National College of Ireland

Lung Cancer Detection Using Machine Learning and Deep Learning

MSc ResearchProject
MSc in Data Analytics

Yash Kutaphale
Student ID: X21195960

School of Computing
National College of Ireland

Supervisor: Furqan Rustam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Yash Kutaphale
Student ID:	X21195960
Programme:	Master of Science in Data Analytics
Year:	2022-23
Module:	MSc Research Project
Supervisor:	Furqan Rustam
Submission Due Date:	14/05/2023
Project Title:	Lung Cancer Detection Using Machine Learning and Deep Learning
Word Count:	8622
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Yash Kutaphale
Date:	14th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Lung Cancer Detection Using Machine Learning And Deep Learning

Yash Kutaphale
X21195960

Abstract

Machine learning (ML) and deep learning (DL) have been combined to create a new way to diagnose lung cancer, which is the top cause of death around the world. In order to improve the accuracy of medical CT scan interpretation and solve issues brought on by the complex nature of these scans, this research presents a novel method that combines deep learning and conventional machine learning. Our work, used a large set of CT pictures from different groups of people in central Iraq. To this dataset, two different techniques were used. The initial stage focused on the cooperative interaction of several ML algorithms, highlighting the significance of careful data calibration. This served as the foundation for a later, more sophisticated approach built on a Convolutional Neural Network (CNN) originally intended for the classification of retinal images. The training accuracy for this CNN methodology was astonishingly high at 98.89% after balancing the data using SMOTE. Our study's central thesis is that combining these computer models with conventional diagnostic techniques has the potential to transform lung cancer diagnosis by giving radiologists effective tools for early and precise identification. Additionally, we demonstrate the revolutionary potential of AI in transforming the landscape of medical picture analysis by contrasting our techniques with current studies.

1 Introduction

Globally, lung cancer continues to be a major health concern. Lung cancer was the main cause of cancer-related deaths in 2020, with over 1.8 million deaths, according to the World Health Organization (2020) [1]. Despite significant improvements in medical imaging and diagnostics [2] lung cancer early diagnosis remains difficult. Treatments are frequently postponed as a result, which has a negative impact on patient outcomes [3]. For instance, the five-year survival rate for lung cancer cases recognized at an early stage might be as high as 56% [4], whereas it substantially decreases for instances discovered at a later stage. With the speed at which technology is developing, there is an increasing drive to use computational tools to support the diagnostic procedure. Our project aims to investigate, assess, and improve the integration of artificial intelligence with conventional diagnostic modalities for more precise lung cancer diagnosis in light of this technological transition.

1.1 Background

One of the most common cancers to be diagnosed is lung cancer, which is also the leading cause of cancer-related death. In order to increase survival rates as well as patient

prognoses, early identification of lung cancer is crucial. Although useful, traditional diagnostic techniques frequently have drawbacks such as invasiveness, delays, and sporadic misdiagnoses. Enter the world of deep learning and artificial intelligence, which promise to revolutionize the way lung cancer is detected. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated exceptional ability in processing medical imagery like CT scans by utilizing computing prowess and methods that mirror human learning processes. These models provide an effective, quick, and non-invasive diagnostic tool by being able to spot patterns, anomalies, and subtle changes that are frequently missed by the human eye. In order to establish the most efficient strategy for accurately and precisely diagnosing this potentially fatal condition, this research explores the potential of deep learning for lung cancer detection.

The manual interpretation of these images has its own set of difficulties, though. First, the sheer number of scans that a radiologist must review can be exhausting, which can result in exhaustion and mistakes. Second, because manual interpretation is subjective, different radiologists may arrive at different diagnoses. These difficulties highlight the need for automated tools that can supplement or possibly replace human expertise in the diagnosis of lung cancer.

1.2 Problem Statement

Traditional diagnostic techniques sometimes fall short in the field of medical imaging because the vast volume of CT scans is mixed with the subtle complexities of each image. This paper presents a novel strategy to close this gap by combining the strengths of deep learning and conventional machine learning. These state-of-the-art computational approaches offer a fresh response to the long-standing problems of automatic and accurate detection in medical CT scans, despite being well-established in other fields.

1.3 Research Objective

The combination of computational methods with conventional diagnostics has enormous potential in the field of medical imaging. The main goal of this study is to investigate this synergy. We intend to extract complex texture-based information from CT scan images using the Gray Level Co-occurrence Matrix (GLCM). These characteristics, which are frequently invisible to the human eye, can include important details about the presence and stage of malignancies. The goal then changes to model selection and optimization using the retrieved features at hand. We assess a range of algorithms that properly classify the CT scan pictures, from basic machine learning models to sophisticated deep learning architectures. With this project, we hope to increase lung cancer detection accuracy while also streamlining the diagnostic procedure to make it quicker and more effective.

- To learn more about the possibilities of merging computational techniques with traditional diagnoses in the field of medical imaging.
- Create a method for exploiting the Gray Level Co-occurrence Matrix (GLCM) to extract intricate texture-based data from CT scan pictures.
- Assess various algorithms for accurately classifying CT scan pictures, including both machine learning and deep learning models.
- Simplify the diagnostic procedure for improved efficiency and precision and improve the accuracy of lung cancer diagnosis using the suggested approach.

- Assist radiologists and lessen the possibility of diagnostic errors by developing an automated, effective, and exact system for lung cancer detection. This will help to transform diagnostics.
- Provide a technology that can help in rapid and accurate diagnosis to address the need for dependable diagnostic tools in disadvantaged areas.
- Significantly advance the use of artificial intelligence in the field of medical imaging by providing standards, recommendations, and insights based on evaluations of various algorithms.

2 Related Work

Author did a thorough study of how deep learning can be used to diagnose lung cancer. They understood how important lung cancer was by pointing out that over 200,000 new cases are found there each year. The disease is marked by uncontrolled growth of lung cells, which leads to the formation of deadly tumors. The authors' review gave special attention to convolutional neural networks (CNNs) for lung cancer diagnosis. They underlined that CNNs have outperformed other conventional methods and have showed promising outcomes in automatically detecting the condition. The evaluation brought attention to the disparate accuracy and sensitivity rates of various models, particularly in the early-stage diagnosis of lung cancer [5].

The importance of early lung cancer diagnosis and the potential of machine learning in assisting this process are stressed by the authors of the paper [8]. In particular, the research looks at the properties of big data and its application to CT and X-ray imaging systems. The application of important machine learning techniques including Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes, and Logistic Regression in the field of lung cancer prediction is highlighted in the details of these methods. The authors stress the revolutionary potential of combining varied data with machine learning methods in cancer prediction [8] in their conclusion.

In the study [9], the team used the National Lung Screening Trial (NLST) dataset, which included 42,290 CT cases from 14,851 patients, to develop a deep learning model capable of anticipating the diagnosis of lung cancer from CT scans. Their model, which was composed of four primary components—lung segmentation, cancer Region of Interest (ROI) detection, full-volume analysis, and a concluding cancer risk prediction—used cutting-edge techniques including Mask-RCNN and RetinaNet for precise identification. Clinical validation using 6,716 cases from NLST and comparisons against six radiologists showed the model's capacity to enhance lung cancer screening. Despite its potential, the model's training was complicated because it was based on vast volumes of data and CT scans are complicated [9]. They offered the data sources as well, but not the confidential code.

