CT Imaging-Based Deep Learning System for Non-Small Cell Lung Cancer Detection and Classification

Devyani Rawat[1], Sachin Sharma[2], Shuchi Bhadula[3]
Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India[1, 3]
Amity School of Engineering and Technology, Amity University Punjab, Mohali, India[2]

***Abstract*—About 85% of all occurrences of lung cancer are classified as Non-Small Cell Lung Cancer (NSCLC), making it a serious worldwide health concern. For better treatment results and patient survival, NSCLC must be detected early and accurately. This research presents an advanced Deep Learning-enabled Lung Cancer Detection and Classification System (LCDCS) aimed at significantly improving diagnostic precision and operational efficiency. Emerging technologies such as artificial intelligence and multi-level convolutional neural networks (ML-CNN) are increasingly being leveraged in CT imaging-based deep learning systems for accurate detection. The outlined framework leverages a multi-layer convolutional neural network to effectively analyse CT scan images and accurately classify lung nodules. Tomek link and Adaptive Synthetic Sampling (ADASYN) are used in a novel way to balance data, address class imbalance, and guarantee strong model performance. Deep learning with a CNN model is utilized to derive features, and the SoftMax function is applied for multi-class classification. Thorough evaluation on datasets like the LUNA16 dataset demonstrates that the system surpasses earlier models and data balancing techniques in accuracy, yielding a training accuracy of 95.8% and a validation accuracy of 96.9%. The findings demonstrate the potential of the suggested method as a trustworthy diagnostic instrument for the prompt identification of lung cancer. The study emphasizes on how crucial it is to combine deep learning architectures with sophisticated data balancing techniques to overcome medical imaging difficulties and raise diagnostic accuracy. Future research attempts to investigate real-time deployment in clinical settings and expand the system's capability to encompass more cancer types.**

Keywords*—*Artificial intelligence; NSCLC; ML-CNN; ADASYN; tomek link

I. Introduction

Lung cancer is the leading type of cancer diagnosed worldwide. As of the latest data, there are approximately 2.2 million new cases each year. It continues to be a leading cause of cancer-related deaths globally, with Non-Small Cell Lung Cancer (NSCLC) comprising roughly 85% of all diagnosed cases. Traditionally, the detection and classification of lung cancer rely on the expertise of radiologists and pathologists who analyze Computed Tomography (CT) images and histopathological samples. However, this process is often time-consuming, subjective, and prone to variability, leading to a demand for more reliable and efficient diagnostic tools. In the realm of therapeutic diagnostics, the integration of progressed innovations like deep learning has assisted a new way of accuracy and proficiency. NSCLC, account for a critical portion of lung cancer cases, urges precise and timely detection for effectual treatment and patient care. Deep learning methods has proven its ability in enhancing the detection and classification processes, contributing to giving vital insights and improving patient outcomes [9].

This study proposes a novel Deep Learning-enabled Lung Cancer Detection and Classification System (LCDCS) particularly for non-small cell. The model leverages multiple scales of CT images to capture the diverse features of lung nodules, enabling a more comprehensive analysis [17]. By integrating the outputs of four CNNs, the suggested framework aims to deliver an efficient and accurate classification of lung nodules into categories such as benign tissue, large cell carcinoma, and squamous cell carcinoma [16]. The effectiveness of the proposed model is demonstrated through rigorous training and validation on a substantial dataset of histopathological images, highlighting its potential to be a valuable tool in the prompt diagnosis and care planning for lung cancer patients.

The study introduces a Deep Learning-enabled Lung Cancer Detection and Classification System (LCDCS), focusing specifically on NSCLC.

- The system employs a multi-level convolutional neural network (ML-CNN) for analyzing CT scan images.
- It highlights the use of ADASYN (Adaptive Synthetic Sampling) combined with Tomek Links for efficient data balancing and enhanced classification performance.
- Multi-scale Image Analysis: Utilizes multiple scales of CT images to capture diverse features of lung nodules for enhanced detection and classification.
- A comprehensive comparative analysis aimed at evaluating the outcome of various class balancing strategies for lung cancer detection.
- The study suggests improving the system so it can be used for more types of cancer or even help predict related health problems like heart disease.

II. Literature Review

The study presents a deep convolutional neural network with multiple levels designed to detect and classify lung cancer by analyzing CT scan images of lung nodules. By leveraging a four-level CNN architecture that processes multiple scales of nodule images, the model effectively distinguishes between benign tissue, large cell carcinoma, and squamous cell

carcinoma. Modeled on a dataset of 25,000 histopathological images, the model achieved a notable accuracy of 78% on the training set and 89.6% on the validation set, highlighting its potential as an efficient tool to aid radiologists in early lung cancer diagnosis [1]. Jenita Subash et al. presents a study which aims to develop a dual-stage classification system for lung cancer detection and staging by integrating hybrid deep learning techniques. The study likely involves preprocessing lung imaging data, such as CT scans, to improve the quality and relevance of the input features. The first stage from the classification involves convolutional neural network (CNN) or a similar deep learning architecture which is used to identify cancerous lesions from the imaging data. Once cancer is detected, the second stage involves determining the stage of lung cancer (e.g., Stage I, II, III, or IV). This stage might use a more complex network or a combination of models to classify the cancer stage based on tumor size, spread, and other relevant clinical features [2].

The study addresses the challenge of diagnosing NSCLC, which accounts for approximately 85 % of lung cancer cases.

The authors employ CNN architectures to analyze the images and identify patterns indicative of NSCLC [3] Approaches such as Gradient-weighted Class Activation Mapping or Local Interpretable Model-agnostic explanations are likely used to highlight regions of the images with the highest influence on the model's outcomes; by making the model's predictions interpretable. The study aims to increase trust in AI systems among medical professionals [4].

1) Lung cancer types: Lung cancer is primarily categorized into two major types, distinguished by the microscopic characteristics of the cancerous cells and their growth patterns. Fig. 1 shows lung cancer under the microscope.

a) Non-Small Cell Lung Cancer (NSCLC): NSCLC is the predominant form of lung cancer, representing approximately 85% of all cases. It encompasses a variety of subtypes, each with distinct characteristics.

