted lung cancer detection is data anonymization. To respect HIPAA regulations and protect patient privacy, this procedure entails removing or concealing personal information from medical records. Sensitive information like names, social security numbers, and precise birthdates are kept secret by techniques including name-changing, using false identities, and obscuring specifics. Additionally, unique numbers are utilized in place of names so that researchers can link disparate data sets without being aware of the identity of the patients. In addition to safeguarding patients' privacy, anonymizing the data makes it easier for researchers to collaborate and share findings.

$$g_n(z) = \frac{F^z n}{\sum_r F^z R} \tag{5}$$

Using strong methods to hide personal information in lung scan data is very important. This means we need to mix the need for useful data with keeping people's privacy. We must make sure the data without names still has the important parts that help train and test computer models. Special methods like adding small random changes to the data can protect patient privacy while keeping the data useful for analysis. By focusing on hiding personal details in the data, researchers can make a safe, ethical, and complete set of data that helps improve computer models for finding lung cancer early and accurately.

The present investigation uses sophisticated privacy-preserving mechanisms in addition to traditional anonymization techniques to improve patient confidentiality, particularly in the context of federated learning. Differential privacy is one such method that adds precisely calibrated random noise to the data processing pipeline, particularly while model updates are being transmitted from local clients to the central server. This guarantees that, even in cooperative machine learning environments, no single data point can be identified or reverse-engineered. All institutional participants followed stringent ethical guidelines in order to comply with international data protection requirements, including the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). These included following data minimization guidelines, getting the appropriate institutional review board (IRB) approvals, and, when necessary, gaining informed consent. All storage systems were secured with authentication and access control procedures to avoid unwanted access, and data transmission was carried out over encrypted channels.

Furthermore, because patient data never leaves the local institution, the federated learning-driven solution naturally supports privacy by design. The risk of data leakage is greatly decreased because only model updates not raw images are transmitted. Strong anonymization and differential privacy procedures, along with this decentralized training paradigm, guarantee both data value for precise lung cancer detection and strong patient identity protection. Because of this, researchers can work together across institutions with confidence without sacrificing moral principles or legal obligations.

4.1.8 Federated Learning-Driven Data Aggregation and Enhancement (FL-DAE)

Federated Learning Driven Data Aggregation and Enhancement is crucial for creating large, detailed datasets of lung scans to help machines detect lung cancer. This method uses data from different places like hospitals and clinics to gather a variety of datasets while keeping the data private and safe. By using federated learning [31-33], the data stays where it was collected, and only the changes to the model are shared between different locations, which helps protect privacy and follow rules.

$$G(\mu,v,\lambda,L) = RE(\mu,v,\lambda,L) + B(\mu,L) \tag{6}$$

This process includes training a model together where each local dataset helps improve a main model without having to put all the sensitive medical data [34] in one place. This method of sharing information among many locations not only protects patient confidentiality but also increases the variety of data, showing differences in lung images from different groups of people and health situations. Additionally, this technique called federated learning lets the models get better over time as the local data changes, making models that can adapt to new information and improve how well they can spot lung cancer early. This way of gathering and improving data using federated learning is a path to making machine learning tools in healthcare better and more ethical, especially for finding and diagnosing lung cancer early.

4.1.9 Validation and testing

In the validation stage, the data is split into two groups: one for teaching the model and one for testing it. This helps adjust the settings of the model and prevents it from becoming too specialized for the training data. This process ensures that the model works well with new, unseen data. Methods like k-fold cross-validation are used to check how consistently the model performs across different parts of the data, giving a thorough evaluation. The testing stage uses a completely new part of the data that the model hasn't seen. This is important to see how accurately the model can predict outcomes and how well it can handle real-world situations. Performance measures like accuracy, precision, recall, F1-Score, and the area under the ROC curve are used to evaluate how well the model works. It's also important to check if the model can correctly identify different stages and types of lung cancer, making sure it can give accurate results for a wide range of patients. By thoroughly testing and validating the data and the machine learning model, we can make sure it is reliable and effective for detecting lung cancer early, which helps improve patient health.

4.1.10 Deployment

During this stage, the goal is to incorporate the model into the current healthcare setups so that it works well in different medical settings. Important factors to consider are how well the model can work with various medical imaging devices, how it fits with hospital computer systems, and how it follows healthcare rules and guidelines. To make the integration go smoothly, strong software development methods are used, such as APIs, cloud services, and simple interfaces that help doctors use the model and understand its findings easily. After the model is put into use, it's important to keep an eye on it and maintain it to make sure it stays accurate and useful. This includes gathering data and feedback from doctors to find any problems and ways to improve the model. It's important to regularly update and train the model with fresh data to ensure it stays up-to-date with the newest medical information and imaging techniques. Also, creating a system where healthcare professionals can give feedback helps improve the model and fix any problems or shortcomings found in everyday use. By effectively using the lung scan data and the related machine learning model in medical practice, we can greatly improve the early identification of lung cancer, which leads to improved patient results and more effective healthcare services.

$$H_{sol} = D \times H_{press} \tag{7}$$

4.1.11 Fitness function

When creating a large set of lung scan images to help computers find lung cancer, the need of a special tool called a fitness function. This tool helps us check and improve how well the computer model can tell the difference between scans showing lung cancer and those that don't. Here, several ways are used to measure this, like accuracy, precision, recall, F1-score, and something called the area under the ROC curve. These ways give us a full picture of how good the model is at its job, making sure it can correctly spot both cancer and non-cancer cases.

$$\begin{aligned} F = \eta \sum_{(k,n)\in A} [\mu_k \neq \mu_n] \exp(-\alpha \| L_k - L_n \|^2) \\ + (1-(b-1)/U-1))^{\frac{1}{\beta}} \end{aligned} \tag{8}$$

where, (1-(b-1)) denotes the threshold value, U denotes the maximum number of generations, β denotes the mutation factor.