This study [10] utilized both machine learning and deep learning techniques to predict the survival or death status of lung cancer patients based on the SEER database. Employing a range of classifiers including logistic regression, Bayes classifier, lazy-classifier, meta-classifier (ASC), rule-learner, and decision-tree, the research found the ASC meta-classifier to be the most accurate among traditional algorithms with an accuracy of 88.51%. However, a deep neural network (DNN) model slightly outperformed it with an accuracy of 88.58%. In comparison to previous research, which achieved

71.18% accuracy using ANN, this study demonstrated superior results. Notably, predictions were based on demographic and clinical features, as opposed to imaging techniques. The study, however, highlighted the limitations of the SEER database in providing detailed therapy specifics [10].

Deep learning was used in the research [11] to forecast overall survival using CT scan data from 268 individuals with stage III NSCLC. The researchers examined 739 scans by combining ResNet CNNs and an RNN. The model's accuracy for predicting 2-year survival for the primary dataset (Dataset A) of 179 patients receiving definitive radiation treatment increased with each subsequent scan, reaching an AUC of 0.74 at 6 months. The model correctly predicted distant metastasis, progression, and locoregional recurrence in 89 patients who were part of an external test set (Dataset B) and received trimodal treatment. The model also differentiated between responders and non-responders for pathologic response prediction, with an AUC of 0.65 [11].

Lung cancer research has greatly benefited from computer-aided pathology and imaging diagnosis, with some models obtaining over 90% classification accuracy. In particular, deep learning has shown potential. With a remarkable AUC of 0.83 for the EGFR mutation, Coudray et al. used the inception-v3 model on TCGA histopathological images to predict mutations in six lung cancer genes. In NSCLC patients, utilized a CNN to map cell distributions, while, deep learning was used to identify critical immune cells. Nevertheless, issues like image heterogeneity and a lack of training datasets continue [13].

The length of stay (LOS) in intensive care units (ICUs) for patients with lung cancer was predicted using machine learning in the research utilizing the MIMIC-III v1.4 dataset from Harvard Medical School. Following extensive feature selection and data preparation, the researchers used six class-balancing strategies to address data imbalance. Random Forest (RF), XGBoost, and Logistic Regression (LR), three machine learning models, were evaluated. The RF model demonstrated good results when paired with class balancing techniques like ADASYN and SMOTE. The SHAP approach was used to make the forecasts easier for healthcare practitioners to understand [14].

AI and ML models were used in a variety of domains in [15] on the usefulness of AI in lung cancer immunotherapy prediction. For long-term PFS advantages in PD-1/PD-L1 immunotherapy candidates, radiomics-based AI models displayed AUCs between 0.75 and 0.82 [15]. An AUC of 0.938 was notable for a Random Forest model. High PD-L1 expression was predicted with an AUC of 0.76 using a deep convolutional neural network. In the field of proteomics, a machine learning system predicted that lung cancer patients will have a 5-year survival rate of 7.6% by looking at serum autoantibody indicators. Prediction models with remarkable AUCs of 0.95 were produced by investigations of the gut microbiota. Neural network models produced a positive predictive value of 76.5% when used to forecast the negative effects of PD-L1 treatment. This study highlights how AI may improve the precision of forecasts for lung cancer immunotherapy [15].

A review of 30 papers on the use of clinical data to predict lung cancer survival presented a variety of data preprocessing and machine learning (ML) methods [17]. In contrast, feature selection was taken into account in 63.3% of the research, with age, sex, and N stage being the most often chosen features. Only 40% of the studies went into depth about handling missing information. The most common machine learning algorithm was Random Forest (RF), which was used in 56.6% of the investigations. K-fold cross-validation was utilized for model evaluation by 40%, while data splitting was chosen by

53.3%. Performance metrics varied, with RMSE being employed in 23.3% of the studies, AUC in 60.6% of them, and the C-statistic in 13.3%. The forecast periods ranged from six months to five years, although the summarized portions did not include exact accuracy values for each model [17].

Authors published a significant work that clarified deep learning's capability for handling large datasets, including radiographic pictures [18]. With a focus on the architectural quirks of Convolutional Neural Networks (CNNs) and their skill at image recognition tasks, their paper provided a thorough overview of deep learning.

However, using CNNs for medical imaging is not without its difficulties. Author In their ground-breaking paper [19] provided a comprehensive exploration of various CNN architectures customized for various diagnostic situations. This study also shed light on the crucial function of transfer learning, particularly in cases where datasets are limited.

Medical imaging's complexities and difficulties have been painstakingly documented. Authors in [20] provided a perceptive analysis of the issues brought on by data scarcity and strong class inequities. Their observations also highlighted the encouraging progress achieved in overcoming these difficulties by deep learning approaches like data augmentation and transfer learning.

The Author's survey provides a comprehensive summary [21] stands out. This thorough investigation shed light on a variety of methods and their particular uses in the field of medical imaging while also highlighting potential dangers, serving as a virtual road map for researchers navigating this complex area. The author implemented CNN, RNN, Autoencoders.

Over the years, a number of empirical investigations have shown the practical uses and consequences of deep learning in radiography analysis. The work by [22] is one significant example because it outlined a comprehensive approach that included model development, validation, and final clinical implementation. The researchers achieved accuracy of 97.5% using deep learning models.

Another ground-breaking investigation by [23], with findings correlating with the knowledge of seasoned dermatologists. Using Deep Neural Network they achieved accuracy of 91%.

The vastness, complexity, and constant evolution of deep learning in radiographic image processing are what make it such a fascinating tapestry. The references mentioned here are but a sample of this vast field. The need for academics and professionals to keep up with new advances remains undiminished as the area continues to grow.

2.1 Overview of Models and Accuracies from Related Works

In the table below, the models used by different researchers and the accuracy achieved are shown.

Reference No.	**Year**	**Models**	**Achieved Accuracy**

[8]	2018	CNN	98
[9]	2019	3D CNN	AUC = 94.5
[10]	2023	Logistic Regression, Decision Trees, Random Forest, SVM, MLP, CNN	CNN 97.6
[11]	2019	Deep Learning	AUC = 77
[13]	2020	CNN	94.53
[14]	2022	Decision Trees, Random Forest, Gradient Boosting, AdaBoost, and Neural Networks	MAE = 2.02 for Gradient Boost
[21]	2017	CNNs, RNNs, Autoencoders	N/A
[22]	2016	Deep Learning	97.5
[23]	2017	Deep Neural Network	91

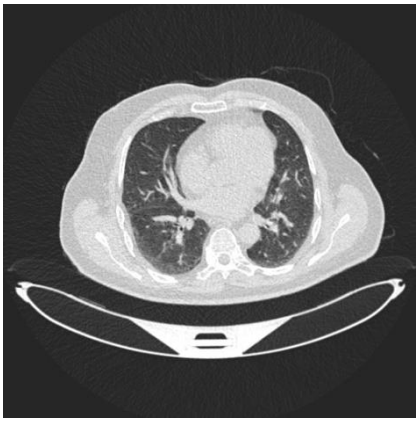
Table 1: Summary of models.