- Adenocarcinoma: This is the most prevalent subtype of NSCLC, typically originating in the outer regions of the lungs. It tends to grow more slowly and is more common in non-smokers compared to other types.
- Squamous Cell Carcinoma: This type usually starts in the airways (bronchi) and is more commonly associated with smoking. It tends to grow in the central parts of the lungs.
- Large Cell Carcinoma – A relatively uncommon form of lung cancer that can develop in any region of the lung, characterized by its rapid growth and aggressive spread.

b) Small Cell Lung Cancer (SCLC): SCLC, also known as small cell carcinoma, makes up about 15% of lung cancer cases. It's characterized by small, round cells and is often linked to smoking. SCLC tends to grow rapidly and is often diagnosed at an advanced stage. Each type of lung cancer can vary in its treatment and prognosis, so accurate diagnosis and staging are crucial for determining the best course of action [5].

2) Lung cancer detection techniques: Detecting lung cancer early is crucial for effective treatment. Here are some common techniques used for detection:

a) Imaging tests:

- Chest X-ray: Often the first test used to look for abnormalities in the lungs.
- Computed Tomography (CT) Scan: Generates high-resolution cross-sectional images of the lungs, enabling the detection of smaller tumors and providing precise evaluation of their size, location, and potential spread.
- Positron Emission Tomography (PET) Scan: Utilized to assess the spread of cancer to other areas of the body by detecting regions with elevated metabolic activity.
- Magnetic Resonance Imaging (MRI): Less commonly used for lung cancer but helpful in assessing spread to the brain or spinal cord.

b) Screening tests:

- Low-Dose Computed Tomography (LDCT): Suggested for individuals at high risk (e.g., heavy smokers or those with a smoking history). It can identify lung cancer at an earlier stage compared to a chest X-ray.

c) Biopsy:

- Needle Biopsy: A small needle is inserted into the chest to extract tissue from the lung.
- Bronchoscopy: A flexible instrument is inserted through the nose or mouth into the lungs to obtain tissue samples.
- Endobronchial Ultrasound (EBUS): A specialized form of bronchoscopy that utilizes ultrasound imaging to precisely guide the biopsy needle for tissue sampling.

d) Sputum cytology: Analysis of mucus (sputum) from the lungs to look for cancer cells, particularly useful in some cases of squamous cell carcinoma.

e) Molecular testing:

- Genetic Testing: Examines cancer cells for specific genetic mutations that can guide targeted therapy options [6].

Table I depicts the physical examination required for lung cancer. Table II depicts the different parameters of blood test with their normal ranges. Table III depicts electrocardiogram (for heart conditions) with its normal ranges. Table IV shows different imaging test required to detect lung cancer. Table V shows the analysis of urine with its normal ranges. Table VI shows the first assessment of NSCLC diagnosis.

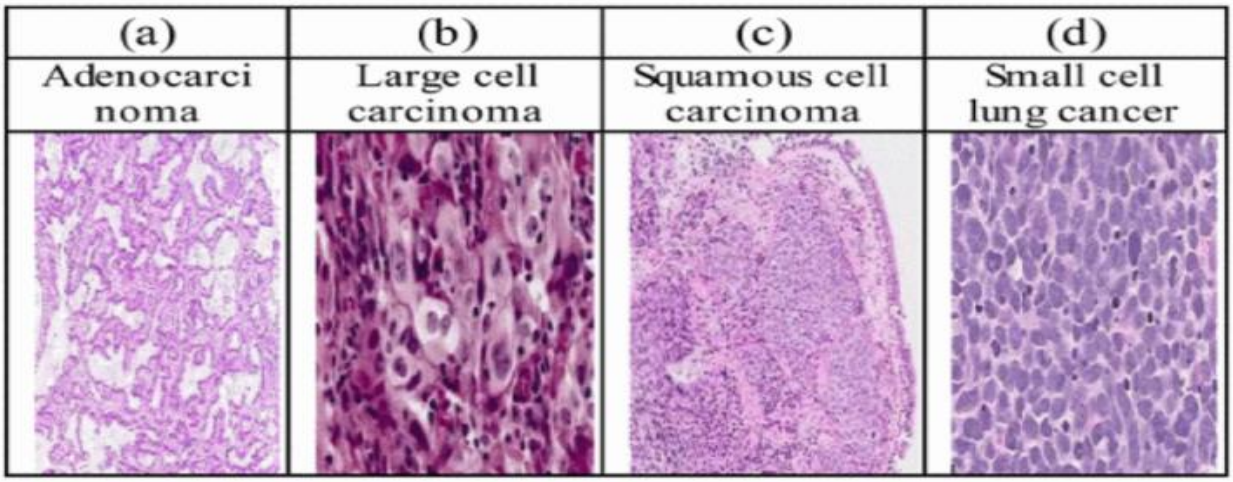


Fig. 1. Categories of lung cancer.

TABLE I PHYSICAL EXAMINATION FOR LUNG CANCER

Test	Parameter
Physical Examination	Palpation: The doctor examines the affected area by touch to identify any tenderness, swelling, or abnormalities.
	Range of Motion Tests: Assesses the movement in joints or muscles.
	Neurological Exam: Checks reflexes, muscle strength, and sensory function.

TABLE II DIFFERENT PARAMETERS OF BLOOD TEST WITH THEIR NORMAL RANGES

Blood Test	Parameter	Range
Complete Blood Count (CBC)	White Blood Cell (WBC)	4,000 to 11,000 cells/µL
	Red Blood Cell (RBC)	M: 4.7 to 6.1 million cells/µL, W: 4.2 to 5.4 million cells/µL
	Hemoglobin (Hb)	M: 13.8 to 17.2 g/dL, W: 12.1 to 15.1 g/dL
	Hematocrit (Hct)	M: 40.7% to 50.3%, W: 36.1% to 44.3%
	Platelet Count	150,000 to 450,000 platelets/µL
	Red Cell Distribution Width (RDW)	11.5% to 14.5%
Differential Leucocyte count	Segmented Neutrophils	40% to 70% of total WBCs
	Lymphocytes	20% to 40% of total WBCs
	Monocytes	2% to 8% of total WBCs
	Mean Platelet Volume (MPV)	7.5 - 12.0 fL

