Deep Learning Based Lung Cancer Detection and Classification

Final Year Project

Session 2021-2025

A project submitted in partial fulfillment of the degree of

BS in Computer Science

Department of Computer Science

Faculty of Computer Science & Information Technology

The Superior University, Lahore

FALL 2024

Type (Nature of project)	[ ] **D**evelopment [ ] **R**esearch [✓] **R&D**
Area of specialization	Deep Learning in Medical Imaging
FYP ID	FYP-BCSM-S24-009

Project Group Members

Sr.#	**Reg. #**	**Student Name**	**Email ID**	**Signature**
(i)	BCSM-S21-011	Waleed Ashraf (Leader)	bcsm-s21-011@superior.edu.pk	
(ii)	BCSM-S21-024	Momal Mushtaq	bcsm-s21-024@superior.edu.pk	
(iii)	BCSM-F20-417	Mahanoor Farooq	bcsm-f20-417@superior.edu.pk	

*The candidates confirm that the work submitted is their own and appropriate credit has been given where reference has been made to work of others

Plagiarism Free Certificate

This is to certify that, I **Muhammed Waleed Ashraf Qureshi** Son of **Muhammed Ashraf Qureshi**, group leader of FYP under registration no **BCSM-S21-011** at Computer Science Department, The Superior University, Lahore. I declare that my FYP report is checked by my supervisor.

Date: 13-12-2024. Name of Group Leader: Muhammed Waleed Ashraf Qureshi Signature:

Name of Supervisor: Dr. Jawad Ahmad
Designation: Assistant Professor

Signature: ________________

Co-Supervisor: Mr. Zeeshan Mubeen
Designation: Senior Lecturer

Signature: __________________

HoD: Dr. Muhammad Azam

Signature: ________________

Project Report

Deep Learning Based Lung Cancer Detection & Classification

Change Record

Author(s)	Version	Date	Notes	Supervisor's Signature
Waleed	1.0	15 Nov 2023	<Designed the initial architecture for the Hybrid CNN model and selected IQ-OTH/NCCD dataset for research>	
Mahanoor, Momal	1.1	20 Dec 2023	<Drafted the literature review and outlined key challenges in lung cancer diagnosis>	
Waleed	1.2	10 Jan 2024	<Developed and optimized the hybrid model using VGG16 and AlexNet for feature extraction>	
Waleed	1.3	25 Feb 2024	<Implemented Flask API for predictions and integrated it with the mobile application>	
Waleed	1.4	10 Mar 2024	<Trained the hybrid model with KNN and SVC classifiers, achieving 99.55% and 98.65% accuracy, respectively>	
Waleed	1.5	15 Apr 2024	<Revised and expanded methodology section>	
Waleed	1.6	30 May 2024	<Deployed the final model and application on AWS ECS for user accessibility>	
Mahanoor, Momal	1.7	15 Jun 2024	<Prepared documentation and added detailed comparison with other research studies>	
Waleed	1.8	20 Sep 2024	<Conducted final accuracy testing and improved preprocessing pipeline for enhanced results>	
Mahanoor, Momal	1.9	10 Nov 2024	<Finalized research paper for submission, including conclusions and future scope>	

APPROVAL

PROJECT SUPERVISOR

Comments: ____

____

Name:______________________________

Date:________________________________ Signature:___________________________

PROJECT MANAGER

Comments: ____

____

Date:________________________________ Signature:___________________________

HEAD OF THE DEPARTMENT

Comments: ____

____

Date:________________________________ Signature:___________________________

Dedication

To all whose lives have been touched by lung cancer patients, families and providers, your courage moves us. This work is dedicated to you, and we further hope that continued advancement of AI-driven near-term diagnosis will increase your outcomes, save more lives and build a better future for all of those whom this disease touches.

Acknowledgements

I would like to express my deepest gratitude to everyone who has supported me throughout the development and completion of this Final Year Project. First and foremost, I thank my project supervisor, Sir Jawad Ahmad, for his invaluable guidance, insightful feedback, and continuous support. His expertise and encouragement were crucial in shaping this project. I am also immensely grateful to my professors and lecturers at Superior University, who provided me with the knowledge and tools necessary for this research, with special thanks to the Department of Computer Science for their resources and support. My heartfelt thanks extend to my family, especially my parents, for their unwavering support, patience, and encouragement throughout my academic journey. Their belief in me has been a constant source of motivation. To my friends and classmates, thank you for your friendship, support, and the many discussions and idea-sharing sessions that greatly helped this project. Lastly, I thank all the participants who took part in my research study. Without their cooperation and willingness to share their experiences, this project would not have been possible.

Thank you all for your invaluable support.

Executive Summary

A new hybrid approach based on VGG16, AlexNet, SVC, and KNN model was developed for lung cancer prediction and classification from CT images. Individual modelling results showed that the model, based on the IQ-OTH/NCCD dataset (1,190 images), achieved accuracies of 99.55% (KNN) and 98.65% (SVC) outperforming the separate modelling. The unique methodology synergizes deep architectures with dimensionality reduction that may facilitate improved feature extraction and early lung cancer diagnostics in clinical practice.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Chapter 1: Introduction

The most fatal diseases in the world, lung cancer greatly influences cancer-related death. With around 2.5 million new cases expected worldwide in 2022, lung cancer is among the most often diagnosed malignancy according to the World Health Organisation (WHO), accounting for over 1.8 million fatalities [1]. The lack of overt symptoms in the early stages adds to the difficulties in early stage lung cancer diagnosis. Most people limit treatment options and lower their chances of survival as most do not show obvious symptoms until the disease has progressed. Consequently, early identification is rather important for raising survival rates since it helps to administer more efficient therapies and produce better therapeutic results.

Improvements in medical imaging and artificial intelligence (AI) during the past few decades have greatly enhanced the speed and accuracy of diagnosis procedures. By allowing computers to automatically examine and interpret imaging data, especially deep learning models, machine learning approaches have transformed medical imaging by often surpassing human capacity in spotting early illness indications. Particularly in identifying and categorizing lung cancer from imaging data like CT scans, convolutional neural networks (CNNs) have become more effective instruments for medical image processing.

1.1. Symptoms of Lung Cancer

Because of its mild and vague symptoms, lung cancer sometimes remains undetectable in its early phases. The cancer may have progressed already by the time the symptoms start to show. Timely diagnosis depends on early identification of important symptoms, which helps diagnostic models and medical practitioners to respond quickly. The most often occurring lung cancer symptoms consist in:

- **Persistent Coughing**: A chronic or worsening cough that may initially be dry and later progress to coughing up mucus or blood.
- **Chest Pain**: Pain that intensifies with coughing, deep breathing, or laughing may indicate lung tissue damage or metastasis.

- **Shortness of Breath**: As tumors grow, they may obstruct airflow, leading to difficulty in breathing, even with mild physical exertion.
- **Unexplained Weight Loss and Fatigue**: These nonspecific symptoms are common in many cancers, including lung cancer.
- **Coughing up Blood (Hemoptysis)**: This is an alarming symptom and usually indicates advanced disease, necessitating immediate medical attention.

Recognizing these symptoms and understanding their potential progression is essential for effective early diagnosis, whether by healthcare providers or AI-based diagnostic tools.

1.2. Types of Lungs Cancer

Lung cancer is categorized into two main types, each with distinct characteristics, treatment options, and prognoses. The two primary classifications of lung cancer are:

1.2.1. Non-Small Cell Lung Cancer (NSCLC)

About 85% of all cases of lung cancer are NSCLC, which is further split into three main subtypes:

- The most prevalent subtype, adenocarcinoma, usually starts in the peripheral lung tissues and is frequently observed in non-smokers.
- Squamous Cell Carcinoma: This subtype typically arises in the lungs' central airways and is closely associated with smoking.
- Large Cell Carcinoma: A less prevalent variety that spreads and grows quickly.

1.2.2. Small Cell Lung Cancer (SCLC)

SCLC is a far more aggressive type of lung cancer, although being less frequent. Early in its course, it has a propensity to spread swiftly to other regions of the body. Smoking is closely linked to SCLC, which necessitates specific treatment regimens that frequently include radiation and chemotherapy.

One of the main problems for AI-based diagnostic models is to distinguish between various categories, which is essential for choosing the right treatments.

1.3. Global Impact and Risk Factors

With 2.5 million new cases yearly and over 1.8 million deaths annually, lung cancer continues to be the primary cause of cancer-related mortality globally [2]. Tobacco usage is clearly connected to lung cancer incidence; smokers have a much higher risk. Non- smokers do, however, also get lung cancer, which emphasizes the need of including environmental elements and genetic predispositions into lung cancer diagnosis. Among them are:

- **Tobacco Smoke:** Still the principal cause of lung cancer, tobacco smoke makes over 85% of the cases. Secondhand smoking also greatly raises danger.
- **Environmental Pollutants:** Particularly in metropolitan areas, long-term exposure to air pollution especially particulate matter and industrial emissions has been shown to increase the risk of lung cancer.
- **Occupational Hazards:** Include carcinogens like radon gas and asbestos, which raises the risk of lung cancer in those environments.
- **Genetic Mutations:** These risk factors highlight the importance of early and reliable diagnostic instruments especially in populations exposed to high-risk situations or genetic predispositions.

1.4. Challenges in Lung Cancer Detection

The diagnosis of lung cancer is hampered by several reasons, including the variety of tumor presentation in medical imaging, especially in CT scans. Some of the challenges include:

- **Variability in Image Quality:** The quality of CT scans may fluctuate due to factors such as machine configurations and patient anatomy, complicating precise interpretation.
- **Distinguishing Between Benign and Malignant Growths:** Differentiating between benign and malignant nodules can be difficult, particularly in the early stages when tumors are small and may overlap with adjacent structures.
- **Restricted Access to Diagnostic Facilities:** In under-resourced areas, the availability of advanced diagnostic equipment and specialized radiologists is constrained, resulting in postponed diagnoses.

- **Reliance on Proficient Radiologists:** The interpretation of radiological images is significantly reliant on the expertise of the radiologist, and the procedure can be laborious and susceptible to human mistake.

These issues require the creation of automated diagnostic solutions, especially those driven by AI, to address these constraints and enhance diagnostic precision.

1.5. The Role of AI in Lung Cancer Diagnosis

Artificial intelligence—more especially, deep learning models—has demonstrated significant promise in recent years for improving the precision, speed, and scalability of lung cancer diagnosis. Convolutional Neural Networks (CNNs), in particular, are machine learning algorithms that automatically learn hierarchical features from medical images. This allows the model to recognize intricate patterns that are frequently invisible to the human eye [3]. Rapid developments in AI and deep learning have created new opportunities for more dependable and effective radiology diagnostic procedures, especially in the diagnosis of lung cancer.

1.6. Important AI Models for Imaging in Medicine

Medical image analysis has effectively used a number of deep learning architectures, especially when it comes to the identification of lung cancer:

- **AlexNet:** One of the first CNNs, AlexNet achieved impressive results in large-scale image classification tasks by introducing innovations like rectified linear units (ReLU) and dropout for regularization [5].
- **VGG16:** Well-known for utilizing tiny convolutional filters in extremely deep networks, VGG16 greatly increased model accuracy and advanced image recognition methods [6].
- **DenseNet:** DenseNet employs dense connectivity across layers to improve feature propagation and lower the amount of parameters, which enhances performance and lessens overfitting [7]. This family of algorithms includes DenseNet121, DenseNet169, and DenseNet201.
- **MobileNetV2:** Because of its computational efficiency and lightweight design, this architecture works well in contexts with limited resources, such as mobile devices [8].