3 Methodology

3.1 Dataset

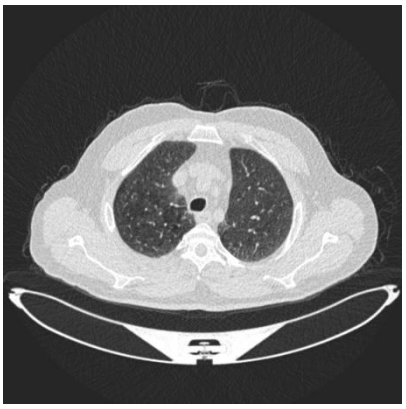
The IQ-OTH/NCCD lung cancer dataset, produced by the National Center for Cancer Diseases and the Iraq-Oncology Teaching Hospital in collaboration, is a symbol of thorough medical data gathering. This dataset was painstakingly put together over the course of three months in the fall of 2019. Through the Kaggle platform, a recognized hub for data science and machine learning resources, it is now made available to the entire audience [29].

This collection is rich in CT scans that depict both healthy lung conditions and lung cancer in all its complexity throughout its various phases. Each scan received meticulous annotation and examination thanks to the commitment of skilled oncologists and radiologists from the contributing centers. The final collection contains 1190 CT scan pictures that were obtained from 110 unique cases. These have been systematically divided into three categories: 40 malignant growths, 15 benign tumors, and 55 normal cases [29].



[A] Bengin Case

[B] Malingnant Case



[C] Normal Case

Figure 1: Types of Cases Detected in CT Scan

3.2 Methods

The methodology acts as a roadmap outlining each action, choice, and approach used during the study and is frequently considered the foundation of a research project. It is a methodical strategy used to guarantee the scientific rigor, coherence, and consistency of the research process. A clearly defined approach might mean the difference between a game-changing discovery and a lost opportunity, particularly in the computational and medical fields.

For the subsequent modeling procedures to be strong and reliable, the data collection phase was of utmost importance. The dataset, a comprehensive collection of CT scan pictures, was obtained from Kaggle, a well-known website renowned for its enormous repository of datasets covering a variety of areas. The dataset was large and of good quality, giving a representative sample of real-world cases because Kaggle was used as the main data source. The research was built on a firm foundation created by the rigorous approach to data collecting, which also ensured that the models developed would be capable of handling the complexities and nuances present in medical imaging data.

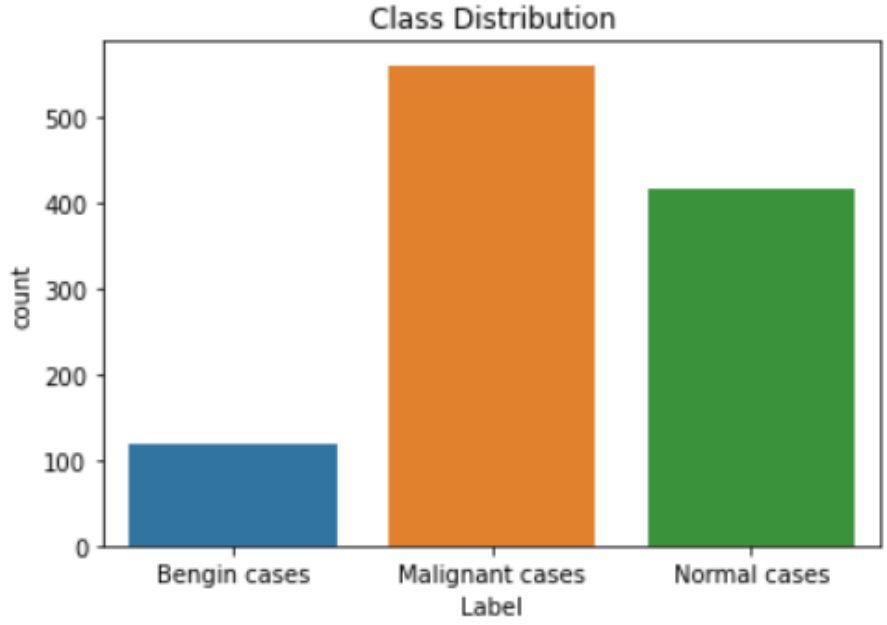


Figure 2: Class Distrbution Bar Chart

In Approach 1, a number of machine learning models are built to perform the task of categorizing CT scan images. The Convolutional Neural Network (CNN) model is one of the models used. A single hidden layer with 128 neurons and a softmax output layer make up this CNN. The number of chosen GLCM (Gray-Level Co-occurrence Matrix) features is represented by the input shape. On the retrieved GLCM characteristics, the CNN is trained and assessed.

Support Vector Machines (SVM), Random Forest, Decision Trees, Gradient Boosting Machines, k-Nearest Neighbors (KNN), Logistic Regression, and variations of neural networks like ResNet and EfficientNet are more models that are covered in Approach 1. The architectural complexity of these models varies, with some relying on decision trees and others, like ResNet and EfficientNet, using convolutional layers. The models are trained on the original unbalanced dataset of CT scan pictures, and it should be noted that SMOTE is not used in this approach.

Importing the libraries that are essential to the methodology is the first stage in this strategy. Python's core libraries for numerical operations and data manipulation, respectively, are numpy and pandas. The os library makes it easier to communicate with the operating system, particularly when working with files and directories.

The skimage library, in particular its graycomatrix and graycoprops functions, is essential for image processing. In order to compute the Gray Level Co-occurrence Matrix (GLCM) and subsequently extract attributes from the matrix, these functions are essential. GLCM is a statistical technique for extracting texture information from photographs by looking at the spatial relationship between pixels.

Tensorflow.keras was used to facilitate deep learning operations. A variety of tools for creating, honing, and assessing neural networks are available in this library. Sklearn was

also heavily used for computation of metrics, feature selection, data preprocessing, and typical machine learning models.

The methodology begins by establishing the basic tools before loading and analyzing a series of CT scan pictures from a pre-defined directory. Each image goes through a number of changes. First off, the reading is in grayscale, which gets rid of any potential complications from color channels. The data is then transformed from the picture to an unsigned integer type to make sure it can be used for the remaining operations.

The computation of the GLCM is the section's most important phase. This matrix measures the frequency with which particular pairings of pixel values appear in a given spatial relationship. In essence, it's a technique for preserving the texture and patterns found in the photographs. The "contrast" characteristic is extracted from this matrix. Not all features, though, are equally informative. As a result, a feature selection method employing the ANOVA F-value is used, with the goal of keeping only the features that have the highest statistical significance.

Understanding the data is crucial before beginning any modeling. The preprocessed data is transferred to a DataFrame format, which makes it simpler to analyze and rich with extracted features. In EDA, visualization is a potent tool. Several graphical representations of the data are plotted using seaborn and matplotlib. The distribution of classes, pairwise correlations between features, and boxplots for specific feature distributions across classes are all included in this. The distribution of the data, any potential correlations, and any abnormalities or outliers are all revealed by these visualizations.

To be utilized in modeling, data must be in a certain format. To allow for model evaluation on unlabeled data, the dataset is split into training and test sets. Additionally, attributes that may be on various scales are standardised using StandardScaler from sklearn.

The first approach uses machine learning techniques and deep learning frameworks to model data comprehensively. The approach leverages on the Convolutional Neural Network's (CNN) built-in benefits for handling picture data.