TABLE III ELECTROCARDIOGRAM WITH ITS NORMAL RANGE

Test	Parameter	Normal Range (Men)	Normal Range (Women)	Description
Electrocardiogram	Heart Rate (Resting)	60-100 bpm	60-100 bpm	Number of heart beats per minute
	P Wave Duration	0.08 to 0.11 seconds	0.08 to 0.11 seconds	Time taken for atrial depolarization
	PR Interval	0.12 to 0.20 seconds	0.12 to 0.20 seconds	Time between P wave start and QRS complex start
	QRS Duration	0.08 to 0.10 seconds	0.08 to 0.10 seconds	Time for ventricular depolarization
	QT Interval	0.35 to 0.45 seconds	0.36 to 0.46 seconds	Time from start of Q wave to end of T wave
	ST Segment	Isoelectric (flat)	Isoelectric (flat)	Represents the period between ventricular depolarization and repolarization
	T Wave	Positive in most leads	Positive in most leads	Represents ventricular repolarization
	RR Interval	0.6 to 1.2 seconds	0.6 to 1.2 seconds	Time between two consecutive R wave peaks
	Axis	−30° to +90°	−30° to +90°	Heart's electrical axis direction in degrees

TABLE IV DIFFERENT IMAGING TEST

Imaging Test	Normal Findings	Purpose
Chest X-ray (CXR)	Clear lungs, normal heart size, no fluid or masses	Assess lung health, detect infections, evaluate heart size
Echocardiogram	Normal heart size and function, no valve abnormalities	Assess heart function, valve abnormalities, heart disease
Electrocardiogram (ECG)	Normal heart rhythm, no signs of arrhythmia or heart damage	Monitor heart rhythm, detect arrhythmias, heart damage

TABLE V URINE ANALYSIS WITH ITS NORMAL RANGE

Test	Parameter	Normal Range	Description
Urine Analysis	Color	Pale yellow to deep amber	Indicates hydration levels and possible health issues
	Clarity	Clear	Cloudy urine may suggest infection or presence of crystals
	Odor	Mild, not strong	Strong odor can suggest infection or diabetes
	Specific Gravity	1.005 to 1.030	Measures concentration of urine; high values may indicate dehydration
	pH	4.5 to 8.0	Reflects acidity or alkalinity of urine
	Protein	Negative to trace (up to 150 mg/day)	Higher levels can indicate kidney issues
	Glucose	Negative	Presence of glucose suggests diabetes
	Ketones	Negative	Presence indicates uncontrolled diabetes or starvation
	Bilirubin	Negative	Indicates liver function; presence can indicate liver disease
	Urobilinogen	0.1 to 1.0 mg/dL	Low or high levels may suggest liver or bile duct issues
	Red Blood Cells (RBCs)	0 to 3 RBCs/HPF	Higher levels can indicate infection, trauma, or stones
	White Blood Cells (WBCs)	0 to 5 WBCs/HPF	Increased levels suggest infection or inflammation
	Nitrites	Negative	Presence suggests bacterial infection
	Leukocyte Esterase	Negative	Indicates white blood cells, which may indicate infection
	Casts	None to rare hyaline casts	Presence of certain casts suggests kidney disease
	Crystals	None to few	High levels may indicate kidney stones or metabolic issues
	Bacteria	None	Presence suggests infection
	Yeast	None	Presence may indicate infection
	Epithelial Cells	Few (0 to 5 cells/HPF)	High numbers may indicate contamination or infection

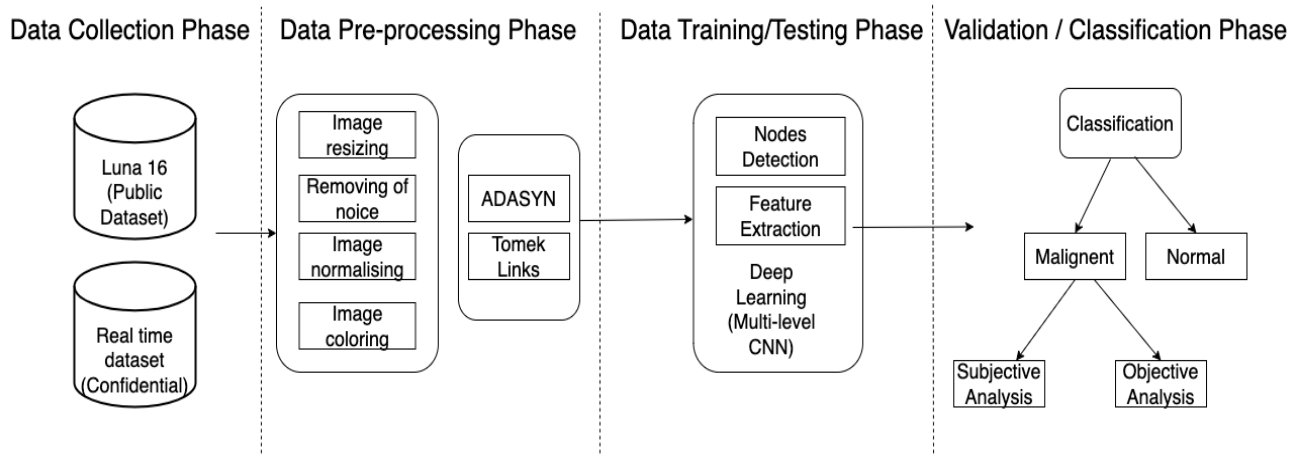


Fig. 2. Deep learning-enabled lung cancer detection and classification system.

3) Deep learning: Deep learning has played a transformative role in lung cancer detection by increasing the precision, responsiveness, and consistency of diagnostic processes [8]. The emergence of AI driven learning models with a focus on CNNs, offers a powerful tool to automate and potentially improve the accuracy of lung cancer classification [2]. The study proposes an automated system that combines deep learning with multiple strategies like (data augmentation, multi-scale analysis, ensemble learning and post processing), to improve the identification and categorization of lung nodules in CT (Computed Tomography) scans [7].