- **ResNet152V2:** Well-known for its residual learning methodology, ResNet152V2 makes training extremely deep networks easier by resolving the vanishing gradient issue [9].
- **EfficientNetV2:** By balancing network depth, width, and resolution, this model optimizes accuracy and efficiency, offering a solution that lowers computing costs without sacrificing performance [10].
- **InceptionV3:** InceptionV3 improves its capacity to identify intricate patterns in medical images by utilizing inception modules that record multi-level features with different kernel sizes [11].

1.7. Research Objectives

The main objective of this study is to assess and contrast the efficiency of several cutting-edge CNN models in identifying CT scan images and classifying lung cancer. Finding the best deep learning architecture for detecting lung cancer in medical imaging involves comparing several models based on important performance measures like accuracy, sensitivity, and specificity. By enhancing the diagnostic process and paving the way for the creation of automated systems, this research hopes to enhance patient outcomes by helping doctors make more accurate diagnoses.

1.8. Problem Statement

The intricate nature of lung cancer diagnosis, along with restricted access to proficient radiologists and new diagnostic tools, poses a considerable obstacle to prompt and precise identification. Contemporary diagnostic techniques, frequently dependent on manual analysis of CT scans, lack scalability and are susceptible to inaccuracies. An urgent necessity exists for AI-driven solutions to automate the diagnostic process, enhance accuracy, and increase accessibility in lung cancer diagnosis.

1.9. Why We Need an AI-Based Model

AI models analyze medical images at speeds and accuracies unattainable by humans. They ensure consistent results and reduce diagnostic delays, especially in regions with limited medical expertise.

1.10. Implementation of Methodology

This project utilizes a hybrid approach, combining AlexNet and VGG16 for feature extraction, Lasso regularization for dimensionality reduction, and SVC/KNN for classification. The IQOTH/NCCD dataset is the cornerstone for training and validating the system, ensuring optimized performance in diverse scenarios.

Chapter 2

Literature Review

Chapter 2: Literature Review

Qadir et al. [12] created the hybrid Lung Cancer Stage Classifier and Diagnosis Model (Hybrid-LCSCDM) by use of CT scan data. Through three categories normal, benign, and malignant, this creative method streamlines the diagnosing procedure. First, the model employs the pre-trained VGG-16 model to detect important characteristics in lung CT scans suggestive of cancer, therefore performing a two-stage procedure. After that, the XGBoost machine learning method divides the scans into suitable groups and labels the produced features. Training and evaluation came from the 1,190 pictures from the three-class IQ-OTH/NCCD Lung Cancer dataset. The results shown the lifetime of the model with an overall accuracy of 98.54% and a classification precision of 98.63%. The hybrid-LCSCDM model beat earlier models according to accuracy, recall, precision, and F1-score tests, so suggesting its efficiency and dependability in CT scan interpretation for lung cancer detection.

Parveen et al. [13] established a machine learning architecture based on CT scan data aiming at producing an accurate and dependable model for lung cancer diagnosis. This work addressed the problem of class imbalance in medical datasets, a typical occurrence causing prediction mistakes. Using the IQ-OTH/NCCD dataset, the study first used Watershed approaches for feature extraction and image segmentation then Scale Invariant Feature Transform (SIFT) for image segmentation. Data augmentation methods were used to ensure a more fair representation for training, hence reducing dataset imbalance. Then, using a convolutional neural network (CNN) framework, categorization attained a 97% accuracy. Emphasizing its ability for early and accurate identification, the suggested approach also concentrated on spotting nodules within malignant lung pictures. Using CT scan pictures, this study emphasizes the efficiency of including CNN classification, data augmentation, and advanced feature extraction approaches for strong lung cancer diagnosis.

Using over 1,100 CT images from two Iraqi hospitals, Kareem at el. [14] studies the IQ-OTH/NCCD lung cancer dataset. Proposed is a computer-based lung cancer detection method using segmentation, picture enhancement, and feature extraction. Evaluated using several kernels and extraction techniques, a Support Vector Machine (SVM) classifier sorts scans into normal, benign, or malignant. With a maximal accuracy of 89.89%, the technique shows the value of this one for efficient lung cancer diagnosis and categorization.

Deepa et al. [15] described the LCC-Deep-ShrimpNet model for CT image-based lung cancer classification using an IQ-OTH/NCCD dataset. Using Bayesian fuzzy clustering and a kernel correlation technique, the method pre-processed by lung nodule area extraction. Deep-ShrimpNet classifier let these locations be classified as either normal, benign, or cancerous. By surpassing present approaches in performance, calculation time, and error rates, the proposed technique greatly enhanced lung cancer detection accuracy and efficiency.

LCCNet architecture was presented by Khaliq et al. [16] utilizing transfer learning with a pre-trained DenseNet-121 CNN model for effective lung cancer classification using the IQ-OTH/NCCD dataset. The model tackled usual difficulties such class imbalance, poor learning rate, overfitting of data, and vanishing gradient by means of data augmentation techniques. Several performance criteria accuracies, F1-score, precision, and recall after CT scan analysis were used to evaluate the system. With an incredible 99% accuracy, the LCCNet model exceeded previous methods in precisely spotting lung cancer.

For the detection of malignant nodules in lung CT scans from the IQ-OTH/NCCD dataset, AL-Huseiny et al. [17] suggested a deep learning method using deep neural networks (DNN), especially transfer learning with GoogLeNet. The technique consisted on centering and normalizing the pictures after a pre-processing stage separating areas of interest (ROIs) by eliminating surrounding tissues and artifacts. To fit the medical data, transfer learning was used by fine-tuning the DNN's last levels while keeping the deeper layers fixed. Showing its potential for early lung cancer diagnosis, the method exceeded previous benchmark methods applied on the same dataset by reaching an accuracy of 94.38%.

Combining ConvNeXt with FocalNet inside a vision transformer architecture, Gulsoy et al. [18] presented FocalNeXt, an enhanced architecture intended for automated lung cancer diagnosis from CT images. Aiming to improve diagnosis accuracy in lung cancer detection, the FocalNeXt model uses the strong attention mechanism of FocalNet and the feature extraction powers of ConvNeXt. FocalNeXt outperformed state-of- the-art models with an amazing accuracy of 99.81% and a sensitivity of 99.78% evaluating the IQ-OTH/NCCD dataset. The ablation research underlined the resilience of the model across many configurations, thus establishing FocalNeXt as a leading method for lung cancer diagnosis and so advancing medical imaging and tailored healthcare.

Al-Yasriy et al. [19] presented AlexNet architecture-based computer-aided system for lung cancer diagnosis using convolutional neural networks (CNN). Using the IQ-OTH/NCCD dataset acquired from Iraqi hospitals, the model was employed to classify CT scans of the lungs as either benign, normal, or malignant. It was clear that the suggested method was effective, as it achieved a sensitivity of 95.614% and a specificity of 95%. This demonstrates how effectively the approach supports early diagnosis, gives a consistent mechanism to help discriminate between several lung illnesses, and improves survival rates through early intervention.

Using machine learning, Gowda M A et al. [20] put forth a computer vision-based technique for lung tumor diagnosis. The approach uses Random Region Segmentation (RSS) for segmentation and SIFT and GLCM for feature extraction. A Triple Support Vector Machine (SVM) categorizes the data as normal, benign, or malignant. After 300 epochs of training on the IQ-OTH/NCCD dataset, the model had 96.5% accuracy with 200 clusters. It got better yet more as the cluster size grew to 500. This method increases classification accuracy, hence it might be a useful instrument for lung cancer diagnosis.

In order to enhance lung cancer detection, Jassim et al. [21] put forth a hybrid framework merging deep learning with multi-criteria decision-making methods. This paper introduced a CNN using three class balancing techniques for feature extraction and classification. The objective was to solve issues about medical dataset class imbalance. We searched and selected the best classification technique using VIKOR and the fuzzy-weighted zero-inconsistency approach. With a 99.27% accuracy rate, the model performed really brilliantly. Any imbalanced dataset across classes can have the ideal classification method found using this approach.

Anand et al. [22] study how deep learning-based models applied to CT scans might enhance early lung cancer detection. The study classifies lung cancer using a handcrafted convolutional neural network (CNN) model in conjunction with transfer learning utilizing the VGG-16 and Inception V3 architectures. These models are enhanced by hyperparameter tweaking, which ensures higher classification accuracy. The study highlights the importance of early identification in increasing survival rates and analyzes model performance to determine the most effective technique. This study uses cutting-edge deep learning approaches to improve lung cancer diagnosis.

Using deep ensemble convolutional neural networks (DECNN), Dass et al. [23] investigate a method for spotting and identifying small objects such as cancers or nodules in medical photographs. The major concentration of the work is semantic segmentation of CT-PET images from organs like the liver and lung. Before they are given to the DECNN for classification, kaze feature extraction and morphological segmentation are performed to assess and extract notable properties. With a 99.8% success rate on stationary data, the proposed approach seems to be a reliable and accurate way for medical imaging microscopic item identification.

Hossain et al. [24] provide a hybrid LeNet-LSTM model for lung cancer classification. Combining LSTM networks with the LeNet architecture recognized for its convolutional and max-pooling layers addresses the vanishing gradient problem and provides sequential data processing. Exceeding existing methods, the model has a 99.27% accuracy. This approach has significant promise to assist doctors in differentiating benign from malignant and normal tumors, therefore improving patient outcomes by helping them to diagnose accurate lung cancer.

Reddy et al. [25] investigate lung cancer using cutting-edge Convolutional Neural Networks (CNNs), including Xception and ResNet50 among previously trained models. The work employs approaches such as SMote and class weighting to address the issue of dataset imbalance and significantly improve model performance. The model achieved 95% accuracy by classifying CT images as normal, benign, or malignant. The study compares custom-built and pre-trained algorithms for identifying malignant nodules. It emphasizes even more how CNN-based techniques might be able to overcome variations in lung lesion size and form.

Using Convolutional Neural Networks (CNNs) tuned using the Sine Cosine Algorithm (SCA), Pathan et al. [26] study offers an automated lung cancer diagnosis tool. The model correctly labels lung CT images into normal, benign, and malignant groups 99% times over. Strong generalization on unknown datasets is shown by using a unique preprocessing technique and tweaking CNN hyperparameters with SCA, therefore providing a potential, non-invasive substitute for conventional biopsies for clinical usage.

The work of Gupta et al. [27] presents UDCT, a new framework for CT image-based lung cancer detection and categorization. Using a modified U-Net architecture with differentiable architecture search (DARTS) for effective classification, the UDCT framework the system also preprocesses images using Multilevel Otsu thresholding and lung nodule segmentation. Experiments on the LIDC-IDRI and IQ-OTH/NCCD datasets demonstrate the framework's high performance, with accuracies of 95.01% and 96.82%, respectively. Ablation research enhances the effectiveness and utility of the framework for detecting lung cancer.

Using the IQ-OTHNCCD CT scan dataset, Mohamed et al. [28] segment lung cancer. The method uses a Convolutional Neural Network (CNN) and comprises extensive preprocessing (resizing, normalizing, Gaussian blurring). SMote is used to solve class imbalance thus improving the capacity of the model to categorize underrepresented groups. With accuracy of 99.64%, the model performs well; precision, recall, and F1-score all surpasses 98%. These findings demonstrate how effectively machine learning may deliver reliable, non-invasive diagnostic tools, allowing for early detection and better patient outcomes.