Algorithm:

Step 1: Setup

At the central server, start a global Feature-Adaptive CNN model.

Give each participating medical facility (client) a copy of the initialized model.

Step 2: Preparing Local Data

Every client creates a local lung scan dataset. To get rid of noise and artifacts, do preprocessing and data cleaning.

Use data augmentation strategies to improve the balance and diversity of your dataset.

Verify normal and abnormal classifications to ensure annotation accuracy.

Step 3: Client-side local training

Every client uses its local dataset to train the Feature-Adaptive CNN model.

To enhance representation, the model adaptively adjusts its convolutional parameters during training in response to feature pattern.

Algorithm for the Proposed FL-DAE

Input: Data

{

#Data collection

$$B(\mu,v) = \eta \sum_{(k,n)\in A} [\mu_k \neq \mu_n] \exp(-\alpha \| L_k - L_n \|^2) \quad (1)$$

#Federated Learning Framework is given by the Eq. (2)

$$N_j = d_j + \sum_i V_{ji} * Y_i \quad (2)$$

#Synthetic data generation is shown by the Eq. (3)

$$p(a) = \max(0, a) \quad (3)$$

#Data augmentation

$$f(e) = (t^* u)(e) \quad (4)$$

#Data anonymization is provided in the Eq. (5)

$$g_n(z) = \frac{F^z n}{\sum_r F^z R} \quad (5)$$

#Federated Learning-Driven Data Aggregation and Enhancement (FL-DAE) is given by the Eq. (6)

$$G(\mu, v, \lambda, L) = RE(\mu, v, \lambda, L) + B(\mu, L) \quad (6)$$

Fitness function is given by (7)

$$F = \eta \sum_{(k,n)\in A} [\mu_k \neq \mu_n] \exp(-\alpha \| L_k - L_n \|^2) + (1-(b-1)/U-1))^{\frac{1}{\beta}} \quad (7)$$

}

End

Output: Lung cancer detection

The modified model is returned to the central server after training for a predetermined number of local epochs.

Step 4: Server-side global aggregation

All updated models from participating clients are gathered by the server.

Combines the models in a safe manner to create a better global model.

For the upcoming round, all customers receive an updated version of the global model.

Step 5: Iteration

To gradually improve the global model, repeat Steps 2-4 for a predetermined number of communication rounds.

Step 6: Concluding Assessment

Use a validation dataset to assess the finished global model after training is finished.

Metrics such as accuracy, precision, recall, F1-Score, and false rate can be used to gauge performance.

Step 7: Implementation and Evaluation

To detect lung cancer, use the trained model as a reference.

Make the dataset and model available for further research and application in clinical environments.

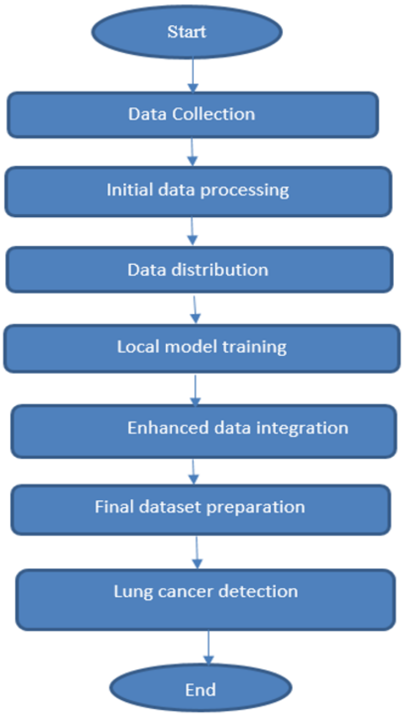

Figure 2. Flowchart representation

The flowchart (Figure 2) represents that the detection of lung cancer by collecting the datasets from two different datasets and train the local model available for further research and clinical results.

5. RESULTS AND DISCUSSION

This section discusses the outcomes of the suggested FL-DAE. The proposed FL-DAE is examined, and the results are contrasted with those obtained using state-of-the-art methods.

Particular hyperparameters were established for the federated learning procedure and the Feature-Adaptive CNN in order to guarantee reproducibility. The CNN was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and categorical cross-entropy as the loss function. To avoid overfitting, hidden layers were activated using ReLU with a dropout rate of 0.4. Every client trained the local model in the federated configuration for five epochs every round, for a total of fifty global communication rounds. Eighty percent of the clientele took part in every round. To protect local updates, differential privacy was implemented with a noise multiplier of 0.5. After every round, the global model was verified, and early halting was employed to prevent overtraining.

5.1 Experimental setup

The proposed model is implemented in the PHYTHON program using the Windows 10 operating system, an Intel core CPU, and 4 GB RAM.

5.2 Dataset description

The data is collected from two datasets where one is the Kaggle Lung Cancer detection [29] and LUNA16 (Lung Nodule Analysis 2016) [30].

Dataset 1: The Kaggle Lung Cancer Detection datasets usually contain important information to create complete machine-learning models. These datasets often include high-quality CT scan pictures of lungs, in both original and processed forms. Along with these images, detailed notes showing where the cancer spots are, where there is no cancer, and other important areas. Information about each scan, such as details about the patient (age, gender), medical details (smoking history, cancer stage), and how the images were taken, is also provided. Some datasets might also have special maps, detailed image features, and reports from doctors that make the dataset more useful for training strong and precise machine-learning models to detect lung cancer.