TABLE VI FIRST ASSESSMENT OF DIAGNOSIS OF NSCLC

Assessment Step	Details
Patient History	- Tobacco use history (duration, pack-years)
	- Exposure to environmental or workplace carcinogens.
	- Genetic predisposition to lung cancer or other neoplasms.
	- Symptoms and warning signs (chronic cough, blood-tinged sputum, unexplained weight loss, chest pain, shortness of breath).
Physical Examination	- Inspection (e.g., clubbing, cachexia)
	- Palpation (e.g., lymphadenopathy)
	- Percussion (e.g., dullness over lungs)
	- Auscultation (e.g., wheezes, crackles, absent breath sounds)
Imaging Studies	- Chest X-ray (initial screening, may show masses, consolidation)
	- Chest CT scan (detailed imaging, tumor size, lymph node involvement)
Biopsy	- Needle biopsy (CT-guided or bronchoscopic biopsy) for histopathological diagnosis
	- Sputum cytology (to detect cancer cells in sputum)
Laboratory Assessments.	- Comprehensive blood analysis (CBC) to evaluate for anemia, infection, or other hematological abnormalities.
Molecular Testing	- EGFR, ALK, ROS1 mutations (for targeted therapy)
Staging Tests	- PET scan (to assess metastasis)
	- Mediastinoscopy (to evaluate mediastinal lymph nodes)
	- Brain MRI (if neurological symptoms are present)
Risk Assessment	- Eastern Cooperative Oncology Group (ECOG) performance status
	- Assessment of comorbidities and general health condition

III. PROPOSED METHODOLOGY

The proposed Deep Learning-enabled LCDCS (see Fig. 2) aims to detect Non-Small Cell Lung Cancer by studying the images obtained by CT-scan, MRI using a multi-level convolutional neural network for feature extraction and SoftMax function for classification purpose. To cater unresolved class imbalance there are different sampling methods are present, like:

Over Sampling Methods:

- SMOTE (Synthetic Minority Over-sampling Technique) - produces artificial instances for the underrepresented class by interpolating between existing minority class examples.
- Random Over sampling - simply duplicates random samples from the minority class to achieve balance.
- ADASYN (Adaptive Synthetic Sampling) - an advanced variant of SMOTE that prioritizes generating synthetic instances for minority class samples that are more challenging to classify [10].

Under-Sampling Methods:

- Random Under-Sampling - randomly eliminates instances from the dominant class to achieve a balanced class distribution.
- NearMiss- chooses instances from the dominant class that are nearest to the minority class examples, thereby decreasing the size of the dominant class.
- Tomek Links -identifies and removes overlapping samples between classes to create a clearer boundary between the majority and minority classes.

In this research work we have used the combination methods i.e. ADASYN (Adaptive Synthetic Sampling) and Tomek Links, to achieve data balancing because the hybrid combination gives better results as compared to the other techniques. The detailed comparison of the proposed Deep learning-enabled LCDCS system with alternative dataset balancing techniques for recognizing the symptoms of lung cancer, accounting for both cases with and without the use of DL, is summed up in Table VII. Algorithm 1 and Algorithm 2 [15] represents the ADASYN algorithm and Tomek link algorithm for achieving data balancing. The suggested system's first module, which includes steps like pre-processing, data balance, and classification, aims to detect lung cancer. Pre-processing stage which involves:

1) Data cleaning: It involves handling missing values and replace missing values with mean, median, mode.

2) Data transformation: It involves normalization, standardization, log transformation and box-cox transformation.

$$Normalization\ x' = \frac{x - \min(x)}{x - \min(x)} \quad (1)$$

3) Data reduction: It involves principal component analysis which reduces the number of features while retaining most of the variance.

If we have a data matrix $\boldsymbol{X}$ with n samples and p features (e.g., pixel intensities from CT scans or biomarkers), and we aim to reduce this to k principal components ($k<p$):

$$Zn \times K = Xn \times p\,.Wp \times k \quad (2)$$

where, Z is the reduced data matrix, X is the original data matrix and W is the matrix of selected eigenvectors (principal components).

4) Data encoding: It has categorical encoding which is further divided into three categories*:*

Label Encoding: Label encoding can be used for ordinal variables such as "Stage of Cancer" (e.g., Stage I, Stage II, Stage III, Stage IV).

Let C= {c1, c2,...,ck} be the set of unique categories, and let x∈ C, be a category for given observation.

One-hot Encoding: For categorical variables without a natural order (e.g., "Type of Symptom", "Smoking History"), one-hot encoding is more appropriate.

Let C={c1,c2,…,ck} be the set of categories for a nominal variable. One-hot encoding generates a binary **v** (x) for each category x∈ C:

$$\boldsymbol{v}(x) = [\, v_1, v_2, v_3 \ldots . v_k \,] \quad (3)$$

where,

$$vi = \begin{cases} 1 \; if \; x = ci \\ 0 \; otherwise \end{cases}$$

5) Data Sampling:

a) Random sampling: Select a random subset of data to reduce computational cost.

b) Stratified sampling: Ensure that the sample represents different strata or groups within the data.

c) Oversampling and under sampling: Adjust the dataset to balance class distribution in imbalanced datasets (e.g., SMOTE).

d) Handling time-series data resampling: Change the frequency of time-series data (e.g., daily to monthly).

Given a time-series $x(t)$ where t represents the time index (e.g., days), resampling to a coarser time frequency (e.g., monthly) can be done by aggregating values over the new time intervals. If you aggregate using a sum, the equation is:

$$xm\,(T) = \sum t \in T \; x(t) \quad (4)$$

where,

$x(t)$ is the original time-series data.

T is the new time interval (e.g., a month).

$xm(T)$ is the resampled data at the new frequency.

- Smoothing- utilize moving means or exponential smoothing to minimize fluctuations and enhance signal clarity. A simple moving average (SMA) over a window of size *N* is calculated as:

$$SMAn(t) = 1/n \sum_{i=0}^{n-1} x(t-i) \quad (5)$$

where,

$x(t)$ is the original time series data

$SMAn(t)$)is the smoothed value at time *t*.

N is the window size.

- Detrending- remove trends to focus on the seasonality and residual components.

6) Data splitting: Efficient preprocessing can substantially boost the efficacy and dependability of machine learning models. The selection of methodologies relies on the characteristics of the data and the particular demands of the analysis or model.

If the trend is linear, you can model it as:

Trend (*t*) = $a.t + b$

where, a is the slope and b is the intercept.