Kumaran et al. [29] use pre-trained models (VGG16, ResNet50, and InceptionV3) to examine a combined deep learning framework for lung cancer detection. The model was trained on the IQ-OTH/NCCD dataset by altering deeper layers and applying SMote and Gaussian Blur to achieve class balance. With 98.18% accuracy, our method improved medical imaging diagnosis precision. A study by Mohamed et al. [30] shows how hard it is to get an accurate diagnosis of lung cancer with CT scans because the images are so complicated that they make it hard to classify them. Traditional convolutional neural network (CNN) models see heavy usage despite their shortcomings in design selection and bias and weight adjustment accuracy. The researchers came up with a hybrid metaheuristic-CNN approach that integrates a customized CNN with the Ebola Optimization Search Algorithm (EOSA) to tackle these problems. On the IQ-OTH/NCCD lung cancer dataset, our strategy outperformed previous strategies, with a success rate of 93.21%. The EOSA-CNN model shows promise as a technique for detecting lung cancer because to its excellent sensitivity and accuracy across many case types.

Majumder et al. [31] study shows how important it is to find lung cancer early in order to improve patient results and lower death rates. Traditional methods of diagnosis aren't always accurate, which is why computer-aided systems are being used more and more. The study proposed a group model named MENet that combines the results of the Xception, InceptionResNetV2, and MobileNetV2 deep learning models. It accomplishes this by employing a fuzzy ranking algorithm based on the Mitscherlich function. This novel approach was tested on two publicly available CT scan datasets, IQ-OTH/NCCD and LIDC-IDRI, and it performed admirably 99.54% of the time. The results reveal that MENet outperforms other models and demonstrate how it may increase diagnostic accuracy by integrating various architectures.

Table 1 Comparative Analysis of the proposed model with other existing studies.

Study	**Approach**	**Year**	**Dataset**	**Accuracy**
Qadir et al. [12]	Hybrid Lung Cancer Stage Classifier and Diagnosis Model (Hybrid-LCSCDM) using pre-trained VGG-16 and XGBoost	2023	IQ-OTH/NCCD Lung Cancer dataset	98.54%
Parveen et al. [13]	Machine learning architecture using Watershed, SIFT, data augmentation, CNN	2023	IQ-OTH/NCCD Lung Cancer dataset	97.00%
Kareem at el. [14]	Lung cancer detection with SVM classifier using segmentation, image enhancement, and feature extraction	2020	IQ-OTH/NCCD Lung Cancer dataset	89.89%

Deepa et al. [15]	LCC-Deep-ShrimpNet with Bayesian fuzzy clustering and kernel correlation	2023	IQ-OTH/NCCD Lung Cancer dataset	97.58%
Khaliq et al. [16]	LCCNet with transfer learning using DenseNet-121, addressing class imbalance and overfitting	2023	IQ-OTH/NCCD Lung Cancer dataset	99.00%
AL-Huseiny et al. [17]	DNN with transfer learning using GoogLeNet, focusing on ROIs and pre-processing	2021	IQ-OTH/NCCD Lung Cancer dataset	94.38%
Al-Yasriy et al. [19]	AlexNet-based system with CNN, focused on early lung cancer diagnosis	2020	IQ-OTH/NCCD	93.548%
Gowda M A et al. [20]	Computer vision-based technique using RSS, SIFT, GLCM, and Triple SVM for classification	2022	IQ-OTH/NCCD Lung Cancer dataset	96.5%
Jassim et al.[21]	Hybrid framework with CNN, class balancing, VIKOR, and fuzzy-weighted zero-inconsistency	2022	IQ-OTH/NCCD Lung Cancer dataset	99.27%
Dass et al. [23]	Deep ensemble CNN (DECNN) using Kaze feature extraction and morphological segmentation	2022	CT-PET images	99.80%
Reddy et al.[25]	CNNs with Xception and ResNet50, using SMote and class weighting	2022	IQ-OTH/NCCD Lung Cancer dataset	95.00%
Pathan et al. [26]	CNN with Sine Cosine Algorithm for hyperparameter tuning	2024	IQ-OTH/NCCD Lung Cancer dataset	99.00%
Gupta et al. [27]	UDCT framework with modified U-Net, DARTS, and Multilevel Otsu thresholding	2024	LIDC-IDRI & IQ-OTH/NCCD datasets	96.82% (IQ-OTH/NCCD), 95.01% (LIDC-IDRI)

Musthafa et al. [28]	CNN with SMote for class imbalance and extensive preprocessing	2024	IQ-OTH/NCCD Lung Cancer dataset	99.64%
Kumaran et al. [29]	Combined deep learning framework with pre-trained VGG16, ResNet50, InceptionV3	2024	IQ-OTH/NCCD Lung Cancer dataset	98.18%,
Mohamed et al. [30]	EOSA-CNN with Ebola Optimization Search Algorithm for weight tuning	2023	IQ-OTH/NCCD Lung Cancer dataset	93.21%

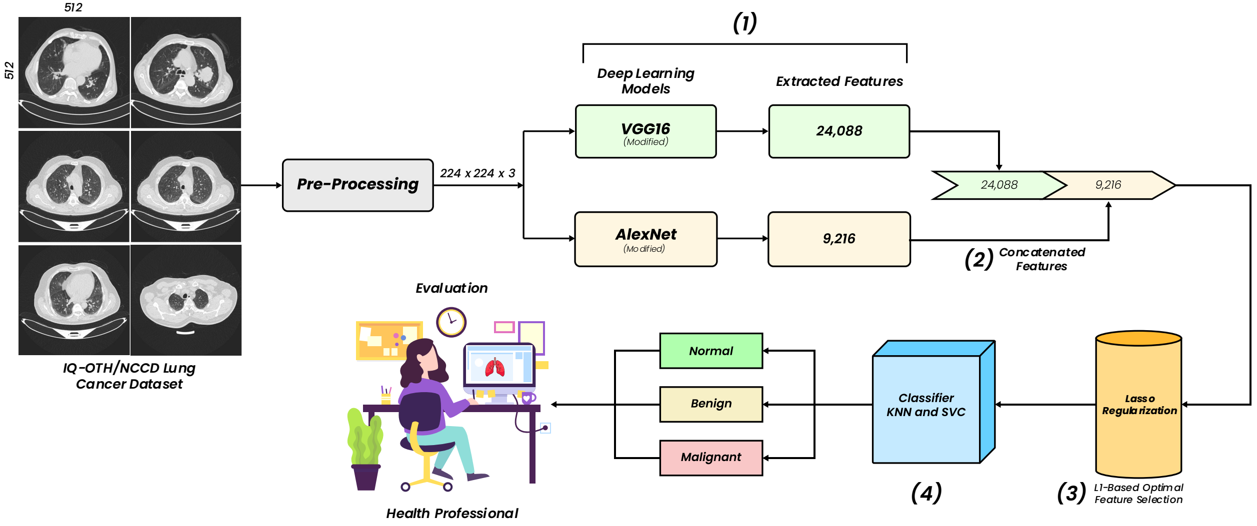
Chapter 3

Methodology

Chapter 3: Methodology

This work generates and tests a new hybrid deep learning framework for early lung cancer diagnosis via a mixed-methods approach. The proposed approach combines feature selection via Lasso regularization and subsequent classification using Support Vector Classification (SVC) and K-Nearest Neighbors (KNN) algorithms with the feature extracting capabilities of two well-known Convolutional Neural Networks (CNNs) AlexNet and VGG16.

Figure 1 Proposed hybrid framework



3.1. Dataset

This paper makes use of the lung cancer dataset from Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) [32], gathered during three months in autumn of 2019. Based on 110 cases, this large dataset comprises 1,190 CT scan slices of both healthy volunteers and lung cancer patients at many phases (Al-Shabi et al., 2019). All standardized to 512×512-pixel size, the photos fall into three categories: 415 normal, 120 benign, and 517 malignant. Included in the data collecting system was a 1 mm slice thickness SOMATOM Siemens scanner at 120 kV. Window center values spanned from 50 to 600 while window widths ranged from 350 to 1200 HU. Every scan was done with a full breath-hold to guarantee constant picture quality. Later, the original DICOM format images were transformed to JPEG while keeping crucial diagnostic information.

Figure 2 CT-Scan image labeled as Benign in dataset

Figure 4 CT-Scan image labeled as Normal in dataset

Figure 3 CT-Scan image labeled as Malignant in dataset

3.2. Data Preprocessing

Preparing the Data is a crucial part of the deep learning pipeline especially when working with image datasets. Deep Learning models' performance is heavily dependent upon the data quality and structure. Preprocessing examples include resizing, normalization, augmentation, and noise reduction, which help the model efficiently learn from the data and generalize to unseen data. To cup on an example of OSG preprocessing procedure, there are numerous iterative steps undertaken in image data form preprocessing, which not only help standardize the input but also help enhance the features by reducing the overfitting along with improving the convergence in training. As part of my research, I incorporated some data preprocessing strategies specific to the properties of the image dataset and will explain it to you in the upcoming sections.

3.2.1. Data Transformation

During this phase we performed the initial and necessary transformations needed to make the lung cancer CT scan images ready for input into the deep learning models. The original images in the dataset varied in size (512x512) as a preliminary stage, all of them were resized into the same shape 224x224x3, as this is a common input size for many convolutional neural networks (CNNs), in particular ResNet or VGG. This helps to make sure that there are not issues related to input dimensions when the model processing the images.

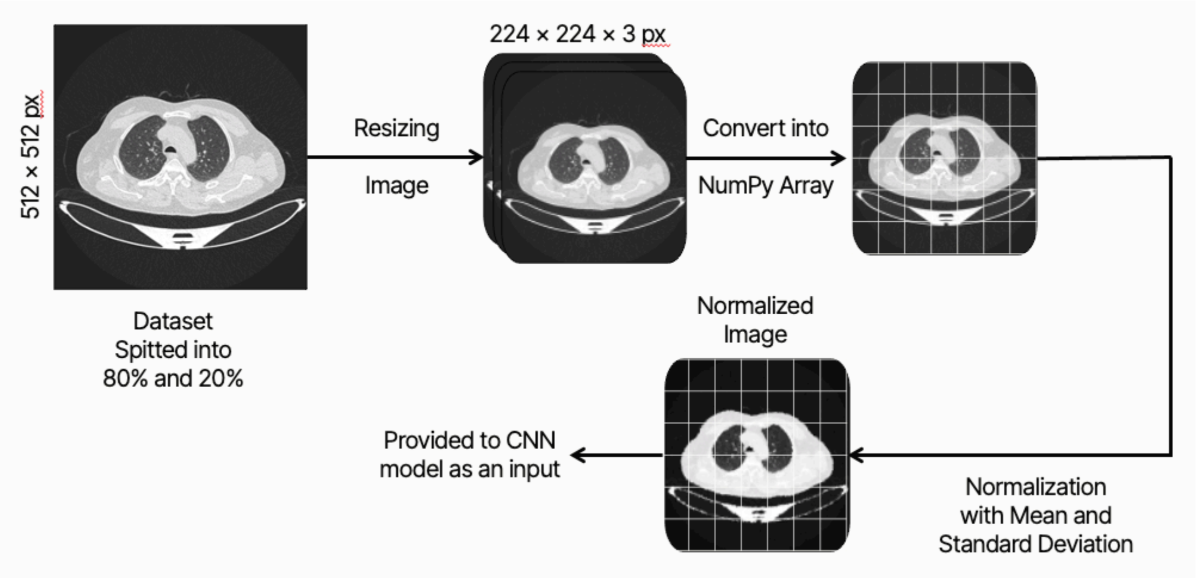
After resizing, the images were transformed into NumPy arrays for fast processing and compatibility with deep learning libraries such as PyTorch and TensorFlow. I then applied the standard mean and standard deviation values for popular pre-trained models to normalize the pixel values:

- Mean = [0.485, 0.456, 0.406]
- STD = [0.229, 0.224, 0.225]

This step in normalization scales the pixel values to a range represented in the typical used ImageNet pre-trained models. It is used for faster convergence during training. Here I have scaled the data so that the input is in the right format to give better results to the model.