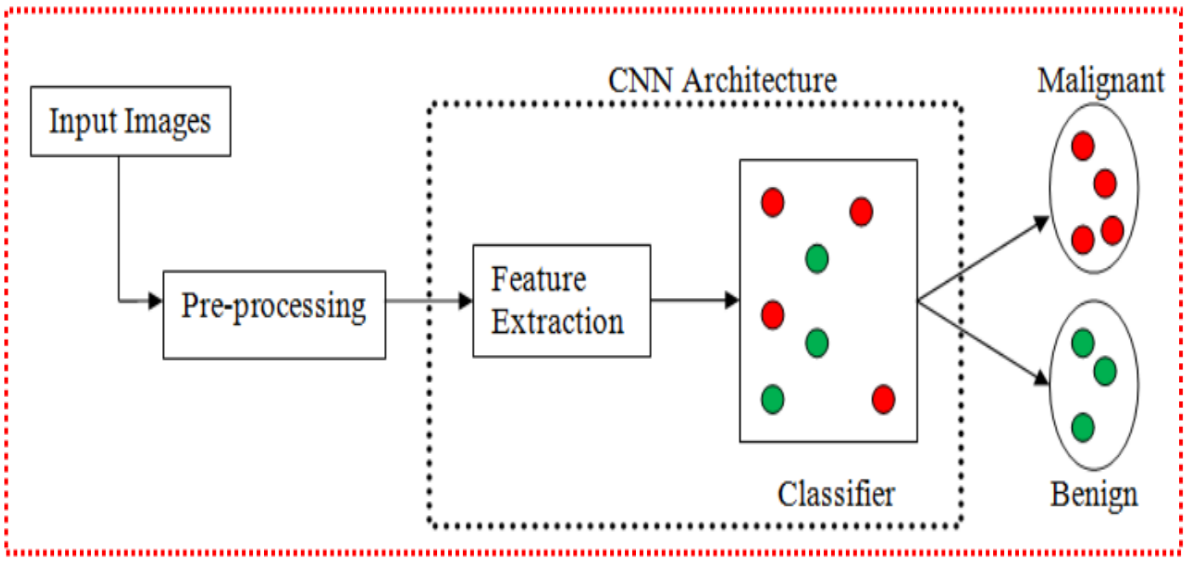


Figure 3: CNN Architecture [21].

However, the methodology diversifies its modeling methods to ensure a strong and thorough comprehension of the CT scan images. SVM, Random Forest, Decision Tree, Gradient Boosting Machine, k-NN, and Logistic Regression are just a few of the well-known machine learning models that are used.

The technique also uses deep learning models like ResNet and EfficientNet. To ensure that the complexities and intricacies of medical imaging data are accurately captured, such thorough modeling is justified.

Evaluation is still a key component of this strategy. Accuracy, precision, recall, and F1-score are just a few of the metrics used to thoroughly evaluate each model's performance.

In short, Data preprocessing and the extraction of Gray-Level Co-occurrence Matrix (GLCM) texture properties, in particular "contrast," from medical images are the first steps in Approach 1. The most useful features are then kept after these are subjected to feature selection using the ANOVA F-value. Then, a variety of machine learning models are used, including transfer learning models like ResNet and EfficientNet, Convolutional Neural Networks (CNN), Support Vector Machine (SVM), Random Forest, Decision Tree, Gradient Boosting, and k-Nearest Neighbors (k-NN). The architecture, activation functions, learning rates, and batch sizes that are specific to each model have been optimized for optimum performance.

In Approach 2, creating a Convolutional Neural Network (CNN) model tailored for image classification tasks is the main goal. Two convolutional layers make up this CNN architecture, which is followed by max-pooling layers for feature extraction. The feature maps are then made flat, and a dense hidden layer made up of 16 neurons is added. Three neurons in the output layer with softmax activity indicate the classification categories. All photos are normalized by scaling pixel values to the range [0, 1] and resized to a set size of 256x256 pixels to ensure data homogeneity.

The use of Synthetic Minority Over-sampling Technique (SMOTE) to alleviate class imbalance is one notable feature of Approach 2. In order to create synthetic examples for the minority classes, SMOTE is applied using the imblearn library. This balancing method helps the model handle uneven data better and improves classification results, especially for underrepresented classes. After training on this SMOTE-augmented dataset, the CNN model is subsequently tested on a different validation dataset. The performance of the model during training is also evaluated and visualized using visualization tools like confusion matrices and training curves.

While building learning models for CT scan image classification is a task shared by both approaches, Approach 1 stresses a variety of models without data balancing, and Approach 2 emphasizes the usage of a CNN model with SMOTE for better performance on unbalanced data. The latter illustrates how important it is to correct class imbalance in order to classify medical images more precisely.

The method begins by importing a large number of libraries essential to its functioning. The fundamental packages for handling numerical calculations and data processing are still numpy and pandas. The os and collections libraries simplify data structures and directory operations.

A combination of libraries, including cv2, imageio, matplotlib, PIL, and seaborn, are used for image processing and visualization. A library for real-time computer vision is called cv2 (OpenCV). A variety of picture data can be read and written with ease with

imageio. Together, they enable the method to modify and process images in an efficient manner.

Sklearn, Imblearn, Tensorflow, and Keras facilitate the core of the modeling step. A machine learning library called sklearn offers quick and effective tools for data analysis. A sizeable portion of the machine learning pipeline is covered. The imblearn library is used to identify probable class imbalance, specifically its Synthetic Minority Over-sampling Technique (SMOTE).

The dataset's images go through a number of preparation procedures. First, these CT scan-derived images are loaded, and the size discrepancies across different categories are carefully examined. To ensure uniformity in input dimensions, each image is downsized to a resolution of 256 by 256 pixels. The methodology also looks at Gaussian blurring, a method for reducing noise and features in photographs and possibly bringing out the most important patterns.

Each image's pixel values are normalized during post-processing to fall within the range of 0 and 1. The model will only accept values that fall inside a consistent scale thanks to this normalization. The dataset is then divided into subgroups for training and validation. The use of SMOTE, which corrects potential class imbalances in the dataset by producing synthetic examples, is a crucial step in this phase.

The Convolutional Neural Network is at the core of this strategy (CNN). Two convolutional layers and a max-pooling layer make up the chosen architecture. These layers are intended to identify and highlight the key elements of the photos. The 2D characteristics are subsequently converted into a 1D vector via the network's flattening layer. After that, a subsequent dense layer assures further pattern extraction before the last dense layer divides the input into one of the three groups. The model is then tested on the validation set after training on the SMOTE-augmented data. To evaluate its performance, metrics including accuracy, precision, recall, and F1-score are generated.

In short, Utilizing a Convolutional Neural Network (CNN), picture classification is the main goal of Approach 2. Data preprocessing, which includes scaling photos, leveling pixel values, and addressing class imbalance by SMOTE oversampling, is the first step in the procedure. Convolutional layers, max-pooling, dense layers, and an output layer make up the CNN model's architecture. There are defined hyperparameters for layer configurations, dropout rates, and optimizer selections. On the preprocessed data with the predetermined batch size and epochs, the model is trained. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score. To maximize model parameters, hyperparameter tuning, training curve visualization, and maybe the usage of methods like grid search are all used.

Both approaches provide a thorough and rigorous way for identifying and evaluating CT scan pictures. From data loading to model evaluation, every step is carried out precisely to guarantee the outcomes' robustness and dependability.

Following a thorough analysis of both techniques, Approach 2 was principally chosen due to its potential to provide a model with a high degree of accuracy. The second approach, which emphasizes convolutional neural networks and careful preparation steps, demonstrated an improved capacity to detect subtle patterns in CT scan pictures, whereas the first methodology provided a broad spectrum of insights through various modeling strategies. The pursuit of optimal model performance, which ensures that the classification of crucial medical images achieves the highest degree of accuracy, underlay

the choice to switch to Approach 2, increasing its application and reliability in real-world medical scenarios.