To detrend the time series:

$$xdetrended(t) = x(t) - Trend(t) \quad (6)$$

where, $xdetrended(t)$ is the time-series data after removing the trend.

Algorithm 1 describes the ADASYN sampling technique for generating synthetic data points in order to achieve data balancing. First of all it checks class imbalance**.**

1) If imbalance d <d_{th}, generate synthetic data

2) In the next step compute required synthetic data count G, this controls how much data is needed.

3) Place more synthetic samples near decision boundaries.

4) Pick a neighbour, create new samples in between $s_i = x_i + (x_{zi} - x_i) \times \lambda$.

Algorithm 1

Input is done when Data set for training is identified. The class identity label associated with x_i is denoted by yi ∈ Y = {1, −1}, where x_{is} is the entity in the n-dimensional feature space X. The number of minority class examples is denoted by m_s, and the quantity of dominant class examples by ml. Therefore, $m_s + m_l$ = m and $m_s \leq m_l$. Degree of class imbalance is calculated by:

$$d = m_s / m_l \quad (7)$$

where, $d \in (0, 1)$.

If d is less than d_{th}, then (d_{th} is a predetermined threshold):

(a) We determine how many examples of synthetic data must be created for the imbalanced class by

$$G = (m_l - m_s) \times \beta \quad (8)$$

Once the synthetic data is generated, the parameter β∈[0,1] is used to control the desired balance level. When β=1, the generalization process yields a fully balanced dataset.

(b) Identify the K nearest neighbours for each instance x_i belonging to the minority class using the Euclidean distance in an n-dimensional space. Then, compute the ratio r_i, defined as follows:

$$r_i = \Delta_i / K, \; i = 1,\ldots,m_s \quad (9)$$

where ri ∈ [0, 1] since Δi is the count of samples in xi's K nearest neighbours that are members of the dominant class; Normalizer *ri* according to r^i = ri/ $\sum_{i=1}^{ms} ri$ so that r^i is a density distribution ($\sum_i r\hat{}i = 1$).

Next, we determine how many samples of synthetic data must be created for every minority example xi by:

$$gi = r\hat{}i \times G \quad (10)$$

Hence, according to Eq. (2), G denotes the overall number of generated data points that must be produced for the outlier class. We create gi synthetic data examples for every outlier class data example xi using the procedures listed below when the loop is done from 1 to gi.

(i) For data x_i, select one sparse data example (x_{zi}) at random from the K nearest neighbours. (ii) Simulated data example is generated by:

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \quad (11)$$

where, λ is a random number: λ ∈ [0, 1], and (xzi − xi) is the displacement vector.

Algorithm 2 presents a data cleaning approach based on Tomek Links, which serves as both a data balancing technique and a method for addressing two key issues: reducing noise in datasets and mitigating class imbalance.

1) For each majority class sample x: it is finding its nearest neighbours y (the closest data point).

2) If y is also from the majority class, do nothing and move to the next sample.

3) If y is from the minority class, check if they form a Tomek Link:

- Find the nearest neighbour of y, call it z.
- If z is the same as x, then (x , y) form a Tomek Link.
  - Remove x (the dominant class sample) from the data repository because it causes overlap between the classes.
  - Repeat this process until no more samples need to be removed.
  - Return the cleaned dataset, which is now better separated and less noisy.

Algorithm 2

Input

A dataset $D = \{ x_1, x_2 \ldots\ldots x_n\}$, where x belongs to either the majority or minority class.

Output

Cleaned dataset D_{clean}.

(1) **Initialize**:

Let $D_{clean} = D$.

(2) **For** each sample $x \in D_{clean}$ where x belongs to the dominant class, **do**:

(3) Locate the nearest point y of x in D_{clean} .

(4) **If** y is the part of dominant class, **then**:

Move to the next instance x and **continue**.

(5) **Else:**

(6) Find the nearest neighbor z of y in D_{clean}.

(7) If $z = x$, **then**:

x and y are nearest neighbors of each other and form a Tomek link.

(8) Remove x from D_{clean}.

(9) Repeat Steps 2–8 until no further modifications occur, or no samples are removed.

(10) **Return** the updated D_{clean}.

End Algorithm

IV. Results and Discussion

A. Data Collection and Experimental Setup

In order to implement Deep Learning-enabled LCDCS system, the Google Collab environment was configured to utilize advanced computational resources. This setup provided robust support for data-intensive operations which included 32 GB of RAM and 1 TB of NVMe SSD storage for faster data handling. For high-performance deep learning and machine learning tasks, an NVIDIA RTX 3080 GPU with 10 GB of GDDR6X VRAM is used. The GPU's architecture enabled efficient parallel processing, significantly reducing training time for large-scale models. Additionally, the GPU-accelerated environment supported real-time experimentation with complex neural networks and computationally expensive tasks, maximizing throughput and performance [19]. This configuration facilitated smooth execution of ML or DL workflows, ensuring scalability and responsiveness for both model development and deployment phases.

In the proposed study, two types of datasets were used to diagnose and forecast lung cancer. The LUNA16 dataset [28], which included more than 1,000 lung CT images in raw DICOM format, served as the source for the initial dataset and a real time dataset, acquired from various stake holders. An annotation file describing the malignant state of each photograph was included. The pictures were saved in PNG format to make processing easier. It included PNG-formatted CT scan images of both healthy people and patients with lung cancer. In all, 979 normal and 1346 malignant pictures were found.

B. Pre-processing Stage

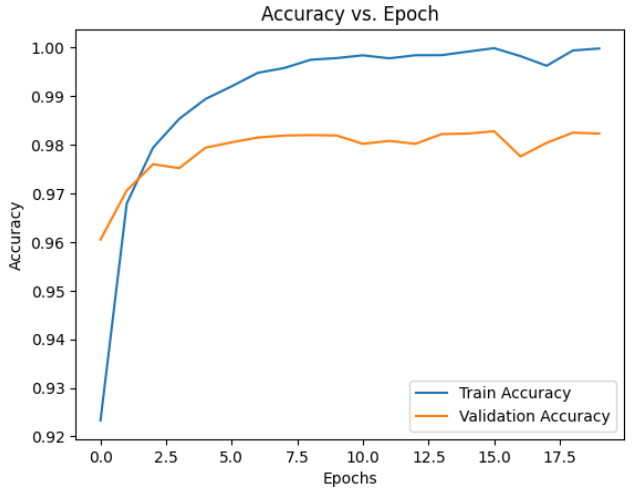
To align with the model architecture, before being input into the CNN model, the original image is converted from BGR and RGBA formats to RGB. Considering that most deep learning models for image classification are trained using RGB images, this conversion is required. After that, the RGB image is scaled to 224 × 24 pixels. Prior to input into the model, the input image is first subjected to a filtering and noise removal process to improve its quality.