Finally, after these transformations, I applied the 80-20 split of the data that has been described in the previous section in order to prepare the data for training and validation.

Figure 5 Illustrating the data preprocessing pipeline



3.2.2. Data Splitting

In the process of training and evaluating deep learning models, data splitting plays a significant role in dataset preparation. By isolating distinct subsets of data for training and validation, ensuring the model can generalize well to new, unseen data. For my research, I used CT scan data for lung cancer which had three classes – **Normal, Benign, Malignant**. Due to the skew nature of the data, it was therefore important that the proportions of each category remain similar across the training and validation splits.

I performed a conventional 80% training 20% validation split of the dataset, ensuring equal representation of classes in each set. By using this strategy, the model was able to be trained on a multitude of images but create an evaluation set that was even across the three groups of images for testing its accuracy.

Table 2 Distribution of lung cancer CT scan images across training and validation sets, with an 80-20 split for each category (Normal, Benign, Malignant).

Category	Total Images	Training Images	Validation Images
Normal	415	332	83
Benign	120	96	24
Malignant	571	457	114
Total	1106	885	221

3.3. Models Selection

Here we experimented with nine of the well-known CNN models, each with their own unique features and architectural innovations appropriate to image classification tasks.

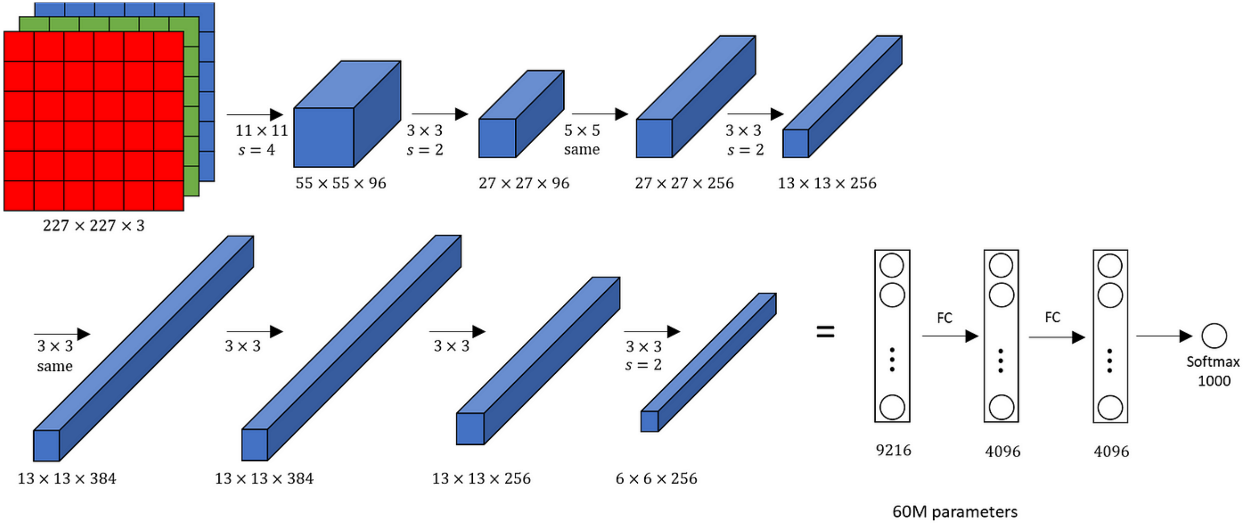
3.3.1. 9 CNN Models for Selection Process

We selected nine state of the art (SOTA) convolutional neural network (CNN) deep learning models for the model selection proceses that have significant reputation among deep learning and medical imaging sector. Below is the brief summery of each model:

3.3.1.1. AlexNet

Developing the first revolutionary application of deep learning to computer vision, AlexNet appeared in 2012. This model is comprised of five convolutional layers and three fully connected layers. AlexNet was amongst the first models to use GPUs efficiently for training and it used ReLU activation and dropout regularization to avoid overfitting. Originally devised for classifying images in the ImageNet contest, it has become the basis of increasingly deeper and more complex models.

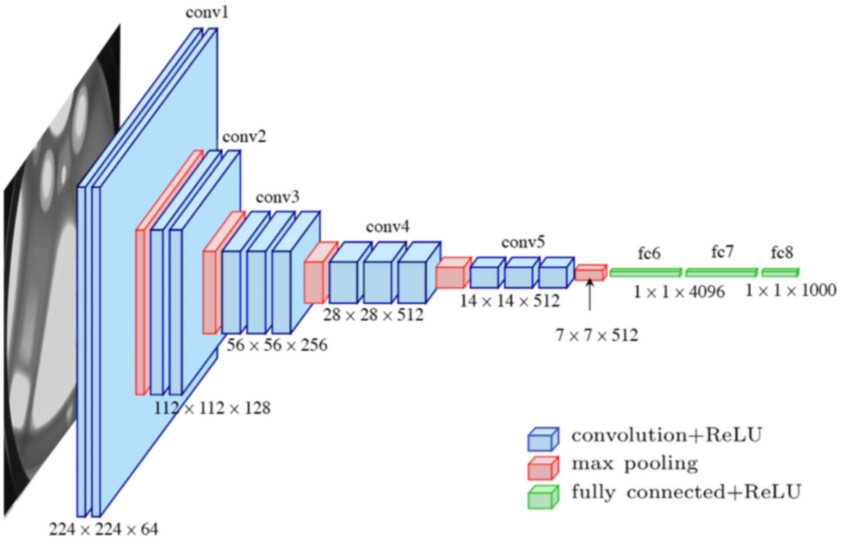
Figure 6 AlexNet Architecture



3.3.1.2. VGG16

Visual Geometry Group (VGG) from Oxford developed VGG16 which is easier but highly effective architecture and it contains 16 layers that mostly make good use of 3×3 convolutional layers stacked on top of each other. VGG16 was remarkable due to its depth and uniform architecture of small receptive fields(like with large filter size and orange arrows) leading to state-of-the-art performance on many vision benchmarks. However, this simplicity comes with a lot of parameters, and thus, it becomes computationally expensive.

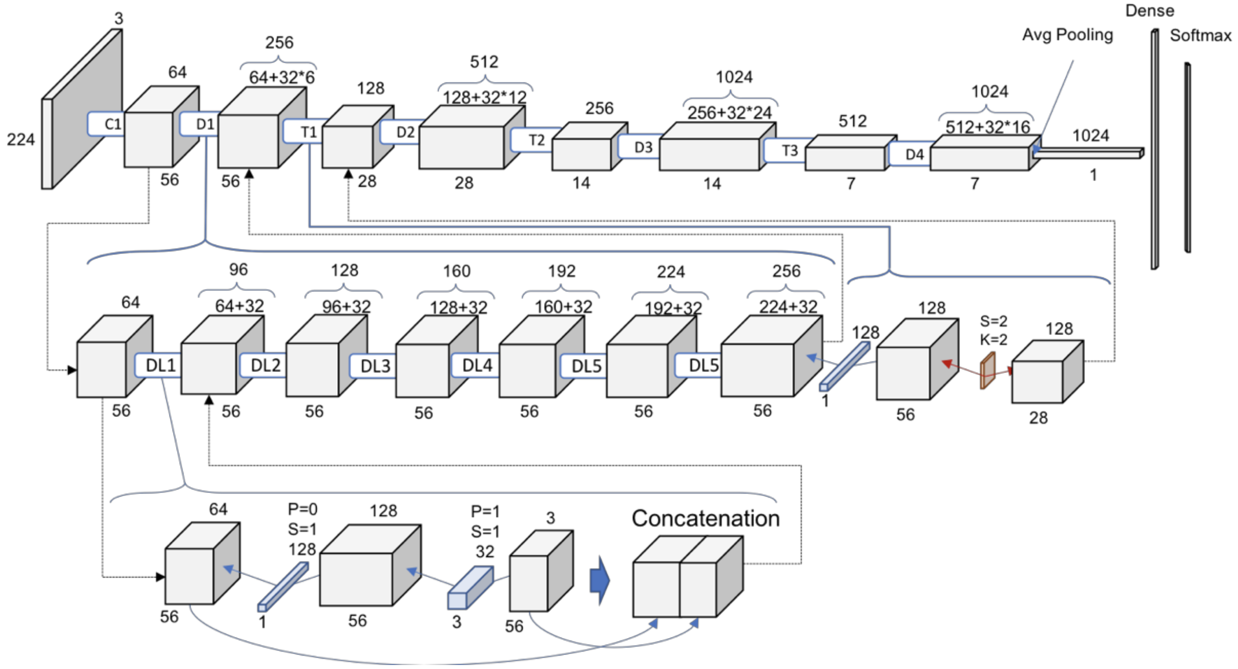
Figure 7 VGG16 Architecture



3.3.1.3. DenseNet121, DenseNet169, DenseNet201

DenseNet (Densely Connected Convolutional Networks), proposed in 2017, are intended to alleviate the vanishing gradient problem: Each layer is connected to every other layer in a feed-forward fashion. This allows each layer to access all preceding layers, enhancing feature reusability and thus efficient model training. The primary difference between DenseNet121, DenseNet169, and DenseNet201 is their depth, meaning DenseNet201 is deeper than the other two with the capability of learning notably more complex features and meaningful representations from the data.

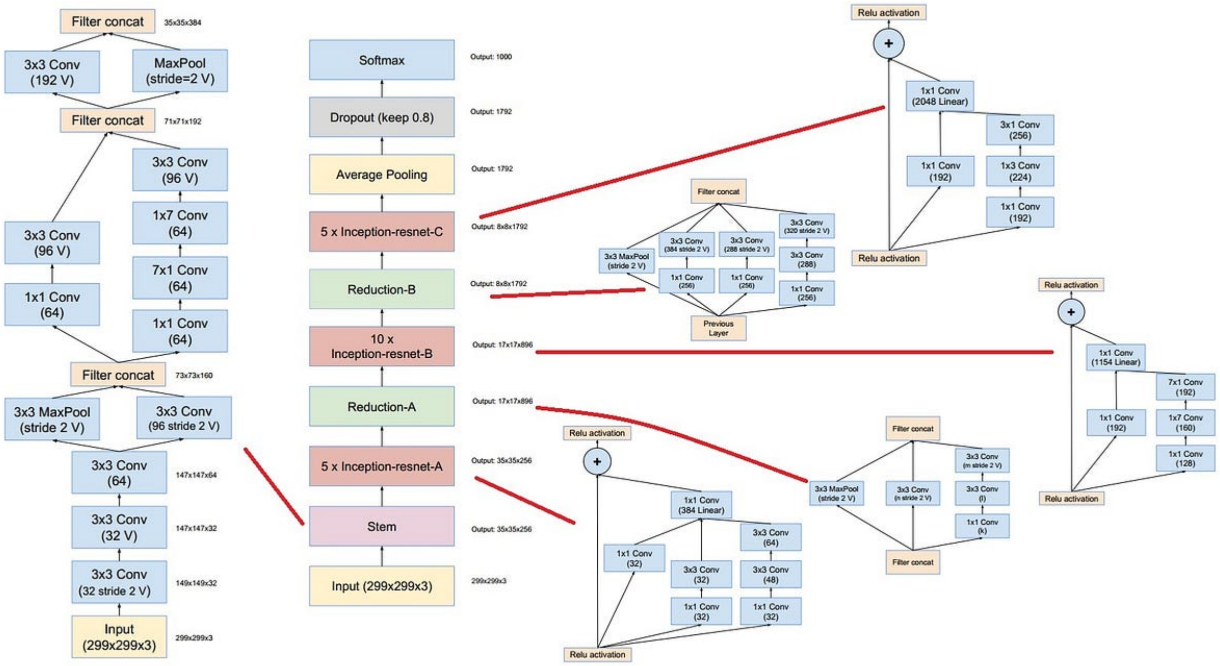
Figure 8 DenseNet121 Architecture



3.3.1.4. ResNetV2

ResNet(Residual Networks) – Skip Connections or residual connections were introduced in ResNet which allows the network to skip certain layers and thus solving the degradation problem in very deep networks. ResNetV2 is a modified version of ResNet that implemented batch normalization after each convolution layer to accommodate better performance. Additionally, this version takes advantage of identity mappings to make training deeper networks both less sensitive and more straightforward.