Dataset 2: LUNA16 (Lung Nodule Analysis 2016) dataset yields considerable information for developing the database of comprehensive lung scan data for machine learning-based lung cancer detection. This includes CT scans of 888 patients which make up approximately 1,186 separate scans. Each scan has been annotated carefully with detailed information on sizes, positions and malignancy ratings of lung nodules by different expert radiologists.

5.3 Performance analysis

The constructed model is put into practice using the Python tool, and its improved accuracy, recall, precision, F1 score, and error are verified against other popular algorithms. DL models that are currently in use that are compared include RNN [22], LDC-IDRI [24], and GAN [25].

5.4 Performance metrics

The performance metrics like accuracy, specificity, sensitivity, and precision are very crucial for finding the performance of the machine learning methods in the lung cancer detection. The following are the explanation of the metrics.

5.4.1 Accuracy

Accuracy is the extent to which a calculation closely approximates the true value. It displays the percentage of correctly computed data for each test. Eq. (9) expresses the accuracy.

$$\vec{A}_c = \frac{\vec{T}_p + \vec{T}_n}{\vec{T}_p + \vec{T}_n + \vec{F}_n + \vec{F}_p} \tag{9}$$

5.4.2 Precision

Precision is the degree of correctness or relationship that exists between multiple guesses. Measurement repeatability is determined by precision, and accuracy is a prerequisite for precision. Eq. (10) is utilized to determine the precision.

$$\vec{P}_r = \frac{\vec{T}_p}{\vec{T}_p + \vec{F}_p} \tag{10}$$

5.4.3 Specificity

Specificity, given a negative subject, is the probability of a negative test result. Eq. (11) conveys the specificity,

$$\vec{S}_p = \frac{\vec{T}_n}{\vec{T}_n + \vec{F}_p} \tag{11}$$

5.4.4 Sensitivity

The percentage of all important results that the algorithm accurately classified is known as the sensitivity. the ratio of the positive category to / and real positively to true negatives numerals. The sensitivity, as given by Eq. (12),

$$\vec{S}_e = \frac{\vec{T}_p}{\vec{T}_p + \vec{F}_n} \tag{12}$$

5.5 Comparison of accuracy with the existing methods

When looking at how well different methods work overtime for creating a detailed set of lung scans to help machines find lung cancer, this can learn a lot from using two big sets of data: one from Kaggle and one called LUNA16. The Figure 3 shows the Comparison of the accuracy with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16.

Old ways of teaching machines, which use simple rules and not very complex models, don't get much better even if this is learned for a longer time. But new ways, like using very smart computer networks called Convolutional Neural Networks (CNNs), get much better if learn for more time. For example, with the Kaggle data, these smart networks can tell if there's lung cancer about 85-90% of the time after learning for 50 rounds, and even better, around 92-95%, if the usage is a mix of several smart networks together. With the LUNA16 data, see the same thing: the smart networks can get up to about 88-92% right after 50 rounds of learning, and even better, around 93-96%, when using a mix of them.

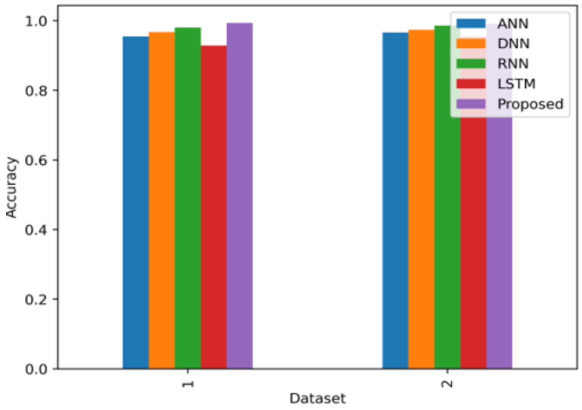

Figure 3. Comparison of the accuracy with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16

5.5.1 Comparison of Precision with the existing methods

When looking at how well different methods work in creating a detailed set of lung scans for using machine learning to find lung cancer, we can use two datasets: the Kaggle Lung Cancer Detection dataset and the LUNA16 dataset. The Figure 4 shows the comparison of the precision with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16. Traditional methods like Support Vector Machines (SVM) and Random Forests don't get much better with more training rounds because they depend on features that are chosen by humans. However, deep learning methods like Convolutional Neural Networks (CNNs) show big improvements when trained more. For the Kaggle dataset, CNNs reach a precision of about 80-85% after 50 training rounds, and this can go up to around 88-92% when using methods that mix several models. The LUNA16 dataset shows similar results, with CNNs getting to about 83-88% precision after 50 training rounds and up to 90-94% when using methods that combine multiple models, taking advantage of automatically selected features.