C. Results of Feature Extraction and Classification

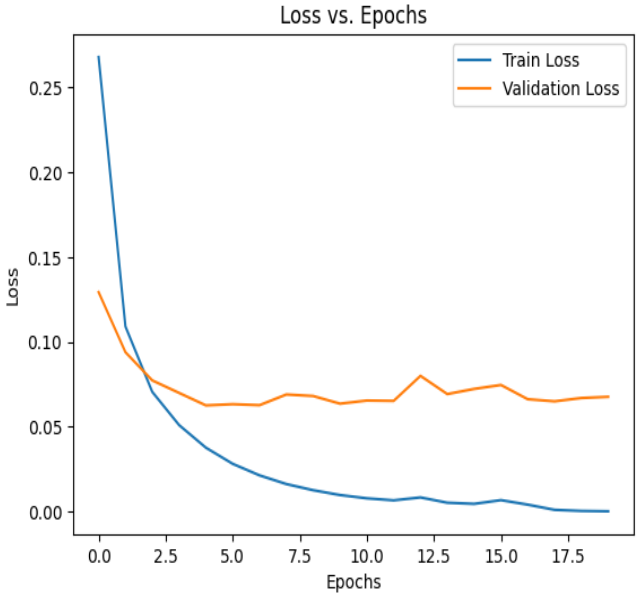
The novel model used deep learning CNN model for the feature extraction purpose and SoftMax function for multi-class classification. The CNN model is built with different layers, Conv2D layer detect important patterns in lung CT scans, such as nodules, textures, and abnormalities and tumour regions (feature extraction). Max pooling layer, it reduces size and focuses on most critical region in the CT scan [18]. Next is flatten and dense layer process which extracted features to classify lung conditions. The model is fine-tuned with feature extraction from the training dataset and is then employed to categorize the testing data, detecting the existence or absence of lung tumors and finally the SoftMax activation provides the final probability distribution over different lung conditions. Fig. 4 illustrates the confusion matrix for the test data, offering a summary of the prediction results obtained by the proposed system.

The model uses a variety of dataset balancing strategies to accomplish classification with and without DL methodology, data augmentation, ADASYN, class-weighted approach, and are some of these methods. The combination of ADASYN and Tomek links works better than the other models, according to an evaluation of the classification findings. The training accuracy of this model is 96.9%. Prior to the application of ADASYN, the dataset reveals a significant class imbalance, with 784 instances representing the minority class (non-cancerous cases) and 10, 10 instances representing the majority class (cancerous cases). Fig. 3 illustrates the enhanced performance of ADASYN, depicting the correlation between

accuracy, loss, and the number of epochs in deep learning. The outcomes using DL are displayed in Fig. 3(a) and Fig. 3(b). Eventually, the classification accuracy of the lung cancer detection module in the suggested system is benchmarked against several existing systems, demonstrating superior accuracy, as illustrated in Table VIII. Further the relationship between training accuracy of different methods are given in Fig. 5, validation accuracy of different methods are given in Fig. 6, and training loss and validation loss with different methods are shown in Fig. 7 and Fig. 8 respectively. Table IX discusses the recovery symptom of NSCLC.



(a)



(b)

Fig. 3. a): Accuracy vs. Epoch with DL, b): Loss Vs. Epoch with DL.

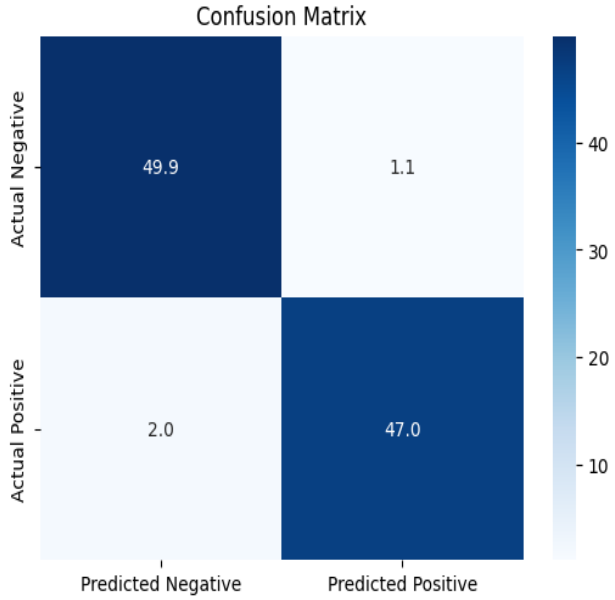


Fig. 4. Confusion matrix.

The model correctly classified mostly non-cancerous case. With 47.06 % most cancer cases are correctly detected.

TABLE VII LUNG CANCER DIAGNOSIS PERFORMANCE COMPARISON: LCDCS SYSTEM WITH DEEP LEARNING AGAINST ALTERNATIVE DATASET NORMALIZATION METHODS WITH AND WITHOUT DEEP LEARNING (DL)

Method	Training accuracy (%)	Validation Accuracy (%)	Training Loss (%)	Validation Loss (%)
Deep Learning enabled LCDCS (ADASYN+ Tomek Link + CNN)	**95.8**	**96.9**	**2.7**	**3.7**
Prox-Smote + CNN	**94.08**	**95.23**	**5.21**	**22.16**
CWA +DL	**93.27**	**94.6**	**11.64**	**8.55**
CWA + CNN	**95.05**	**93.34**	**7.07**	**22.92**
DA + DL	**92.24**	**95.26**	**19.62**	**8.65**
DA +CNN	**85.53**	**78.73**	**31.01**	**39.48**