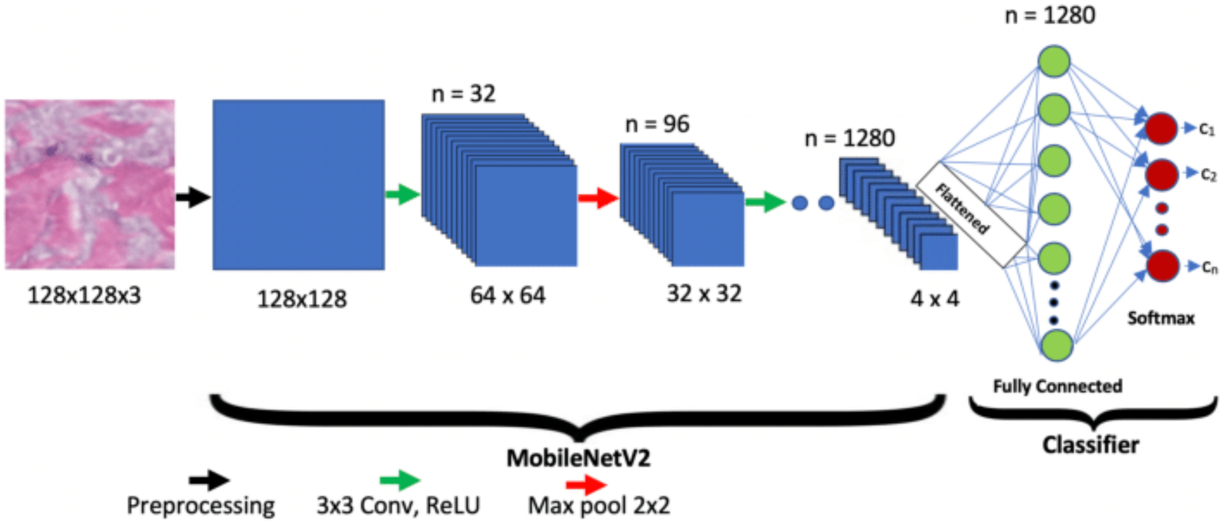
Figure 9 ResNetV2 Architecture



3.3.1.5. MobileNetV2

Mobile models designed to be lightweight, MobileNetV2 employs depthwise separable convolutions that greatly reduce number of parameters and computational cost while preserving high performance. It also employs linear bottleneck layers and an inverted residual architecture, which makes it suitable for resource-constrained mobile and embedded systems.

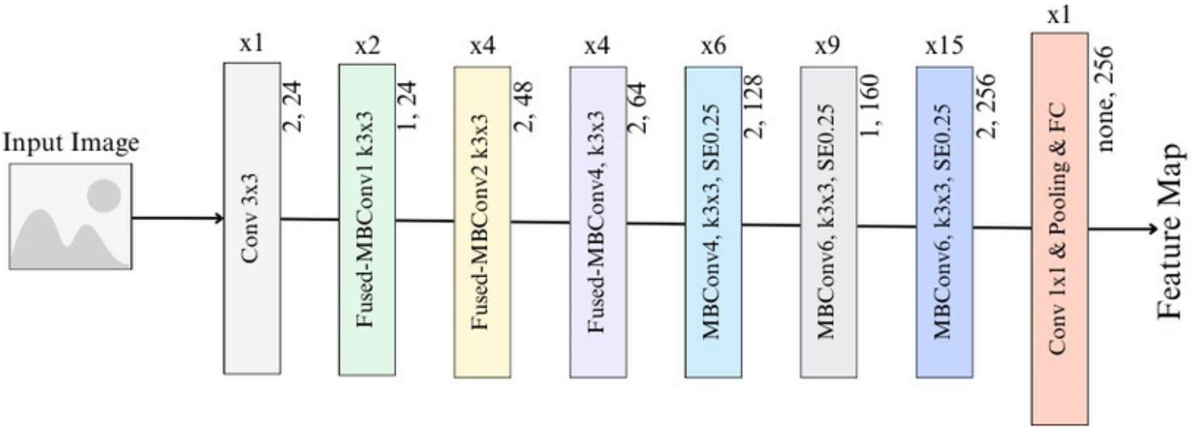
Figure 10 MobileNetV2 Architecture



3.3.1.6. EfficientNetV2

EfficientNetV2 is an advancement of the EfficientNet family of models, which minimised both precision and performance by avoiding aggregate pitching. EfficientNetV2 addresses this shortcoming by using an improved compound scaling method to build models that can be efficiently scaled in the depth, width, and input resolution dimensions, resulting in better performance with fewer parameters and faster training speed. This makes it relatively well-suited for applications in which computational cost is a concern but not the primary driver of performance.

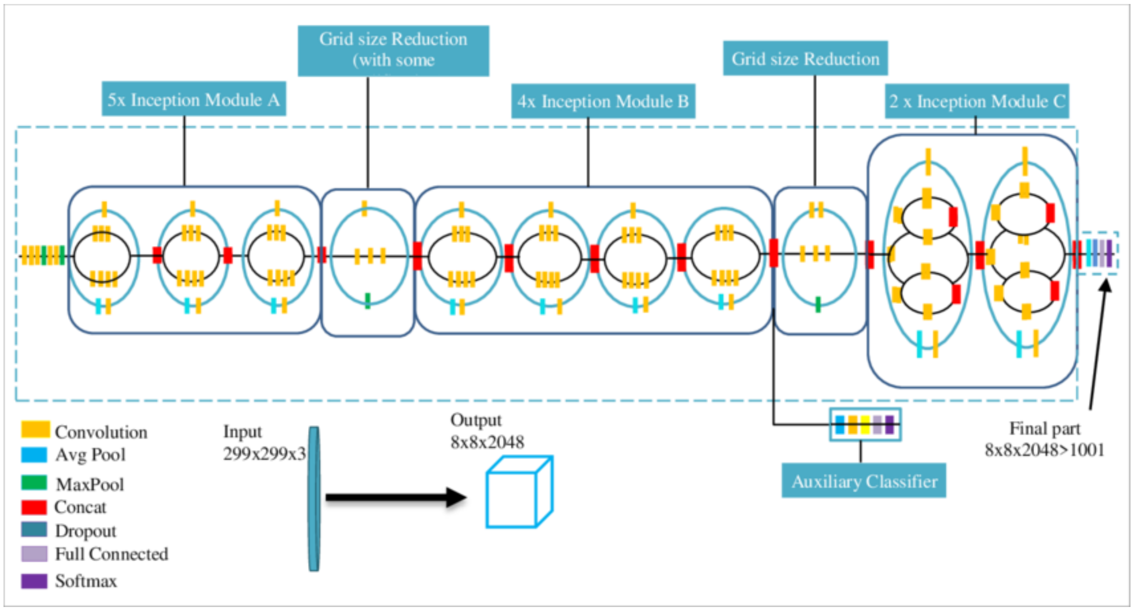
Figure 11 EfficientNetV2 Architecture



3.3.1.7. InceptionV3

A member of the GoogleNet family, Inception V3 is characterized by its modular architecture, allowing the model to effectively use varying types of convolutional layers to extract features from the image. They are applied, combined and repeated, similar to "Inception blocks" that use filters of three sizes to combine the results so as to pick up multiple spatial hierarchies of the image. This architecture also included techniques such as auxiliary classifiers and batch normalization to facilitate training and improve accuracy.

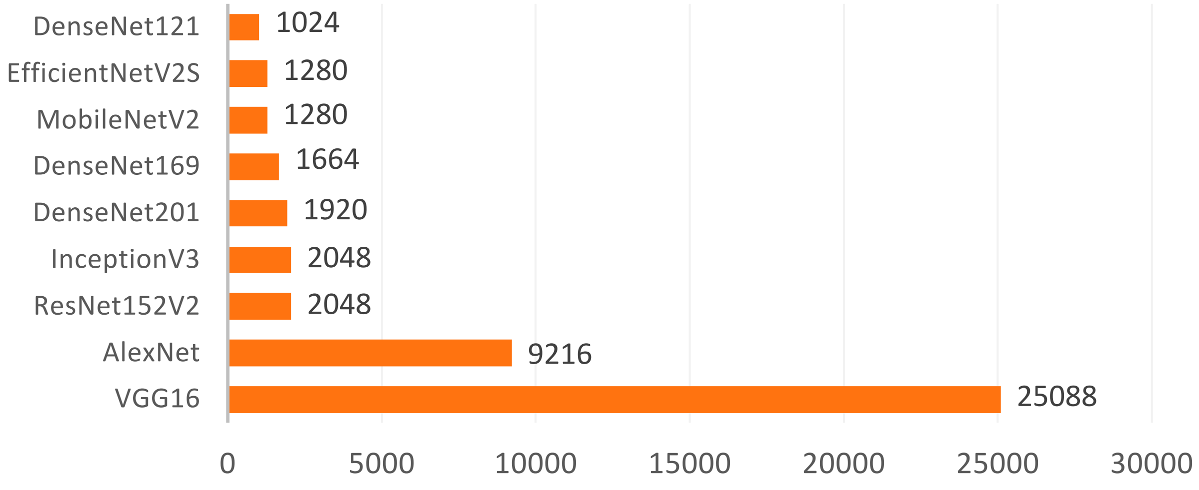
Figure 12 Inception-V3 Architecture



3.3.2. Models Evaluation Process

Having chosen these nine CNN models, we now used them to extract features while also evaluating them with external classifiers. Instead of just the classifiers already present within these models, we took the feature maps from the convolutional layers from each model, and used these as the input to 2 different external classifiers, SVC and K-Nearest Neighbors (KNN). This helped us to separate out the feature extraction capability of each model and compare the performances of SVC and KNN classifies for the last classification task.