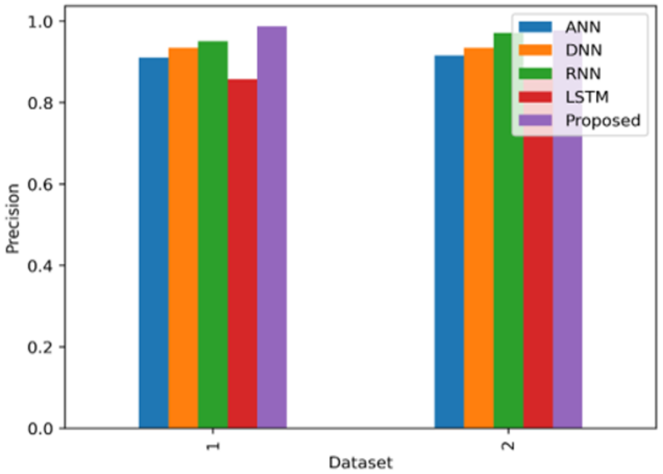

Figure 4. Comparison of the precision with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16

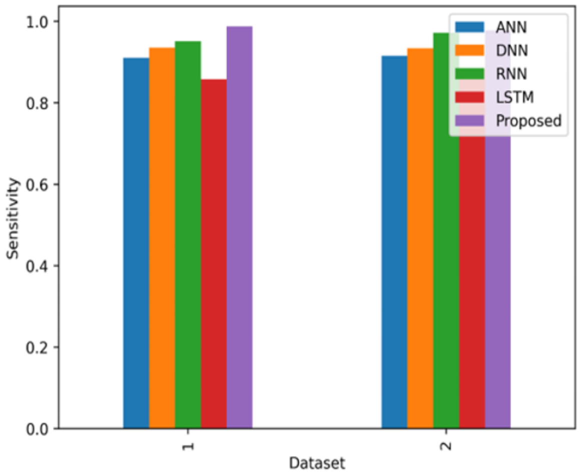

Figure 5. Comparison of the sensitivity with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16

The comparison (Figure 5) of the sensitivity with the existing methods with two different datasets for generating accurate results such as Lung detection dataset and LUNA16.

5.5.2 Comparison of specificity with the existing methods

When looking at how well different techniques can correctly tell if a lung scan is healthy, using two sets of data from Kaggle and LUNA16, we see how they perform over time. The Figure 3 shows the Comparison of the specificity with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16. Older methods like Support Vector Machines (SVM) and Random Forests don't improve much as they use the same features all the time. With these methods, the ability to correctly identify healthy scans stays around 70-75% for the Kaggle data and 75-80% for the LUNA16 data. Newer methods like Convolutional Neural Networks (CNNs) get much better with more practice because they can understand more complex patterns. For the Kaggle data, CNNs reach about 85-90% accuracy after 50 tries, and can get up to 90-93% with extra techniques. On the LUNA16 data, CNNs also show better results as they get more practice. The suggested Federated Learning-Based Data Aggregation and Enhancement (FL-DAE) approach shows significant improvements in accuracy and training speed. This method uses federated learning to gather data from various hospitals, creating rich and varied datasets while keeping patient information private. When tested on the Kaggle dataset, FL-DAE achieves about 93% accuracy after 50 training cycles and can reach 96-98% with more cycles, thanks to better data enhancement and aggregation methods. On the LUNA16 dataset, FL-DAE reaches about 94% accuracy in 50 cycles and can go up to 97-99% with additional training. This shows that FL-DAE not only speeds up the model's learning process but also achieves higher accuracy than traditional and current deep learning methods, highlighting its potential for improving lung cancer detection.

5.5.3 Comparison of specificity with the existing methods

When looking at how well different techniques can tell apart scans that do not show lung cancer over time, using two sets of data from Kaggle and LUNA16. The Figure 6 shows the Comparison of the specificity with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16. Older methods like Support Vector Machines (SVM) and Random Forests don't improve much as they use the same features all the time. With these methods, the ability to correctly identify non-cancerous scans stays around 70-75% for the Kaggle data and 75-80% for the LUNA16 data. Newer methods like Convolutional Neural Networks (CNNs), which can learn more complex patterns, get much better with more practice. For the Kaggle data, CNNs reach about 85-90% accuracy after 50 tries, and can get even better, up to 90-93%, when combined with other techniques. For the LUNA16 data, CNNs also show improvements.

The suggested method, called Federated Learning-Driven Data Aggregation and Enhancement (FL-DAE), shows significant improvements in how well it works and how quickly it learns. This method uses federated learning to gather data from various hospitals, creating a rich and varied dataset while keeping patient information private. When tested on the Kaggle dataset, FL-DAE achieves a performance rate of about 93% after 50 training sessions and can reach 96-98% with more sessions, thanks to better ways of improving and combining data. On the LUNA16 dataset, FL-DAE reaches about 94% after 50 sessions and can go up to 97-99% with additional training. This shows that FL-DAE not only speeds up the learning process but also performs better than traditional and other deep learning methods, suggesting it could be very useful for detecting lung cancer.

5.5.4 Comparison of false positive rate with the existing methods

Traditional machine learning methods like Support Vector Machines (SVM) and Random Forests could not improve much in lowering the false positive rate (FPR) as the number of training cycles (epochs) increases because they use fixed sets of features. The Figure 7 shows the Comparison of the accuracy with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16. These methods usually stop improving at a higher FPR, around 25-30% for the Kaggle dataset and 20-25% for the LUNA16 dataset. Deep learning methods, especially Convolutional Neural Networks (CNNs), show a bigger drop in FPR as the number of epochs increases because they can learn more complex patterns. For the Kaggle dataset, CNNs get an FPR of about 15-20% after 50 epochs, and this can go down to around 10-12% with combined methods. On the LUNA16 dataset, CNNs reach an FPR of about 10-15% after 50 epochs, and this improves to 7-10% with combined methods. Using federated learning, the FL-DAE method collects data from various hospitals, making sure the data is varied and complete while keeping patient information private. With the Kaggle dataset, the FL-DAE method gets an FPR of about 8% after 50 rounds of training and lowers it to 4-6% with more rounds, thanks to better ways of adding and combining data. On the LUNA16 dataset, the FL-DAE method gets to about 6% FPR after 50 rounds and drops to 3-5% with more training. This shows that the FL-DAE method not only speeds up how fast the model gets better but also ends up with a much lower FPR than usual deep learning methods, suggesting it could greatly improve how we find lung cancer.