TABLE VIII COMPARISON OF THE SUGGESTED DEEP LEARNING ENABLED LCDCS SYSTEM'S MODULE WITH CURRENT SYSTEMS

Evaluation Metric	Deep learning enabled LCDCS	Multisection CNN [1]	SVM [11]	3D CNN [12]
Accuracy (%)	**96.9%**	92.17%	92%	83.7%

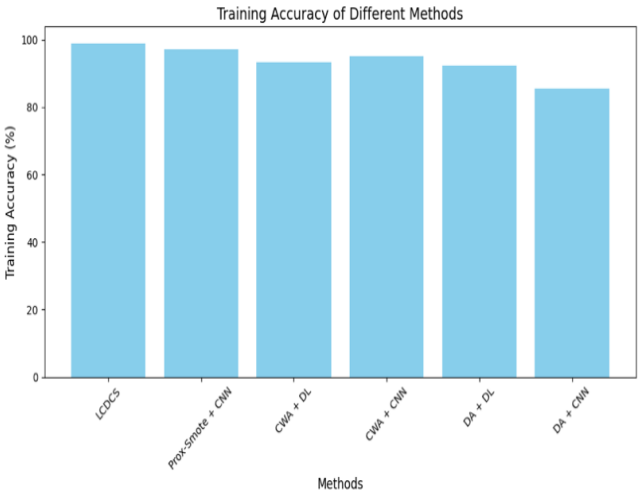


Fig. 5. Training accuracy of different methods.

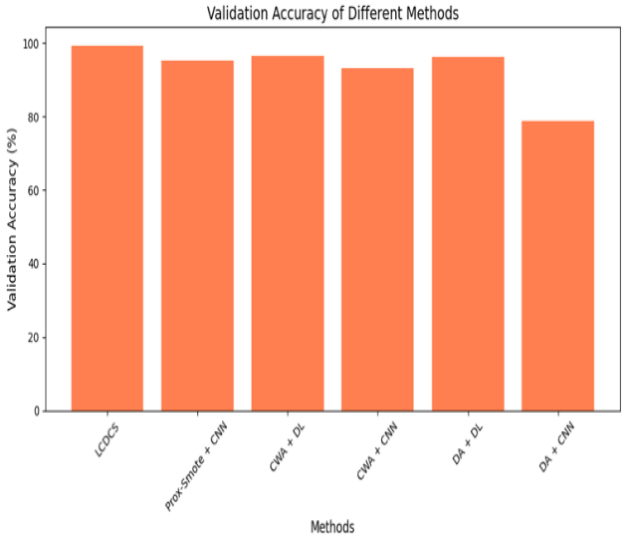


Fig. 6. Validation accuracy of different methods.

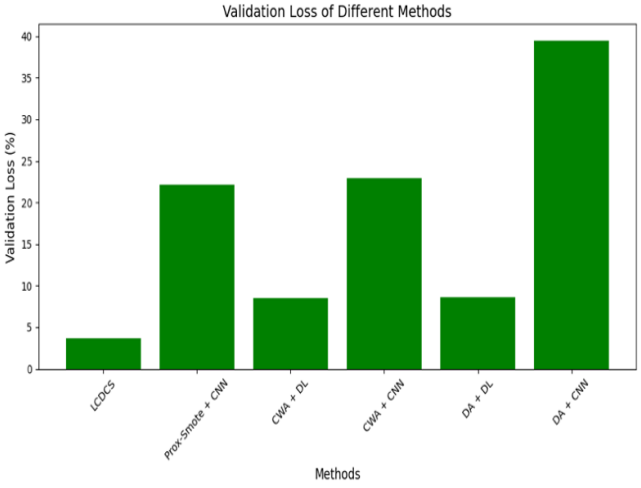


Fig. 7. Validation loss of different methods.

Fig. 8. Training loss of different methods.

TABLE IX RECOVERY SYMPTOM OF NSCLC

Phase	Symptoms/Effects	Management
Immediate	- Fatigue, pain, nausea, appetite loss	- Rest, pain management, anti-nausea meds
Short-Term (1–3 months)	- Dyspnea, cough, weakness, emotional distress	- Pulmonary rehab, physical therapy, counseling
Mid-Term (3–6 months)	- Sleep issues, lymphedema, chest discomfort	- Breathing exercises, compression garments
Long-Term (6+ months)	- Chronic cough, neuropathy, emotional stress	- Long-term rehab, pain management, counseling
Emotional Recovery	- Fear, anxiety about recurrence	- Counseling, support groups

D. Comparison of Model Performance

Evaluating a model is an essential step to determine how well it performs. Various metrics can be used for this purpose, such as accuracy, recall, F1-score, and precision, specificity, FPR each offering different interpretation of the model's effectiveness. TABLE X displays a comparison of the evaluation metrics for the deep learning-enabled LCDCS with those of other machine learning models.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{47.0+49.9}{47.0+49.9+1.1+2.0} = 96.9\%$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{47.0}{47.0+2.0} = 95.9\%$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{47.0}{47.0+1.1} = 97.8\%$$

$$\text{F1- Score} = 2\ X\ \frac{Precision\ X\ Recall}{Precision + Recall} = \frac{97.8*95.9}{97.8+95.9} = 96.8\ \%$$

$$\text{Specificity (True Negative Rate)} = \frac{TN}{TN+FP} = \frac{49.9}{49.9+1.1} = 98.0\%$$

$$\text{False Positive rate (FPR)} = \frac{FP}{TN+FP} = \frac{1.1}{49.9+1.1} = 2.2\%$$

TABLE X COMPARISON OF THE EVALUATION METRICS FOR THE DEEP LEARNING-ENABLED LCDCS WITH THOSE OF OTHER MACHINE LEARNING MODELS

Performance Metric (%)	Deep learning enabled LCDCS	Random Forest[13]	KNN [14]	Logistic Regression[13]
Accuracy	**96.9**	89.5	87.1	90.3
Recall	**95.9**	86.9	80.3	89.6
Precision	**97.8**	89.1	91.1	90.1
F1-score	**96.8**	88.0	85.3	89.9
Specificity	**98.0**	91.5	93.1	90.9
FPR	**2.2**	8.4	6.8	9.0

V. ADVANTAGES

The proposed LCDCS system offers several significant advantages, including high diagnostic accuracy (96.9%) in detecting and classifying NSCLC, making it a reliable tool for supporting early clinical decision-making. The innovative integration of ADASYN and Tomek Links for data balancing resolves class imbalance issues, enhancing model robustness. The system outperforms existing models in both accuracy and operational efficiency.