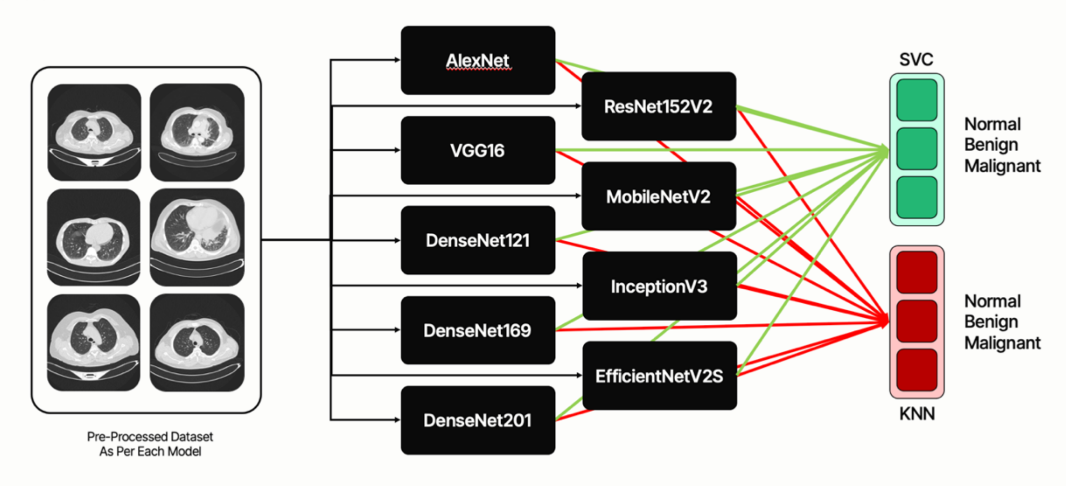
Figure 13 Number of features each model extracts without Dense Layer



Its procedure is as follows:

- **Feature Extraction:** The models were both trained based on the lung cancer CT scan dataset, and the have output outputs from the last convolutional layers. This output captured the rich, high-dimensional features the models had learned to extract from the images.
- **Classification:** The extracted features passed to SVC and KNN classifiers. The SVC was applied to determine the hyperplane that best separates the classes, while KNN assigned classes based on the distances from the nearest neighbors in the feature space.

Figure 14 Illustrating the data flow through the CNN models for feature extraction, followed by classification using SVC and KNN.



- **Evaluation:** For evaluation purposes, both the number of features extracted by each CNN model and the accuracy of the external classifiers (SVC and KNN) used for classification was compared information. These results gave us a deeper sense of each model’s ability to extract salient features and perform well in the final classification task.

Table 3 Classification accuracy achieved by each model when paired with SVC and KNN.

Models	SVC Accuracies (%)	KNN Accuracies (%)
AlexNet	98.20	97.30
VGG16	98.19	97.74
DenseNet121	96.39	96.84
DenseNet169	95.94	96.39
MobileNetV2	95.04	95.49

ResNet152V2	95.94	95.04
DenseNet201	97.29	94.14
EfficientNetV2S	94.14	95.94

Based upon the whole model selection process, results are shown in the *Table 2*, We got the best results with AlexNet and VGG16 as they both gave us excellent accuracies on both SVC and KNN classifier. So, for our proposed method/framework we selected both AlexNet and VGG16 as CNN architectures those could be utilized in Hybrid approach to get better results.

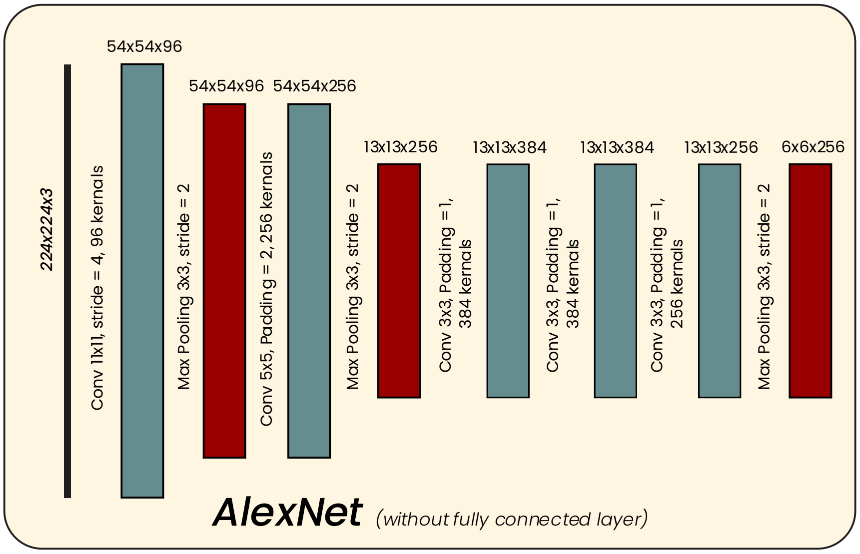
3.4. Proposed Hybrid Framework

After the whole critical but important phase of models' selection for our hybrid approach, it's time to develop our own hybrid framework. As we had reviewed multiple literatures that tells about the Hybrid CNN Models methodology but still there are some flaws/gaps those could be filled. As dataset preprocessing is explained earlier, now it's time to first modify the existing AlexNet and VGG16 architecture details are below:

3.4.1. Modified AlexNet

Feature selection is an essential part of any model, and our suggested hybrid architecture aims to use deep learning for both feature extraction and feature selection. For this, we tweaked the original AlexNet architecture, which has shown to be quite effective in picture classification tasks. Convolutional, pooling, and fully connected layers are all part of AlexNet's architecture, which allows it to learn hierarchical features from unprocessed image data more effectively. To implement our methods, we made specific adjustments to the network, removing the thick layers that AlexNet typically uses for final classification and decision-making. These layers were deemed unnecessary in our technique because an external mechanism within our framework was already responsible for managing their primary function, which was to execute feature selection.

Figure 15 Modified AlexNet Architecture



3.4.2. Modified VGG16

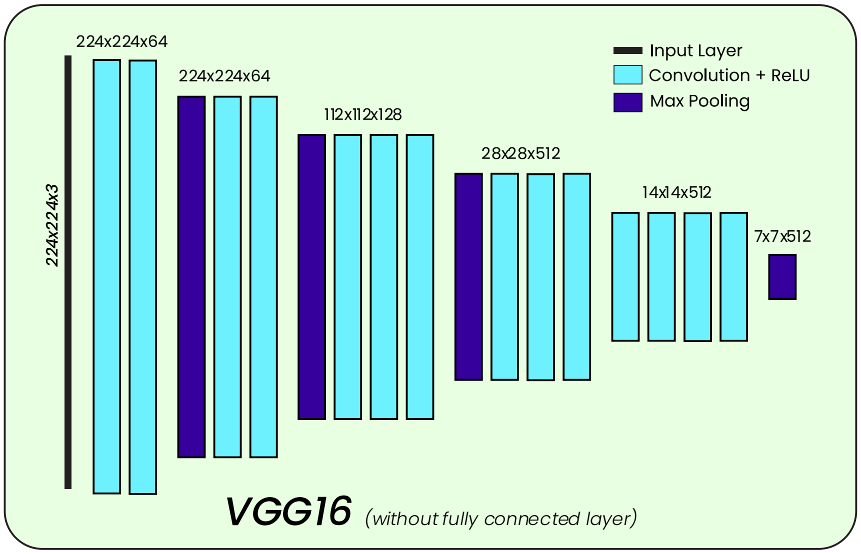
The VGG16 architecture was utilized as part of the hybrid framework for lung cancer detection and classification. VGG16, a convolutional neural network developed by the Visual Geometry Group (VGG) at the University of Oxford, is well-known for its uniform and deep architecture, which significantly enhances feature extraction capabilities in image classification tasks. It employs a straightforward approach using multiple 3×3 convolutional layers stacked together, ensuring small receptive fields while preserving spatial hierarchies.

The original VGG16 model's dense (completely connected) layers were eliminated for our modified framework, which only uses the convolutional layers to extract features. By removing these dense layers, we made sure that the model produced high-dimensional feature maps without adding the classification bias that came with the original dense layers' pre-trained weights. External classifiers, such as Support Vector Classification (SVC) and K-Nearest Neighbors (KNN), then used these extracted features to enhance classification performance.

The modified VGG16 demonstrated a noteworthy ability to capture crucial traits required for lung cancer categorization. Prior to being fed into the classification models, the convolutional layers extracted spatial and hierarchical features from input images that had been expanded to 224×224×3 dimensions. By reducing computing complexity without compromising the accuracy of generated features, this straightforward adaptation enhanced the model's appropriateness for medical imaging tasks.

The VGG16 model's seamless integration with the broader hybrid framework and support for efficient feature selection made it possible to achieve the high classification accuracies in this study.

Figure 16 Modified VGG16 Architecture



3.4.3. Features Selection

In order to identify important characteristics that support precise lung cancer diagnosis and classification, feature selection was essential to this study. We accomplished this by using modified versions of the popular convolutional neural networks (CNNs) AlexNet and VGG16 models, which are renowned for their effectiveness in processing picture data. To concentrate just on the convolutional layers for feature extraction, dense (completely connected) layers were eliminated from both models.

Computed tomography (CT) scans made up the preprocessed dataset, which was split into training (80%) and validation (20%) subsets. The improved AlexNet and VGG16 models were used to extract features from the training dataset. Each CT scan was fed through the models' convolutional layers after being scaled to a consistent 224 x 224 x 3 pixel size. These layers created feature maps that represented important spatial and hierarchical aspects of the images by applying several convolutional procedures, pooling, and activation functions.

With its comparatively simpler architecture, the improved AlexNet generated feature vectors of size 9,216 that contained the key characteristics found in each of its five convolutional layers. AlexNet's lightweight design allowed it to extract features effectively while preserving enough depth to identify pertinent patterns in the image data.

However, the updated VGG16 produced feature vectors of size 24,088 because to its deeper design and more convolutional layers. VGG16 is a very good model for identifying complex patterns in the dataset because of the consistent structure of its convolutional layers, which guaranteed the extraction of detailed and hierarchical features.

3.4.4. Features Concatenation

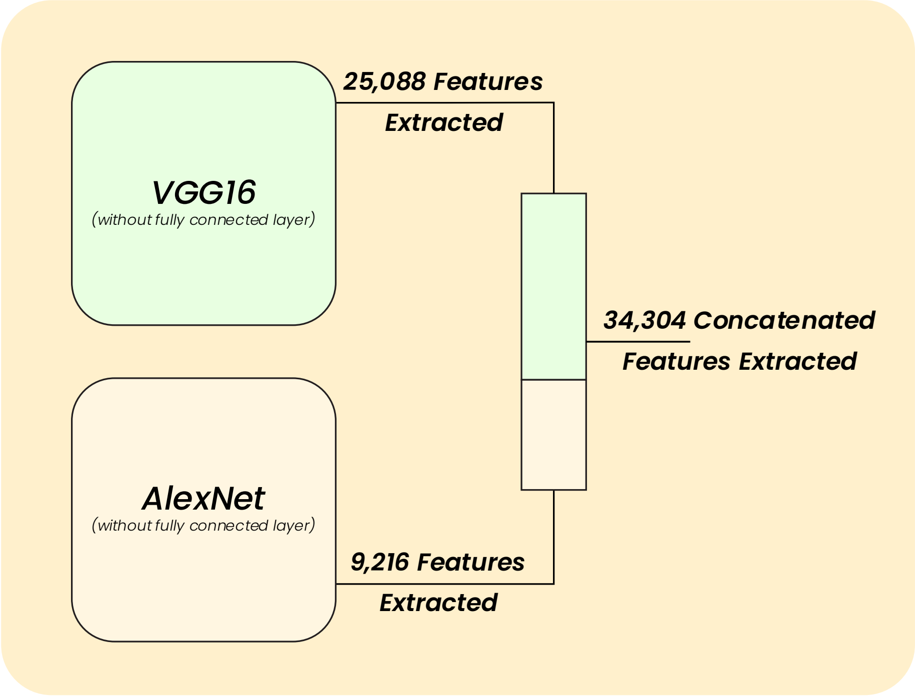
Following feature extraction, a unified feature representation was created by concatenating the feature vectors produced by the updated AlexNet and VGG16 models. This stage sought to improve the overall classification performance by combining the advantages of both models and utilizing their unique strengths.