4 Design Specification

The architecture of the system is expressed through a methodical approach that carefully combines cutting-edge computer techniques for the complex task of radiographic image analysis. The use of deep learning architectures, which are specifically specialized for image datasets and have the inherent capacity to capture the granular and complicated information contained in radiography pictures, is at the core of this approach.

The Convolutional Neural Network (CNN) provides the main support for the integrated models. Convolutional filters are used by CNNs, which are renowned for their proficiency in image classification tasks, to accurately identify spatial hierarchies inside images. They excel at identifying increasingly abstract patterns, which are essential for efficient radiographic image processing, thanks to their multi-layered architecture.

The Support Vector Machine (SVM) is a method that supports deep learning. SVMs are essential because of their shown proficiency in high-dimensional spaces, which are a defining feature of picture data. They are especially well suited for binary classification requirements because they can choose the best hyperplane that clearly divides data into two categories.

The Random Forest method enhances the ensemble much more. This approach of ensemble learning builds numerous decision trees and combines their results, providing a strong defense against overfitting and ensuring a more robust and dependable model.

The Gradient Boosting Machine is a key element in the system's architecture. By sequentially incorporating weaker learners, this iterative approach gradually corrects flaws from past models to steadily improve prediction accuracy.

While the aforementioned models are the foundation of the design, other models, such as k-Nearest Neighbors (KNN), are used because of their non-parametric character, which is extremely useful when the dataset displays a logical structure. Due to its mathematical rigor and capacity to estimate categorical probabilities, the Logistic Regression model, normally used for binary classification issues, has also been included.

EfficientNet is one of the most recent and promising enhancements to the concept. It has redefined accuracy and efficiency benchmarks by utilizing compound scaling, making it a crucial part of radiographic image studies.

Beyond the models, the design places a focus on painstaking data preprocessing. Techniques like Gaussian blurring are used to improve feature detection since it is crucial to maintain high data quality across all images. In order to avoid potential hazards of class imbalances, the Synthetic Minority Over-sampling Technique (SMOTE) has also been implemented, guaranteeing a fair representation in the training data.

A variety of indicators are used to rate the effectiveness of this complex design. Each metric offers insights into various aspects of the system's effectiveness, from the RMSE, which gauges the prediction accuracy in regression tasks, to the AUC, which assesses classification abilities, and the C-statistic, which gauges survival analysis.

Fundamentally, the concept is a carefully coordinated synthesis of cutting-edge computational models and data-centric approaches that aim to maximize diagnostic precision in radiographic imaging analysis. This all-encompassing strategy guarantees that the complex radiographic data, model construction, and preprocessing facets synchronize flawlessly to give accurate and high-fidelity outcomes.

5 Implementation

The research's implementation phase represents the change from abstract ideas to practical recommendations. On this trip, a wide range of cutting-edge technologies came together, setting the groundwork for the complex models and algorithms created for retinal picture classification. A collection of potent libraries, including scikit-image, scikit-learn, and TensorFlow, were used to give life to our models using Python's reliability as the core programming language.

The methodical use of the ImageDataGenerator from the TensorFlow and Keras ecosystems was essential to the implementation process. The available dataset was greatly increased thanks to this technology, which was created for the purpose of augmenting image data. ImageDataGenerator guarantees that the models are trained on a wide set of data, enhancing their generalization skills, by introducing controlled modifications like as rotations, flips, and zooms.

A key method for feature extraction emerged: the Gray Level Co-occurrence Matrix (GLCM). Utilizing the scikit-image package, this technique examines spatial correlations in grayscale photos to extract crucial textural data. These properties serve as the foundation for subsequent models, capturing elements including contrast, dissimilarity, homogeneity, and energy.

The CNN was painstakingly adjusted for optimum performance in retinal image categorization thanks to its natural capacity to recognize hierarchical patterns in images. With the former excelling in high-dimensional spaces and the latter utilizing an ensemble of decision trees for reliable classifications, Support Vector Machines (SVM) and Random Forests each brought their own specialties to the table. The model ensemble was further extended by algorithms including Decision Trees, Gradient Boosting Machines, k-Nearest Neighbors (KNN), Logistic Regression, and the sophisticated EfficientNet, each of which brought with it unique benefits.

Data pretreatment became crucial to the smooth execution of these models. This required interpreting the photos, transforming them into a format suitable for machine learning models, and normalizing them. Images were scanned, enhanced, and their features were retrieved using the scikit-image package. The next step was normalization, which is crucial for ensuring model convergence and top performance.

The evaluation was the last stage of the implementation process. The performance of each model was carefully examined using a number of criteria, including accuracy, precision, recall, and the F1-score. These measures, which were put into practice using scikit-learn, provided a comprehensive view of each model's strengths and potential development areas, directing further iterations and improvements.

Collaboration tools like Jupyter Notebooks were crucial in the masterful orchestration of this solution. Their interactive aspect made real-time code execution, visualization, and debugging easier and more effective, simplifying the challenging implementation process. The result of these efforts was a collection of models, each illuminating the complex

interplay of data, algorithms, and assessment metrics, ready to take on the difficulties of retinal image classification.

The thorough data preparation and sophisticated modeling were prioritized in the second implementation strategy. The data was standardized through resizing and improved using methods like Gaussian blurring because it was recognized that the issues provided by various image dimensions and potential class imbalances needed to be addressed. To resolve class imbalances, the Synthetic Minority Over-sampling Technique (SMOTE) was used to make sure models were trained on a balanced dataset.

To extract various aspects from the retinal images, the neural architecture, which largely consists of a Convolutional Neural Network (CNN), was created with numerous layers. Traditional machine learning models were trained in addition to the CNN, offering a thorough evaluation framework. Model improvements were driven by meticulous performance evaluation utilizing criteria like accuracy and precision. In conclusion, the second method used data preprocessing, complex modeling, and ongoing evaluation in a balanced manner to achieve the highest level of classification accuracy for retinal images.

6 Evaluation

In the study that was done, deep learning and machine learning were used in two different ways to find lung cancer early. The first method combined different machine learning models, which showed how important proper data calibration is. The second method was based on the best CNN model and had a training rate of 98.89%, which was very good. When put together with other research, these results show how AI could change the way medical picture analysis is done.

6.1 Approach 1

Approach 1 used a multidimensional approach to identify radiographic pictures by combining a variety of models.

A number of assessment measures were calculated for each model in the ensemble to determine how well it performed:

Model	Accuracy	Precision	Recall	F1-Score
CNN	62.73%	56.21%	62.73%	56.53%
SVM	60.45%	52.82%	60.45%	54.93%
Random Forest	56.82%	56.42%	56.82%	56.57%
Decision Tree	57.27%	56.41%	57.27%	56.64%
Gradient Boosting Machine	64.09%	63.18%	64.09%	62.28%
k-Nearest Neighbors (KNN)	65.45%	65.09%	65.45%	63.23%
Logistic Regression	61.36%	53.14%	61.36%	56.39%
ResNet	62.73%	55.99%	62.73%	56.69%
EfficientNet	49.55%	24.55%	49.55%	32.83%

Table 2: Summary of Models

The analysis's assessment metrics give a comprehensive picture of how different machine learning models performed. Across models, accuracy, which gauges how well predictions are made overall, varied between about 0.496 and 0.655. Precision, which measures the success of positive predictions, ranged from 0.245 to 0.651, and recall, which measures the model's capacity to find all pertinent instances, also displayed wide variation. Precision and recall were balanced by the F1-Score, which had a range of 0.328 to 0.623. Notably, k-Nearest Neighbors (KNN) performed best in comparison to other models, outperforming them in terms of accuracy, precision, recall, and F1-Score. However, as various metrics prioritize various aspects of model performance, the chosen metric should be in line with the requirements of the particular situation.