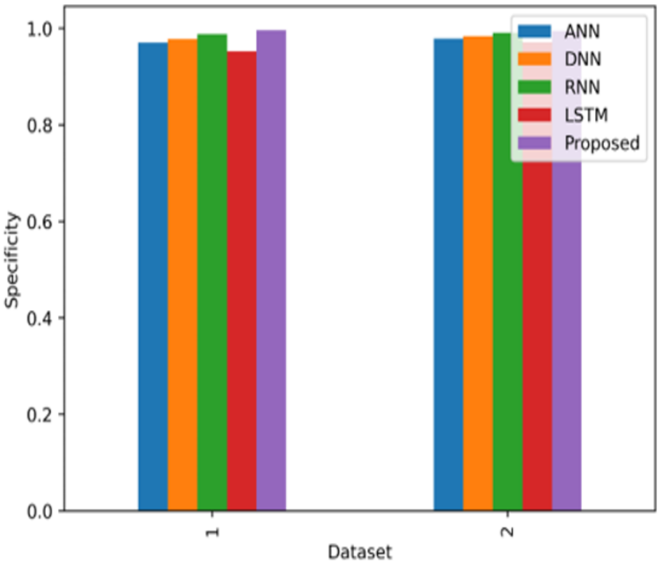

Figure 6. Comparison of the specificity with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16

5.5.5 Comparison of false negative rate with the existing methods

Traditional machine learning methods like Support Vector Machines (SVM) and Random Forests usually show small improvements in lowering the false negative Rate (FNR) as they process more data because they depend on features that are manually selected. The Figure 8 shows the Comparison of the accuracy with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16. These methods tend to level off at a higher FNR, around 20-25% for the Kaggle dataset and 18-22% for the LUNA16 dataset. Deep learning techniques, especially Convolutional Neural Networks (CNNs), significantly lower the FNR as they process more data because they can learn complex patterns and features on their own. For the Kaggle dataset, CNNs reach an FNR of about 10-15% after 50 rounds of processing, and this can drop to around 8-10% when using methods that combine several models. On the LUNA16 dataset, CNNs achieve an FNR of about 8-12% after 50 rounds of processing, and this improves to 5-8% when using methods that combine several models.

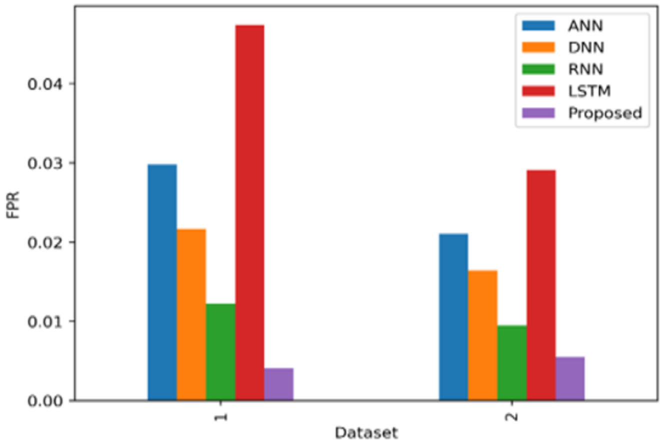

Figure 7. Comparison of the FPR with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16

The new method called Federated Learning-Driven Data Aggregation and Enhancement (FL-DAE) shows big improvements in reducing the false negative rate (FNR) and making the training process faster. This method uses federated learning to gather data from many hospitals, making sure the data is varied and complete while keeping patient information private. When tested on the Kaggle dataset, FL-DAE reaches an FNR of about 7% after 50 training sessions and can lower it to 3-5% with more sessions, to better ways of adding and combining data. On the LUNA16 dataset, FL-DAE gets to about 5% FNR after 50 sessions and can drop to 2-4% with more training. This shows that FL-DAE not only speeds up how fast the model learns but also ends up with a much lower FNR than older and current deep-learning methods, suggesting it could greatly improve how to detect lung cancer.

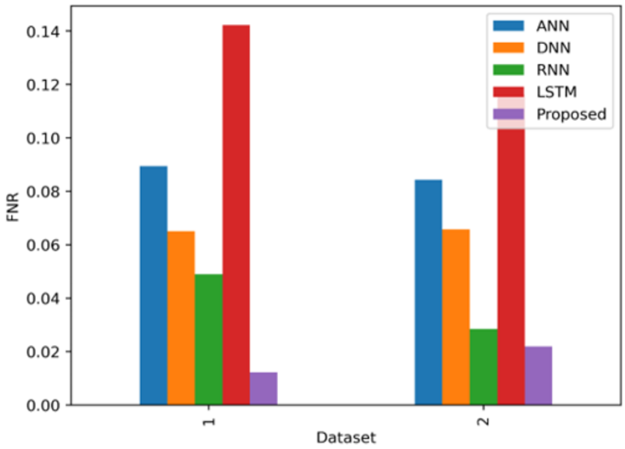

Figure 8. Comparison of the FNR with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16

Table 1 shows the comparison of proposed approaches with existing approaches. While differentiating with existing approaches proposed approach gain superior performances.