The updated VGG16 model extracted 24,088 features from the training dataset, whereas the modified AlexNet model yielded 9,216 features. Each image in the training dataset had a composite feature vector of size 33,304 after these features were concatenated horizontally. The spatial and hierarchical properties that were recorded by both models were maintained in a consistent way thanks to the horizontal concatenation.

This concatenation was done in order to take advantage of the complementary nature of the features that the two models were able to extract. While VGG16's deeper design collected more complex and hierarchical features, AlexNet's comparatively simpler architecture concentrated on catching broad patterns. The hybrid framework produced a more comprehensive and discriminative representation of the input data by merging these feature sets.

This concatenation process not only enhanced the performance of the hybrid framework but also highlighted the effectiveness of integrating features from multiple deep learning models to achieve superior results in medical image analysis tasks.

Figure 17 Illustration Features Concatenation



3.4.5. Optimal Features Selection via Lasso

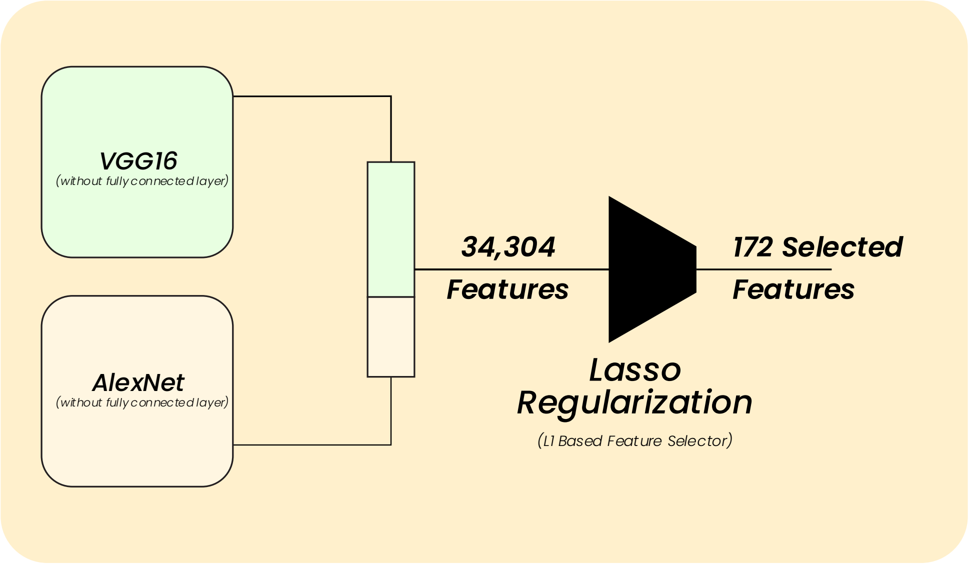
Concatenating features was essential since it enabled us to combine the advantages of both models. While VGG16, which is renowned for its deeper and more complex layers, extracted finer and more detailed characteristics, AlexNet, with its simpler design, was excellent at identifying broad patterns in the CT scan pictures. We successfully integrated many levels of abstraction into a single feature space by merging these two feature sets, guaranteeing a more varied and richer representation of the data.

Nevertheless, dealing directly with a feature set of this high dimensionality can be costly computationally and may result in redundancy, which could affect the system's overall performance. In order to solve this, Lasso L1 regularization was used to send the concatenated features to an ideal feature selection procedure. A popular dimensionality reduction method called Lasso assists in identifying and keeping only the most important features while removing less important ones.

$$\min \beta \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The feature vector was drastically shrunk from 33,304 dimensions to just **172 optimum features** as a result of this procedure. In addition to being extremely pertinent, these traits were chosen to guarantee that the classifiers employed in later phases could function more effectively and efficiently. By concentrating on these crucial elements, the system was able to reduce computing complexity and retain excellent accuracy, which made it more useful for real-world applications. A crucial part of our hybrid architecture was the combination of Lasso-based selection and feature concatenation, which allowed it to optimize the feature set for better classification performance while utilizing the greatest features of both VGG16 and AlexNet.

Figure 18 Illustration of Lasso Regularization Optimal Feature Selector in our proposed Method



3.4.6. Classification and Classifiers

Sorting data into predetermined classes according to its attributes is the aim of classification, a fundamental activity in machine learning. The objective of this study was to categorize lung CT scans into three groups: benign, malignant, and normal. Classifiers are algorithms that identify a given data point's class based on input features. These classifiers predict unknown data with variable accuracy by identifying patterns in the training data.

We used two well-known classifiers for our hybrid framework: Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN). These algorithms were selected because to their unique categorization strategies and high-dimensional data handling capabilities. A thorough description of each classifier and its function within our system may be found below.

3.4.6.1. K-Nearest Neighbors (KNN)

KNN is a proximity-based classification technique that is straightforward but efficient. By looking at the labels of a data point's closest neighbors in the feature space, it classifies the data point. One parameter that can be adjusted to maximize the classifier's performance is the number of neighbors, represented by k.

In KNN, the most common class among the k nearest neighbors is allocated to the data point, and the classification choice is usually done by majority voting. Metrics like the Euclidean distance are used in mathematics to determine the distance between data points:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Here, x and y are feature vectors, and nnn is the number of features. In our study, KNN was applied to the 172 optimal features selected through the Lasso-based feature selection process. The classifier's simplicity and reliance on local data distribution made it a suitable choice for this high-dimensional feature set.

3.4.6.2. Support Vector Classifier (SVC)

SVC is a strong and adaptable classification technique that is a variation of the Support Vector Machine (SVM). It operates by identifying the feature space hyperplane that best divides data points from various classes. The ideal hyperplane maximizes the margin, or the separation between the hyperplane and the support vectors—the closest data points from each class.

The hyperplane's equation is:

$$w \cdot x + b = 0$$

Where w is the weight vector, x is the input feature vector, and bbb is the bias term. For classification, the decision rule is:

By employing kernel functions (such as linear, polynomial, or radial basis function kernels) to convert non-linear data into a higher-dimensional space, SVC may also handle non-linear data. Because of its adaptability, SVC can efficiently categorize intricate datasets.

$$Class = sign(w \cdot x + b)$$

The same 172 features were classified using SVC in our framework, which took use of its capacity to manage high-dimensional data and identify the best decision boundaries.

Chapter 4

Results & Discussion

Chapter 4: Results & Conclusion

The results of our study on the hybrid deep learning framework for lung cancer classification and detection are shown in this chapter. The main goal is to assess the performance of the suggested model, which combines the best feature selection and classification methods with KNN and SVC with the feature extraction capabilities of modified AlexNet and VGG16. We evaluate our method's performance in identifying lung cancer from CT scan pictures by looking at accuracy, precision, recall, and other pertinent parameters.

With an amazing accuracy of **99.55% for the KNN** classifier and **98.65% for the SVC** classifier, our model's findings were quite encouraging. These outcomes show how resilient our hybrid framework is, accurately dividing cases of lung cancer into Normal, Benign, and Malignant categories. Such performance demonstrates the classifiers' ability to handle complex medical imaging data and the effectiveness of the feature extraction and selection process.

This chapter also highlights the research's contributions to the field of medical image analysis and provides a summary of the main conclusions derived from the findings. Along with suggestions for more advancements in lung cancer detection and classification systems, it also provides insights into possible directions for future research. The performance assessment and findings reported here will confirm the feasibility of the suggested hybrid framework for practical uses in early detection of lung cancer.

4.1. Performance Metrices

Several performance criteria were employed to thoroughly evaluate the efficacy of the KNN and SVC classifiers in our hybrid deep learning system for lung cancer detection and classification. These metrics consist of the confusion matrix, F1-score, recall, accuracy, and precision. We may gain a better understanding of the classifiers' performance and potential areas for improvement by examining these values. A detailed examination of the main performance matrices included in our research is provided below.

- The quantity of accurately predicted positive instances—that is, cases of lung cancer that are correctly diagnosed as either benign or malignant—is known as the True Positive (TP).
- The quantity of accurately classified normal instances, or correctly forecasted negative cases, is known as the True Negative (TN).
- The percentage of benign cases that are mistakenly identified as malignant or benign is known as the False Positive (FP) rate.
- False Negative (FN): The proportion of benign or malignant cases that are mistakenly labeled as normal.

For every class, the matrix aids in the computation of several other performance metrics, including precision, recall, and F1-score. For a multi-class situation like ours, the confusion matrix is usually structured as follows:

Table 4 Confusion Matrix, used to calculate the performance matrices

	Normal	**Benign**	**Malignant**
Normal	TP_1	FP_1	FP_2
Benign	FN_1	TP_2	FP_3
Malignant	FN_2	FN_3	TP_3

The values in this matrix are used to calculate:

- **Precision** refers to the percentage of positive predictions that are actually correct. It is calculated for each class using the formula:

$$Precision = \frac{TP + FP}{TP}$$

- **Recall** (also known as Sensitivity or True Positive Rate) measures the percentage of actual positives correctly identified by the classifier. It is calculated as:

$$Recall = \frac{TP + FN}{TP}$$

- **F1-Score** is the harmonic mean of precision and recall, providing a single metric that balances both concerns, particularly useful when dealing with imbalanced datasets. It is calculated as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **Acuracy** is one of the most straightforward and widely used performance metrics. It calculates the overall percentage of correct predictions (both true positives and true negatives) out of all predictions. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN \ + FP + FN}$$

These metrics provide a comprehensive view of the classifier's performance. High precision indicates a low false positive rate, while high recall means that the classifier is effectively identifying all true positive cases. The F1-score balances these two metrics, offering a robust measure of the classifier's overall performance.

4.2. Experimental Results

The effectiveness of the suggested hybrid deep learning framework for lung cancer detection and classification is confirmed by the experimental findings of this study. We made notable performance increases by combining the feature extraction capabilities of modified AlexNet and VGG16, followed by Lasso L1 regularization for optimal feature selection, then KNN and SVC for classification. The outcomes demonstrate the hybrid approach's strengths as well as its potential for use in practical settings for early and precise lung cancer diagnosis.

4.2.1. Classifiers Performance

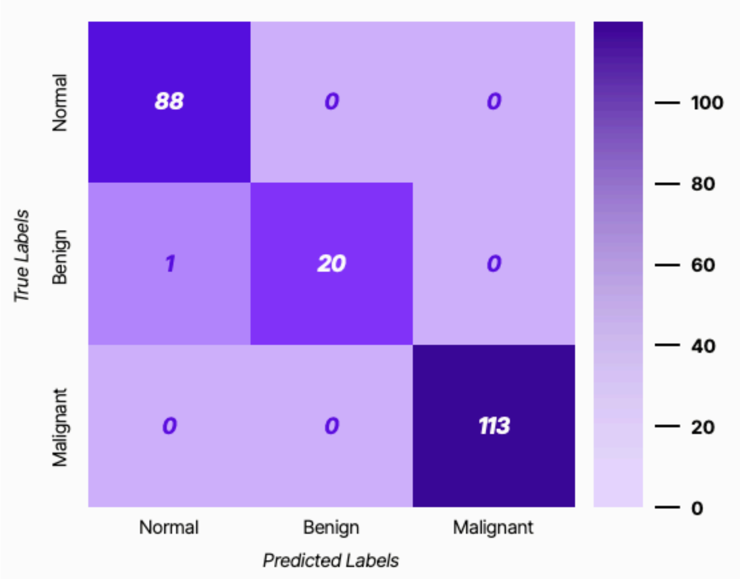
The 172 ideal features chosen by the Lasso L1 regularization were assessed using two classifiers, KNN and SVC. Accuracy, precision, recall, and F1-score were among the measures used to evaluate each classifier's performance.