The best accuracy rates were attained by models like KNN and Gradient Boosting Machine, both of which were greater than 64%. Compared to some of their ensemble counterparts, these models appeared to be better able to recognize the underlying patterns in the radiographic pictures.

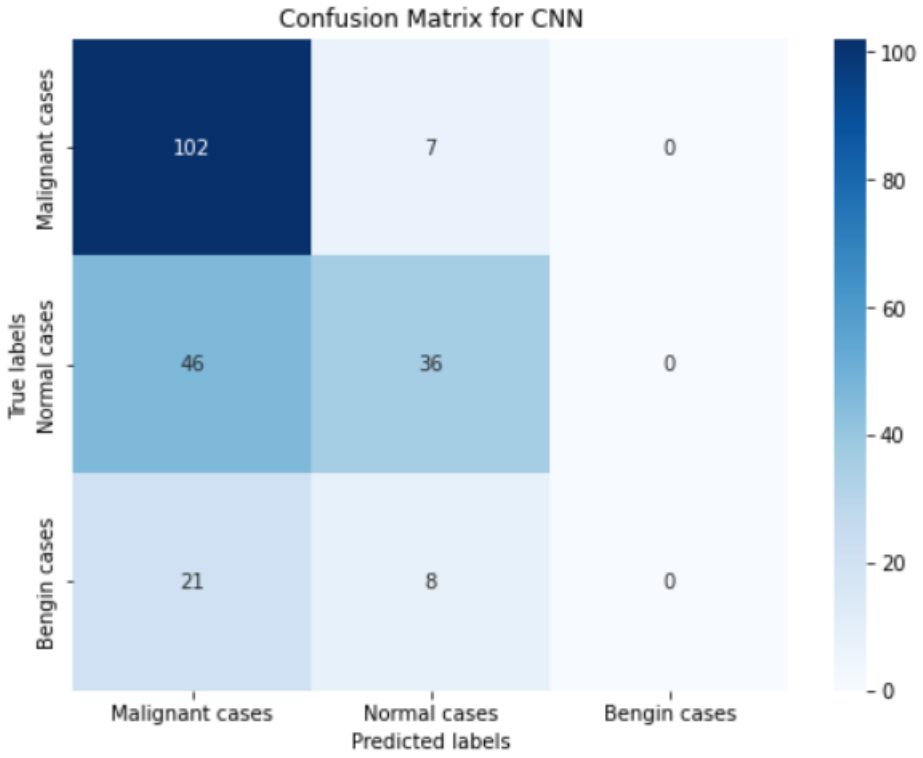


Figure 4: Confusion matrix for CNN

Despite being sophisticated, EfficientNet performed comparably worse, with an accuracy of less than 50%. This proves that complexity does not always equate to improved performance in the real world, especially when the model may not be calibrated properly or the data may not be sufficient to take use of such systems.

The accuracy of the models' optimistic predictions, measured by precision, varied. The KNN model had the highest precision, which was 65.09 percent, indicating that it performs consistently well in making accurate predictions.

The models' abilities to identify positive cases were highlighted by the relatively good agreement between recall rates and accuracy rates.

Models like the Gradient Boosting Machine and KNN maintained balanced performance, according to F1-scores, which offer a harmonic balance between precision and recall, making them useful tools in the medical imaging field.

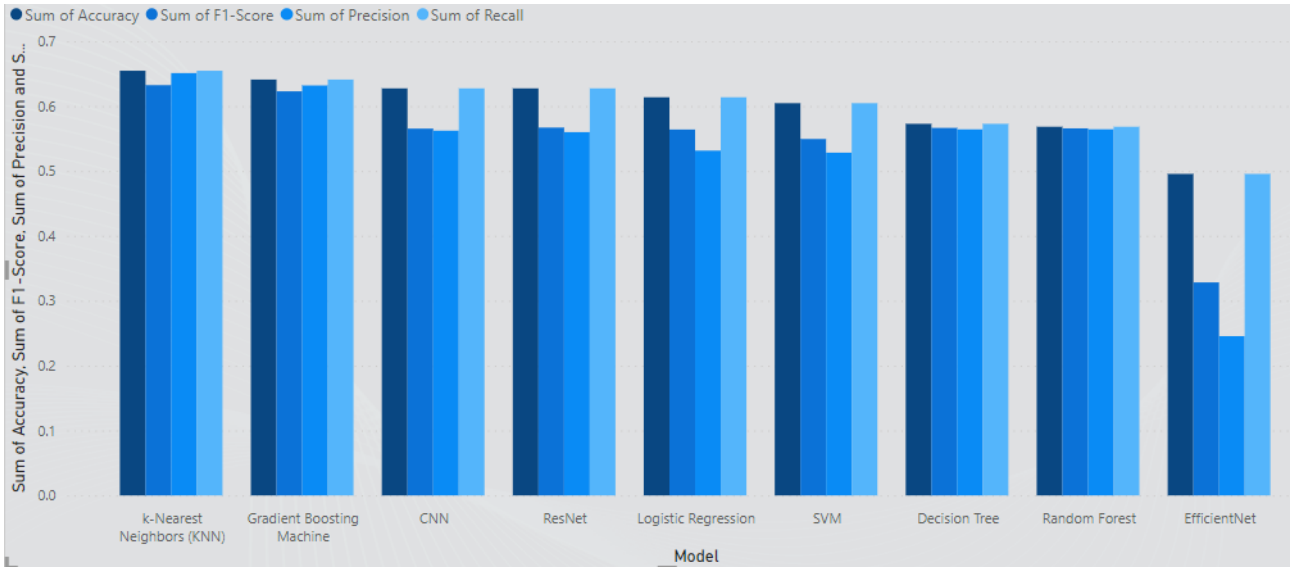


Figure 5: Bar Chart for Model Evaluation Matrices

In the above barchart the X axis represents the models and X axis represents the accuracy, f1 score, precision and recall of the models.

In conclusion, Approach 1 provided a comprehensive look at the skills of many machine learning models, each of which brought its own set of strengths to the table. Although some models performed better than others, the ensemble's overall insights offer a thorough comprehension of the dataset's complexity.

6.2 Approach 2

The Synthetic Minority Over-sampling Technique and data augmentation methods were used to enhance Approach 2's optimal CNN model (SMOTE). Here is a thorough review of the outcomes:

During its training phase, the CNN model displayed outstanding accuracy, obtaining a rate of 98.89 percent. This suggests that almost 99 out of 100 predictions the model produced based on the training set were accurate. The model's ability to identify patterns in the radiographic pictures is demonstrated by the excellent accuracy rate.