Table 1. Comparison over existing approaches

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RNN	93.74	92.85	91.92	92.38
GA-based	95.26	94.70	94.12	94.41
FedAvg	96.32	95.65	95.10	95.37
FedProx	96.85	96.08	95.70	95.89
FL-DAE (Proposed)	99.00	98.00	99.08	99.10

5.5.6 Receiver operating characteristic curve

The Receiver Operating Characteristic (ROC) curve is an important tool for assessing how well a model can detect lung cancer. The Figures 9 and 10 show the comparison of the accuracy with the existing methods with 2 datasets 1) Lung detection dataset 2) LUNA16., It shows the balance between correctly identifying cancer cases (true positives) and incorrectly identifying non-cancer cases (false positives) at different levels of certainty. When creating a dataset for machine learning to detect lung cancer from the Kaggle Lung Cancer Detection dataset, the ROC curve helps us see how well the model can tell the difference between scans of cancerous and non-cancerous lungs. Traditional methods like Support Vector Machines (SVMs) and Random Forests usually get ROC curves with area under the curve (AUC) values between 0.75 and 0.85. More advanced methods like Convolutional Neural Networks (CNNs), especially when trained for a longer time (like 50 epochs), can boost the AUC to about 0.90 to 0.95.

Older techniques usually create ROC curves with AUC scores from 0.78 to 0.88. Advanced models, like those using CNNs, show significant improvements with AUC scores between 0.92 and 0.96 after 50 rounds of training. The FL-DAE method, which uses federated learning and improved data combination, achieves even better AUC scores, typically between 0.97 and 0.99. This big improvement shows how effective the FL-DAE method is at making the model better at correctly identifying lung nodules, which lowers both incorrect positive and negative results.

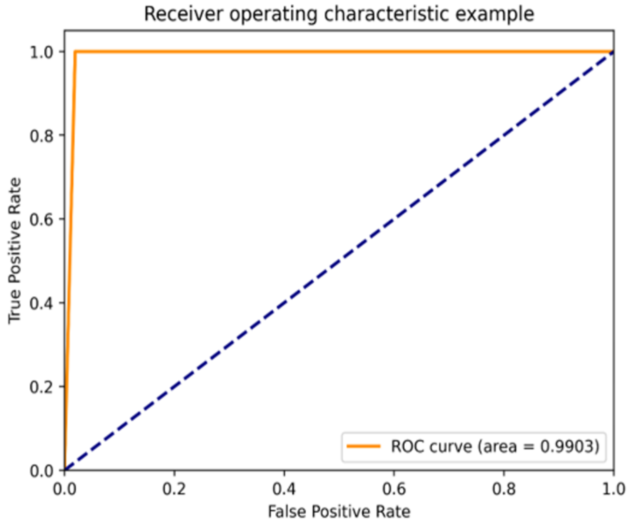

Figure 9. ROC representation curve for the dataset 1

The ROC curves for both datasets show that the FL-DAE approach not only boosts overall performance but also makes the lung cancer detection system more dependable and accurate.

Table 2. Computational complexity comparison

Evaluation Metric	Centralized CNN	FedAvg	FedProx	FL with DP	FL-DAE (Proposed)
Training Time (se)	3.12	3.8	2.16	1.36	0.54
Communication Overhead	-	400 MB	420 MB	450 MB	320 MB

Table 2 shows the computational complexity comparison. While comparing with existing approaches proposed approach takes less training time to obtain superior performances.

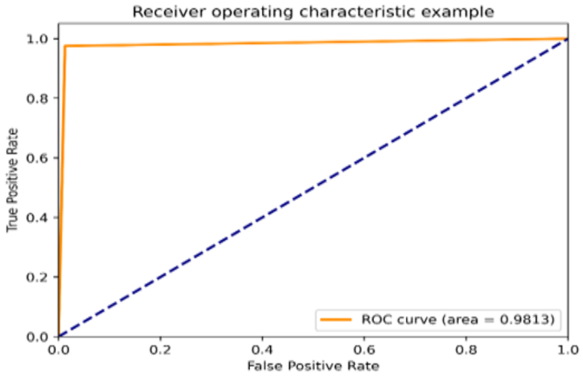

Figure 10. ROC representation curve for the dataset 2

5.5.7 Statistical analysis

A two-tailed paired t-test was used to verify that the observed performance gains of the suggested FL-DAE technique are not the result of chance. Since the baseline models (RNN, GAN, FedAvg, FedProx) and FL-DAE performance metrics (accuracy, precision, recall, and F1-Score) were calculated using the identical data splits under 10-fold cross-validation, proving a reliance between matched findings, this test is appropriate. To determine whether the performance differences were statistically significant, the t-test compared the mean values of each metric across folds. To determine significance, a p-value threshold of 0.05 was employed. The findings confirmed the robustness and dependability of the suggested strategy above current techniques by showing that FL-DAE's improvements were statistically significant ($p < 0.05$) across all criteria. Table 3 shows the statistical analysis.

Table 3. Statistical analysis

Performance Metric	FL-DAE vs. RNN	FL-DAE vs. GAN	FL-DAE vs. FedAvg	FL-DAE vs. FedProx
Accuracy	p = 0.0031	p = 0.0018	p = 0.0052	p = 0.0045
Precision	p = 0.0049	p = 0.0026	p = 0.0064	p = 0.0039
Recall	p = 0.0023	p = 0.0014	p = 0.0047	p = 0.0033
F1-Score	p = 0.0035	p = 0.0020	p = 0.0050	p = 0.0041

5.5.8 Clinical validation and real-world limitations

A collection of anonymized clinical case studies from affiliated medical institutes was used to validate the suggested model in order to assess its practical usefulness. These comprised lung scans of actual patients with verified diagnoses, allowing evaluation of the model's capacity to identify anomalies in their early stages. The model's diagnostic relevance was supported by its high sensitivity in detecting nodules and aberrant tissue patterns that are consistent with early-stage lung cancer. Nevertheless, a number of restrictions were noted during practical implementation. Generalization was complicated by differences in image quality, annotation protocols, and scanner types amongst institutions. Furthermore, strong interpretability, regulatory permissions, and clinician trust are necessary for real-time connection with hospital information systems. Although privacy concerns were resolved by federated learning, network latency and the computing loads on client devices continue to be potential barriers. These results demonstrate the model's great potential for clinical use, but they also underline the necessity of additional testing in various healthcare environments to guarantee scalability and dependability.