VI. FUTURE WORK

The study suggests extending the system for broader applications, potentially encompassing additional cancer types or integrating predictive capabilities for associated conditions like cardiovascular disease. Further improvements in interpretability and live implementation are recommended to enhance its clinical applicability.

VII. CONCLUSION

The study concludes that the proposed LCDCS system, powered by deep learning, achieves exceptional accuracy in detecting and classifying NSCLC. By integrating a multi-level convolutional neural network (ML-CNN) with advanced data balancing techniques, the system demonstrates notable accuracy and resilience in handling imbalanced datasets. The incorporation of multi-scale image analysis further enhances the model's ability to detect and classify lung nodules with precision. Through comprehensive comparative evaluation, the research underscores the effectiveness of strategic class balancing in improving diagnostic outcomes. The system proves to be a highly reliable diagnostic tool, offering critical support to radiologists in the early detection of conditions and enabling timely, more effective treatment planning.

REFERENCES

[1] Jabir, K., and A. Thirumurthi Raja. "A Comprehensive Survey on Various Cancer Prediction Using Natural Language Processing Techniques." In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 1880-1884. IEEE, 2022.

[2] AbuSamra, Aiman Ahmad, and Areej MR Al-Madhoun. "Applying Deep Learning and Natural Language Processing in Cancer: A Survey." In *2021 Palestinian International Conference on Information and Communication Technology (PICICT)*, pp. 103-115. IEEE, 2021.

[3] Juhn, Young, and Hongfang Liu. "Artificial intelligence approaches using natural language processing to advance EHR-based clinical research." *Journal of Allergy and Clinical Immunology* 145, no. 2 (2020): 463-469.

[4] Devyani Rawat, Sachin Sharma, Shuchi Bhadula, "Case Based Reasoning Technique in Digital Diagnostic System for Lung Cancer Detection", In 2023 8th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2023.

[5] Menasalvas Ruiz, Ernestina, Juan Manuel Tuñas, Guzmán Bermejo, Consuelo Gonzalo Martín, Alejandro Rodríguez-González, Massimiliano Zanin, Cristina González de Pedro et al. "Profiling lung cancer patients using electronic health records." *Journal of Medical Systems* 42 (2018): 1-10.

[6] Devyani Rawat, Sachin Sharma, and Shuchi Bhadula. "Digital Clinical Diagnostic System for Lung Cancer Detection." In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), pp. 535-540. IEEE, 2023.

[7] Wang, Shidan, Donghan M. Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao. "Artificial intelligence in lung cancer pathology image analysis." *Cancers* 11, no. 11 (2019): 1673.

[8] Chen, Po-Hao. "Essential elements of natural language processing: what the radiologist should know." Academic radiology 27, no. 1 (2020): 6-12.

[9] Devyani Rawat, Sachin Sharma, and Shuchi Bhadula. "Deep Learning Techniques in Digital Clinical Diagnostic System for Lung Cancer." In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 1232-1237. IEEE, 2023.

[10] Nageswaran, Sharmila, G. Arunkumar, Anil Kumar Bisht, Shivlal Mewada, JNVR Swarup Kumar, Malik Jawarneh, and Evans Asenso. "[Retracted] Lung Cancer Classification and Prediction Using Machine Learning and Image Processing." *BioMed Research International* 2022, no. 1 (2022): 1755460.

[11] Puts, Sander, Martijn Nobel, Catharina Zegers, Iñigo Bermejo, Simon Robben, and Andre Dekker. "How Natural Language Processing Can Aid With Pulmonary Oncology Tumor Node Metastasis Staging From Free-Text Radiology Reports: Algorithm Development and Validation." JMIR Formative Research 7 (2023): e38125.

[12] Wang, Liwei, Lei Luo, Yanshan Wang, Jason Wampfler, Ping Yang, and Hongfang Liu. "Natural language processing for populating lung cancer clinical research data." BMC medical informatics and decision making 19 (2019): 1-10.

[13] Gupta, Khushbu, Ratchainant Thammasudjarit, and Ammarin Thakkinstian. "NLP automation to read radiological reports to detect the stage of cancer among lung cancer patients." In WNLP@ ACL, pp. 138-141. 2019.

[14] Do, Richard KG, Kaelan Lupton, Pamela I. Causa Andrieu, Anisha Luthra, Michio Taya, Karen Batch, Huy Nguyen et al. "Patterns of metastatic disease in patients with cancer derived from natural language processing of structured CT radiology reports over a 10-year period." Radiology 301, no. 1 (2021): 115-122.

[15] Negi, Shubham, Poornima Mittal, and Brijesh Kumar. "Modeling and analysis of high-performance triple hole block layer organic LED based light sensor for detection of ovarian cancer." IEEE Transactions on Circuits and Systems I: Regular Papers 68, no. 8 (2021): 3254-3264.

[16] Guan, Qing, Xiaochun Wan, Hongtao Lu, Bo Ping, Duanshu Li, Li Wang, Yongxue Zhu, Yunjun Wang, and Jun Xiang. "Deep convolutional neural network Inception-v3 model for differential diagnosing of lymph node in cytological images: a pilot study." Annals of translational medicine 7, no. 14 (2019): 307.

[17] Sahu, Pranjal, Dantong Yu, Mallesham Dasari, Fei Hou, and Hong Qin. "A lightweight multi-section CNN for lung nodule classification and malignancy estimation." IEEE journal of biomedical and health informatics 23, no. 3 (2018): 960-968.

[18] Thanoon, Mohammad A., Mohd Asyraf Zulkifley, Muhammad Ammirrul Atiqi Mohd Zainuri, and Siti Raihanah Abdani. "A review of deep learning techniques for lung cancer screening and diagnosis based on CT images." Diagnostics 13, no. 16 (2023): 2617.

[19] Sundar, R., Sudhir Ramadass, D. Meeha, Balambigai Subramanian, S. Siva Shankar, and Gayatri Parasa. "Evaluating the Solutions to Predict the Impact of Lung Cancer with an Advanced Intelligent Computing Method." In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1733-1737. IEEE, 2023.