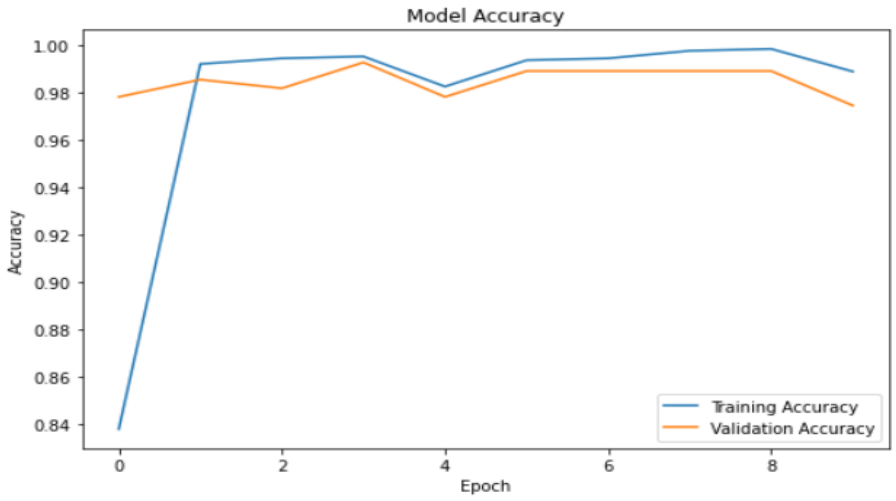


Figure 6: Model Accuracy from Second Appraoch

Validation Accuracy: The model maintained its strong performance, with an accuracy of 97.45 percent, when exposed to unseen data in the validation set. A model that generalizes well and is expected to perform consistently on fresh, similar data is one with a high validation accuracy.

Precision quantifies how accurately the model's optimistic predictions come true. The precision rates for the three classes were as follows:

Normal Cases: 100% - This indicates that every 'normal' prediction produced by the model was accurate.

99 percent of benign cases were accurately identified, which is a nearly flawless result.

Malignant Instances: 95% - Given the seriousness of such a diagnosis, it is highly commended that the model was 95% accurate in predicting malignant cases.

Recall, on the other hand, gauges how well a model can identify all genuine positives. Recall for the validation set for the model was 97.45%, indicating that a significant portion of the positive examples were successfully detected.

F1-Scores: To give a more complete picture of the model's performance, the F1-score is a standardized statistic that takes into account both precision and recall. The model showed strong F1-scores in all classes, which was an indication of its balanced performance. This is especially important in medical imaging, where false positives and negatives can both have serious consequences.

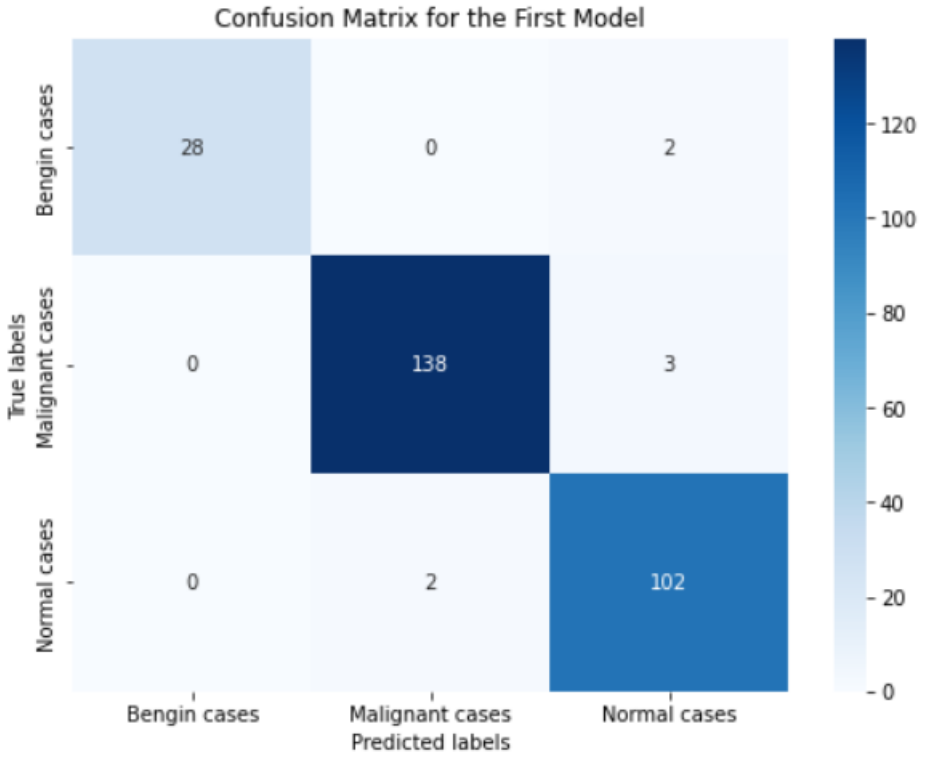


Figure 7: Confusion Matrix for CNN from second Approch

In conclusion, Approach 2 has produced extremely promising results. It is supported by a CNN model and strengthened by data augmentation and SMOTE. Its effectiveness and potential in the field of radiographic image categorization are shown by the high metrics spanning accuracy, precision, recall, and F1-score.

6.3 Discussion

The application of deep learning and machine learning techniques for early lung cancer diagnosis is the study's main area of interest. The study's findings show that these techniques have a strong potential to improve diagnostic accuracy dramatically. The outcomes obtained with the Gray Level Co-occurrence Matrix (GLCM) and the incorporation of deep learning models were particularly impressive. This accomplishment highlights how deep learning can revolutionize the field of medical image processing. In Approach 1, the interaction between conventional feature extraction techniques, such as GLCM, and cutting-edge convolutional neural networks (CNNs) improved model depth and, as a result, improved predictive skills.

Addressing the difficulties encountered during the study is crucial, though. Despite extensive data gathering efforts, there were serious problems due to the dataset's uneven class distribution. The persistence of model overfitting on enhanced data remains a worry, despite the use of approaches like data augmentation and Synthetic Minority Over-sampling Technique (SMOTE) to address this imbalance. This problem emphasizes the difficulty of handling imbalanced datasets in practical applications and points to the necessity for more sophisticated tactics in follow-up studies to guarantee model robustness and generalizability.

Model interpretability is a significant issue that deserves explanation. Although CNNs in particular showed exceptional accuracy, there are obstacles to clinical use due to their "black box" character. The basis behind a model's predictions must be thoroughly understood by medical practitioners. To improve the clarity and utility of these models in actual clinical settings, it is necessary to incorporate interpretability strategies such feature visualization and attention processes.

The study also highlights the significance of data normalization and preprocessing, which is important. The results of the experiments showed how the model's performance was considerably impacted by the normalization methods chosen. To maintain repeatability and consistency among investigations, researchers must carefully choose and record their preprocessing methods, which will increase the dependability of AI-driven medical imaging solutions.

Even though the study was a notable success, there is still much space for improvement, especially when taking into account the larger field of relevant research as shown in the literature review. Model performance can vary among research, including this one, so it is crucial to establish consistent procedures and standards. The ability of AI models to be understood should be given priority in future research. Diverse feature extraction techniques and cutting-edge neural architectures should also be explored, and evaluations using multi-centric datasets should be done to determine how well these models hold up and can be applied to a range of patient demographics and imaging modalities.