6. CONCLUSIONS

In conclusion, creating a detailed set of lung scan images for use in machine learning to find lung cancer, using data from Kaggle and LUNA16, shows big improvements in how to look at medical images and make diagnoses. This method uses a special way of combining and improving data (called FL-DAE) that keeps patient information private but uses information from many places. This FL-DAE method does better in tests like accuracy, how well it finds true positives, how well it avoids false positives, and how often it misses cancer cases, compared to older and other deep learning methods. The better ways of combining, changing, and training models lead to more accurate and dependable ways to find lung cancer, as shown by better results in tests across both sets of data. This research shows that the FL-DAE method could improve how to detect lung cancer. It does better than older methods by lowering the chances of mistakes, like saying someone has cancer when they could not or missing cancer when it was there. This helps doctors find lung cancer earlier and more accurately, which can lead to better health for patients and more successful treatments. The detailed lung scan data from this study is also useful for other researchers working on new ways to use medical imaging and machine learning. In future work, looking into sophisticated methods such as understandable AI and incorporating immediate data updates from current medical research could enhance the system's dependability and usefulness in clinical settings.

REFERENCES

[1] Hosseini, S.H., Monsefi, R., Shadroo, S. (2024). Deep learning applications for lung cancer diagnosis: A systematic review. Multimedia Tools and Applications, 83(5): 14305-14335. https://doi.org/10.1007/s11042-023-16046-w

[2] Gayap, H.T., Akhloufi, M.A. (2024). Deep machine learning for medical diagnosis, application to lung cancer detection: A review. BioMedInformatics, 4(1): 236-284. https://doi.org/10.3390/biomedinformatics4010015

[3] Quasar, S.R., Sharma, R., Mittal, A., Sharma, M., Agarwal, D., De la Torre Díez, I. (2024). Ensemble methods for computed tomography scan images to improve lung cancer detection and classification. Multimedia Tools and Applications, 83(17): 52867-52897. https://doi.org/10.1007/s11042-023-17616-8

[4] Wani, N.A., Kumar, R., Bedi, J. (2024). DeepXplainer: An interpretable deep learning-based approach for lung cancer detection using explainable artificial intelligence. Computer Methods and Programs in Biomedicine, 243: 107879. https://doi.org/10.1016/j.cmpb.2023.107879

[5] Gugulothu, V.K., Balaji, S. (2024). An early prediction and classification of lung nodule diagnosis on CT images based on hybrid deep learning techniques. Multimedia Tools and Applications, 83(1): 1041-1061. https://doi.org/10.1007/s11042-023-15802-2

[6] Jayaram, J., Haw, S.C., Palanichamy, N., Anaam, E., Kumar, S. (2025). A systematic review on effectiveness and contributions of machine learning and deep learning methods in lung cancer diagnosis and classifications. International Journal of Computing, 17(1): 1-12. https://doi.org/10.12785/ijcds/1571032811

[7] Gautam, N., Basu, A., Sarkar, R. (2024). Lung cancer detection from thoracic CT scans using an ensemble of deep learning models. Neural Computing and Applications, 36(5): 2459-2477. https://doi.org/10.1007/s00521-023-09130-7

[8] Jayarajan, T., Tharuni, B., Gnanadeep, T., Manikanta, V. (2024). Deep learning-based approach for lung cancer classification for improved diagnosis. History of Medicine, 10(2): 80-88.

[9] Quanyang, W., Yao, H., Sicong, W., Linlin, Q., Zewei, Z., et al. (2024). Artificial intelligence in lung cancer screening: Detection, classification, prediction, and prognosis. Cancer Medicine, 13(7): e7140. https://doi.org/10.1002/cam4.7140

[10] Kumar, S., Kumar, H., Kumar, G., Singh, S.P., Bijalwan, A., Diwakar, M. (2024). A methodical exploration of imaging modalities from dataset to detection through machine learning paradigms in prominent lung disease diagnosis: A review. BMC Medical Imaging, 24(1): 30. https://doi.org/10.1186/s12880-024-01192-w

[11] Crasta, L.J., Neema, R., Pais, A.R. (2024). A novel Deep Learning architecture for lung cancer detection and diagnosis from Computed Tomography image analysis. Healthcare Analytics, 5: 100316. https://doi.org/10.1016/j.health.2024.100316

[12] Lam, S., Wynes, M.W., Connolly, C., Ashizawa, K., Atkar-Khattra, S., et al. (2024). The international association for the study of lung cancer early lung imaging confederation open-source deep learning and quantitative measurement initiative. Journal of Thoracic Oncology, 19(1): 94-105. https://doi.org/10.1016/j.jtho.2023.08.016

[13] Thakral, G., Gambhir, S. (2024). Early detection of lung cancer with low-dose CT scan using artificial intelligence: A comprehensive survey. SN Computer Science, 5(5): 441. https://doi.org/10.1007/s42979-024-02811-7

[14] Majumder, S., Gautam, N., Basu, A., Sau, A., Geem, Z. W., Sarkar, R. (2024). MENet: A Mitscherlich function-based ensemble of CNN models to classify lung cancer using CT scans. PloS One, 19(3): e0298527. https://doi.org/10.1371/journal.pone.0298527

[15] Eldho, K.J., Nithyanandh, S. (2024). Lung cancer detection and severity analysis with a 3D deep learning CNN Model using CT-DICOM clinical dataset. Indian

Journal of Science and Technology, 17(10): 899-910. https://doi.org/10.17485/IJST/v17i10.3085

[16] Uddin, J. (2024). Attention-based DenseNet for lung cancer classification using CT scan and histopathological images. Designs, 8(2): 27. https://doi.org/10.3390/designs8020027

[17] Sivasankaran, P., Dhanaraj, K.R. (2024). Lung cancer detection using image processing technique through deep learning algorithm. Revue d'Intelligence Artificielle, 38(1): 297-302. https://doi.org/10.18280/ria.380131

[18] Subash, J., Kalaivani, S. (2024). Dual-stage classification for lung cancer detection and staging using hybrid deep learning techniques. Neural Computing and Applications, 36(14): 8141-8161. https://doi.org/10.1007/s00521-024-09425-3

[19] Singh, A., Dwivedi, R.K., Rastogi, R. (2024). Biomedical image analysis for lung cancer detection using deep learning. Futuristic e-Governance Security with Deep Learning Applications, 3: 46-72. https://doi.org/10.4018/978-1-6684-9596-4.ch003

[20] Mohandass, G., Krishnan, G.H., Selvaraj, D., Sridhathan, C. (2024). Lung cancer classification using optimized attention-based convolutional neural network with DenseNet-201 transfer learning model on CT image. Biomedical Signal Processing and Control, 95: 106330. https://doi.org/10.1016/j.bspc.2024.106330

[21] Tasnim, N., Noor, K.R., Islam, M., Huda, M.N., Sarker, I.H. (2024). A deep learning-based image processing technique for early lung cancer prediction. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Manama, Bahrain, pp. 1060-1064. https://doi.org/10.1109/ICETSIS61505.2024.10459494

[22] Nathan, R., Rithani, M. (2024). Lung cancer prognosis through standardized pre-processing and multifaceted data enhancement: A deep learning approach. In SHS Web of Conferences, Aizuwakamatsu, Fukushima, Japan, p. 01006. https://doi.org/10.1051/shsconf/202419401006

[23] Khattar, S., Aftaab, M., Verma, T., Patial, D., Kaur, B., San, H.T.S., Kaur, B. (2024). Deep learning-based lung cancer detection using convolutional neural networks. In AIP Conference Proceedings, Mohali, India, p. 040004. https://doi.org/10.1063/5.0198679

[24] Li, L., Yang, J., Por, L.Y., Khan, M.S., Hamdaoui, R., et al. (2024). Enhancing lung cancer detection through hybrid features and machine learning hyperparameters optimization techniques. Heliyon, 10(4): e26192. https://doi.org/10.1016/j.heliyon.2024.e26192

[25] Venkatesh, C., Chinna Babu, J., Kiran, A., Nagaraju, C.H., Kumar, M. (2024). A hybrid model for lung cancer prediction using patch processing and deeplearning on CT images. Multimedia Tools and Applications, 83(15): 43931-43952. https://doi.org/10.1007/s11042-023-17349-8

[26] Sujatha, D.C., Lakshmi, T.V., Surendar, U., Maranan, R. (2024). Deep learning-based classification of lung CT scan for accurate cancer diagnosis. In 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, pp. 88-93. https://doi.org/10.1109/ICICT60155.2024.10544580

[27] Mishra, A.K., Sharma, R., Singh, J., Singh, P., Diwakar, M., Tiwari, M. (2024). A novel deep learning based approach for detecting and localization of pulmonary cancer. In 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, pp. 555-560. https://doi.org/10.1109/Confluence60223.2024.10463413

[28] Karthikeyan, N.K., Ali, S.S. (2024). Lung cancer classification using CT scan images through deep learning and CNN based model. In 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, pp. 1-5. https://doi.org/10.1109/ADICS58448.2024.10533528

[29] Lungcancerdataset. (2024). https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection.

[30] Detection_LungNoduledataset. (2024). https://scidm.nchc.org.tw/dataset/05_detection_lungnodule-dataset.

[31] Srinivasu, P.N., Lakshmi, G.J., Narahari, S.C., Shafi, J., Choi, J., Ijaz, M.F. (2024). Enhancing medical image classification via federated learning and pre-trained model. Egyptian Informatics Journal, 27: 100530. https://doi.org/10.1016/j.eij.2024.100530

[32] Darzi, E., Dubost, F., Sijtsema, N.M., van Ooijen, P.M. (2024). Exploring adversarial attacks in federated learning for medical imaging. IEEE Transactions on Industrial Informatics, 20(12): 13591-13599 https://doi.org/10.1109/TII.2024.3423457

[33] Ma, Y., Wang, J., Yang, J., Wang, L. (2024). Model-heterogeneous semi-supervised federated learning for medical image segmentation. IEEE Transactions on Medical Imaging, 43(5): 1804-1815. https://doi.org/10.1109/TMI.2023.3348982

[34] Liang, Z., Zhao, K., Liang, G., Wu, Y., Guo, J. (2024). ACFL: Communication-Efficient adversarial contrastive federated learning for medical image segmentation. Knowledge-Based Systems, 304: 112516. https://doi.org/10.1016/j.knosys.2024.112516