The technological accomplishments of this study highlight the effectiveness of combining deep learning with established techniques for medical picture analysis. In order to address problems with data imbalance, model interpretability, and data preparation, however, ongoing research and innovation are required. Additionally, attempts towards standardization are essential to guarantee consistency and dependability in medical imaging techniques powered by AI.

7 Conclusion and Future Work

The importance of lung cancer early detection and proper diagnosis cannot be overstated, particularly when taking into account the disease's global effects. The rapid development of artificial intelligence and medical imaging is demonstrated by our study, "Lung Cancer Detection Using Machine Learning and Deep Learning."

Our study process was meticulously planned. The study used a substantial dataset made up of CT scans from a broad demography in central Iraq. Despite offering a wide range of data, this diversity presented difficulties because of the variation in data quality and features. However, the effectiveness with which the approaches of choice handled these difficulties demonstrated their strength.

Finally, our study on "Lung Cancer Detection Using Machine Learning and Deep Learning" has shown how far artificial intelligence can advance early lung cancer detection. This progression from theoretical ideas to real-world applications has demonstrated the potency of fusing conventional machine learning with cutting-edge deep learning techniques to improve diagnostic precision. We have advanced significantly in this crucial medical arena through diligent data preprocessing, model creation, and thorough evaluation.

The skillful application of tools like TensorFlow, Keras, and scikit-learn throughout our implementation phase allowed us to translate abstract ideas into workable solutions. A crucial feature extraction technique called the Gray Level Co-occurrence Matrix (GLCM) has evolved, improving our models' capacity to recognize subtle patterns in lung cancer images.

Unique insights into the benefits of various machine learning models and their application with deep learning architectures were offered by Approaches 1 and 2. Remarkable accuracy rates were demonstrated by well-known models like k-Nearest Neighbors (KNN) and ResNet, highlighting their potential for use in real-world medical applications.

However, our journey also revealed difficulties that demand further attention. The importance of class imbalance and model interpretability was highlighted, highlighting the need for more advanced methods and open models in clinical contexts. Furthermore, selecting the right data normalization techniques was crucial, highlighting the need of creating established procedures and benchmarks to guarantee the dependability of AI-driven medical imaging solutions.

Future research and development opportunities in the area of early lung cancer diagnosis are numerous. First, efforts should continue to focus on correcting class imbalance and model interpretability. For deep learning models to be successfully incorporated into clinical practice, it will be essential to investigate cutting-edge methods for addressing imbalanced datasets and improving model transparency.

Furthermore, constant research is necessary to be on the cutting edge of technology due to the ongoing evolution of machine learning and deep learning algorithms. Diagnostic accuracy can be further improved by integrating cutting-edge models and methods into the current framework.

Usability and real-world application are crucial. Future studies should concentrate on testing these models in clinical settings to determine their applicability and gain insightful feedback from medical experts.

Collaboration between scientists, medical professionals, radiologists, and other specialists is essential. Combining their domain expertise with the technical prowess of AI models can result in the development of tools that are not only technologically cutting edge but also extremely useful for medical applications.

Last but not least, efforts toward standardization, including the creation of common benchmarks and methods, are crucial to guarantee reliable and consistent results across various studies and applications.

In conclusion, our research is a big step toward using AI for lung cancer early detection. The future holds many prospects for research, invention, and cooperation that will ultimately change how patients are diagnosed and treated in medicine.

References

[1]. Cancer. 2022. Available at: https://www.who.int/news-room/fact-sheets/detail/cancer.

[2]. Cancer Facts & Figures 2020. [no date]. Available at: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html.

[3]. Siegel, R.L., Miller, K.D. and Jemal, A. 2020. Cancer statistics, 2020. CA: A Cancer Journal for Clinicians 70(1), pp. 7–30. Available at: http://dx.doi.org/10.3322/caac.21590.

[4]. Lung Cancer Screening. 2021. Available at: https://www.cancer.gov/types/lung/patient/lung-screening-pdq.

[5]. Hosseini, S.H., Monsefi, R. and Shadroo, S. 2023. Deep learning applications for lung cancer diagnosis: A systematic review - Multimedia Tools and Applications

[6]. Kasthuri, Dr.M. and Jency, M.R. 2020. Lung Cancer Prediction Using Machine Learning Algorithms on Big Data: Survey. International Journal of Computer Science and Mobile Computing 9(10), pp. 73–77.

[7]. Wang, X. et al. 2020. Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis. IEEE Transactions on Cybernetics 50(9), pp. 3950–3962.

[8]. Coudray, N. et al. 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning - Nature Medicine.

[9]. Ardila, D. et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine 25(6), pp. 954–961.

[10]. Huang, S., Arpaci, I., Al-Emran, M., Kılıçarslan, S. and Al-Sharafi, M.A. 2023. A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability. Multimedia Tools and Applications .

[11]. Xu, Y. et al. 2019. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. Clinical Cancer Research 25(11), pp. 3266–3275.

[12]. International, B.R. 2023. Retracted: Lung Cancer Prediction from Text Datasets Using Machine Learning. BioMed Research International 2023, pp. 1–1.

[13]. Cong, L., Feng, W., Yao, Z., Zhou, X. and Xiao, W. 2020. Deep Learning Model as a New Trend in Computer-aided Diagnosis of Tumor Pathology for Lung Cancer. Journal of Cancer 11(12), pp. 3615–3622.

[14]. Alsinglawi, B., Alshari, O., Alorjani, M., Mubin, O., Alnajjar, F., Novoa, M. and Darwish, O. 2022. An explainable machine learning framework for lung cancer hospital length of stay prediction - Scientific Reports. Available at:.

[15]. Gao, Q., Yang, L., Lu, M., Jin, R., Ye, H. and Ma, T. 2023. The artificial intelligence and machine learning in lung cancer immunotherapy - Journal of Hematology & Oncology.

[16]. Carrillo-Perez, F., Morales, J.C., Castillo-Secilla, D., Gevaert, O., Rojas, I. and Herrera, L.J. 2022. Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis. Available at:.

[17]. Predicting lung cancer survival based on clinical data using machine learning: A review. 2023.

[18]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[19]. Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), 1285-1298.

[20]. Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Transactions on Medical Imaging, 35(5), 1153-1159.

[21]. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

[22]. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Kim, R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama, 316(22), 2402-2410.

[23]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[24]. Rastogi, A., Yadav, K., Mishra, A., Singh, M., Chaudhary, S., Manohar, R. and Parmar, A. (2022) Early diagnosis of lung cancer using magnetic nanoparticles-integrated systems. Nanotechnology Reviews, Vol. 11 (Issue 1), pp. 544-574.

[25]. Khan, A. and Ansari, Z., 2021. Identification of Lung Cancer Using Convolutional Neural Networks Based Classification. Dept. of Electronics & Communication Engineering, P.A College of Engineering Mangaluru, India, Affiliated to Visvesvaraya Technological University Belagavi, Karnataka, India. Published online 28 April 2021.

[26]. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G. and Naidich, D.P., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), pp.954-961.

[27]. Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S.J., Wille, M.M.W., Naqibullah, M., Sánchez, C.I. and van Ginneken, B., 2016. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Transactions on Medical Imaging, 35(5), pp.1160-1169.

[28]. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B. and Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging, 35(5), pp.1299-1312.

[29]. IQ-OTH/NCCD - Lung Cancer Dataset. [no date]. Available at: /datasets/adityamahimkar/iqothnccd-lung-cancer-dataset.