4.2.1.1. KNN Performance

The KNN classifier's remarkable 99.55% accuracy rate showed how robust it is at differentiating between the Normal, Benign, and Malignant classes. These outstanding outcomes were made possible by KNN's proximity-based decision-making methodology and its superior feature set.

The distribution of accurate and inaccurate predictions among the three classes is further depicted by the KNN confusion matrix (Figure 19).

Figure 19 Confusion matrix, Illustrating the performance of KNN Classifier



4.2.1.2. SVC Performance

The SVC classifier demonstrated its capacity to efficiently classify the dataset using optimal features by achieving a competitive accuracy of 98.65%. Even with complicated and overlapping features, SVC performed well because of its built-in hyperplane optimization.

Chapter 5

Conclusion

Chapter 5: Conclusion

The research on our hybrid deep learning framework for the detection and classification of lung cancer is concluded in this chapter. It considers the main conclusions, talks about the ramifications, and suggests possible paths of inquiry for further research. The remarkable outcomes attained highlight how important it is to combine feature selection and classification methods for precise medical picture analysis.

5.1. Summary of Findings

The main objective of this study was to create a reliable and effective system that could use CT scan pictures to identify and categorize lung cancer. The suggested framework produced ground-breaking results by utilizing modified AlexNet and VGG16 models for feature extraction, Lasso L1-based optimal feature selection, and classification using KNN and SVC.

Key outcomes include:

- Extraction of **9,216 features** from AlexNet and **24,088 features** from VGG16, combined into a unified feature set of **33,304 features**.
- Reduction of the feature set to **172 optimal features** using Lasso L1 regularization, significantly reducing computational complexity while preserving performance.
- Exceptional classification accuracy of **99.55%** using KNN and **98.65%** using SVC, highlighting the framework's ability to differentiate between Normal, Benign, and Malignant lung CT scans.

These results validate the effectiveness of the hybrid approach and its potential to aid early diagnosis in clinical settings.

5.2. Contributions of the Research

This study significantly advances the fields of deep learning and medical image processing in a number of ways:

5.2.1. Hybrid Framework Development

The framework creates a revolutionary approach to lung cancer detection and classification by integrating external classifiers, optimal feature selection, and feature extraction.

Modified Deep Learning Models: AlexNet and VGG16's feature extraction capabilities were improved by removing their thick layers, which increased the models' suitability for use with medical datasets.

5.2.2. Optimal Feature Selection

By systematically reducing dimensionality through the use of Lasso L1 regularization, the classifier was able to function effectively without compromising accuracy.

5.2.3. Evaluation of Classifier Performance

A thorough examination of KNN and SVC revealed their advantages and offered a comparison, which helped to clarify these algorithms' application to high-dimensional medical datasets.

5.3. Implications of the Study

The research's conclusions have several theoretical and practical ramifications:

1. **Early Detection:** The framework's high accuracy highlights its potential to help detect lung cancer early, which would greatly improve patient outcomes by enabling prompt diagnosis.
2. **Clinical Integration:** The framework can be used to actual clinical settings due to the ease of the feature selection and categorization procedure.
3. **Scalability:** The framework's modular architecture enables its expansion to further medical imaging uses, such as the identification of additional illnesses.

5.4. Limitations of the Study

Despite its success, this research has certain limitations:

1. **Dataset Size:** While the results were promising, the framework’s performance could be further validated using larger and more diverse datasets.

2. **Feature Dependency:** The accuracy of the classifiers heavily depends on the quality of the selected features. Future research could explore additional feature selection methods to enhance robustness.
3. **Model Interpretability:** Deep learning models are often criticized for their lack of interpretability. This framework could benefit from techniques that explain the decision-making process of the classifiers.

5.5. Future Directions

The following directions for further research are suggested in order to build on the advantages and overcome the shortcomings of this study:

- **Dataset Expansion:** To broaden the framework's usefulness, test it on bigger datasets with different demographics and imaging settings.
- **Explainable AI (XAI) Integration:** Using XAI methodologies to help physicians better understand the framework's predictions will increase their confidence in AI-driven decision-making.
- **Real-Time Implementation:** Creating a system that operates in real-time and incorporates the framework into clinical workflows to guarantee smooth functioning and prompt diagnosis.
- **Investigating Other Classifiers:** To increase classification efficiency and accuracy, additional machine learning techniques, like Random Forests or Gradient Boosting, are being tested.
- **3D Imaging Analysis:** Using the framework to examine 3D CT scans could yield more thorough information on anomalies in the lungs.

References

References

[1] “Lung Cancer Key Fact,” World Health Organization (WHO). Accessed: Nov. 10, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[2] “Lung Cancer Awareness Month 2024. World Health Organization,” International Agency for Research on Cancer (IARC). Accessed: Nov. 10, 2024. [Online]. Available: https://www.iarc.who.int/news-events/lung-cancer-awareness-month-2024/

[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[4] K. Chung *et al.*, “Lung-RADS Category 4X: Does It Improve Prediction of Malignancy in Subsolid Nodules?,” *Radiology*, vol. 284, no. 1, pp. 264–271, Jul. 2017, doi: 10.1148/radiol.2017161624.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[6] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 2014, *arXiv*. doi: 10.48550/ARXIV.1409.1556.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” 2016, *arXiv*. doi: 10.48550/ARXIV.1608.06993.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” 2018, doi: 10.48550/ARXIV.1801.04381.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[10] M. Tan and Q. V. Le, “EfficientNetV2: Smaller Models and Faster Training,” Jun. 23, 2021, *arXiv*: arXiv:2104.00298. Accessed: Nov. 11, 2024. [Online]. Available: http://arxiv.org/abs/2104.00298

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[12] A. Mohammed Qadir, P. Ahmed Abdalla, and D. Faiq Abd, “A Hybrid Lung Cancer Model for Diagnosis and Stage Classification from Computed Tomography Images,” *Iraqi J. Electr. Electron. Eng.*, vol. 20, no. 2, pp. 266–274, Dec. 2024, doi: 10.37917/ijeee.20.2.23.

[13] R. Parveen, U. Saleem, K. Abid, and N. Aslam, “Identification of Lungs Cancer by using Watershed Machine Learning Algorithm,” *VFAST Trans. Softw. Eng.*, vol. 11, no. 2, pp. 70–79, Jun. 2023, doi: 10.21015/vtse.v11i2.1500.

[14] H. F. Kareem, M. S. AL-Huseiny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, “Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 3, p. 1731, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1731-1738.

[15] V. Deepa and P. M. Fathimal, “Deep-ShrimpNet fostered Lung Cancer Classification from CT Images,” *Int. J. Image Graph. Signal Process.*, vol. 15, no. 4, pp. 59–68, Aug. 2023, doi: 10.5815/ijigsp.2023.04.05.

[16] K. Khaliq, A. Naeem, N. Aslam, A. Malik, and K. Abid, “LCCNet: A Deep Learning Based Method for the Identification of Lungs Cancer using CT Scans,” *VFAST Trans. Softw. Eng.*, vol. 11, no. 2, pp. 80–93, Jun. 2023, doi: 10.21015/vtse.v11i2.1497.

[17] M. S. AL-Huseiny and A. S. Sajit, “Transfer learning with GoogLeNet for detection of lung cancer,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, p. 1078, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1078-1086.

[18] T. Gulsoy and E. Baykal Kablan, “FocalNeXt: A ConvNeXt augmented FocalNet architecture for lung cancer classification from CT-scan images,” *Expert Syst. Appl.*, vol. 261, p. 125553, Feb. 2025, doi: 10.1016/j.eswa.2024.125553.

[19] H. F. Al-Yasriy, M. S. AL-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, “Diagnosis of Lung Cancer Based on CT Scans Using CNN,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 928, no. 2, p. 022035, Nov. 2020, doi: 10.1088/1757-899X/928/2/022035.

[20] S. G. M. A and A. Jayachandran, “Triple SVM Integrated with Enhanced Random Region Segmentation for Classification of Lung Tumors,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, 2022, doi: 10.14569/IJACSA.2022.01310103.

[21] M. M. Jassim and M. M. Jaber, “Hybrid selection framework for class balancing approaches based on integrated CNN and decision making techniques for lung cancer diagnosis,” *East.-Eur. J. Enterp. Technol.*, vol. 4, no. 9(118), pp. 69–76, Aug. 2022, doi: 10.15587/1729-4061.2022.263644.

[22] Anand, Rameshwari, and D. Sumukh, “Lung cancer detection and prediction using deep learning,” pp. 313–320, May 2022.

[23] J. M. Arockia Dass and K. S. Magesh, “A Novel Approach for Small Object Detection in Medical Images through Deep Ensemble Convolution Neural Network,” *Int. J. Adv. Comput. Sci. Appl.*, vol. Vol. 13, No. 3, pp. 1–7, 2022.

[24] Md. I. Hossain, S. Ahmad, S. Hasan, Md. A. Rahim, and J. Shin, “Lung cancer detection and classification using LeNet-LSTM model on computed tomography images”, doi: 10.31893/multiscience.2025188.

[25] S. V. G. Reddy, V. Bhuvaneshwari, A. Kumar Tikariha, Y. Aniket, and B. Sesha Sahith Ra, “Automatic pulmonary nodule detection in ct scans using xception, resnet50 and advanced convolutional neural networks models,” *Int Res J EngTechnol*, vol. 09, Apr. 2022.

[26] S. Pathan, T. Ali, S. P G, V. K. P, and D. Rao, “An optimized convolutional neural network architecture for lung cancer detection,” *APL Bioeng.*, vol. 8, no. 2, p. 026121, Jun. 2024, doi: 10.1063/5.0208520.

[27] A. Gupta, A. Kumar, and K. Rautela, “UDCT: lung Cancer detection and classification using U-net and DARTS for medical CT images,” *Multimed. Tools Appl.*, Jul. 2024, doi: 10.1007/s11042-024-19801-9.

[28] M. M. Musthafa, I. Manimozhi, T. R. Mahesh, and S. Guluwadi, “Optimizing double-layered convolutional neural networks for efficient lung cancer classification through hyperparameter optimization and advanced image pre-processing techniques,” *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 142, May 2024, doi: 10.1186/s12911-024-02553-9.

[29] Y. Kumaran S, J. J. Jeya, M. T. R, S. B. Khan, S. Alzahrani, and M. Alojail, “Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam,” *BMC Med. Imaging*, vol. 24, no. 1, p. 176, Jul. 2024, doi: 10.1186/s12880-024-01345-x.

[30] T. I. A. Mohamed, O. N. Oyelade, and A. E. Ezugwu, “Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm,” *PLOS ONE*, vol. 18, no. 8, p. e0285796, Aug. 2023, doi: 10.1371/journal.pone.0285796.

[31] S. Majumder, N. Gautam, A. Basu, A. Sau, Z. W. Geem, and R. Sarkar, “MENet: A Mitscherlich function based ensemble of CNN models to classify lung cancer using CT scans,” *PLOS ONE*, vol. 19, no. 3, p. e0298527, Mar. 2024, doi: 10.1371/journal.pone.0298527.

[32] Hamdalla Fadil Al-Yasriy, “The IQ-OTH/NCCD lung cancer dataset.” Mendeley, Jan. 03, 2023. doi: 10.17632/BHMDR45BH2.3.

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Crossref database
- Crossref posted content database

Match Groups

114 Not Cited or Quoted 12%
Matches with neither in-text citation nor quotation marks

2 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5%	Internet sources
4%	Publications
10%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **114** Not Cited or Quoted 12%
  Matches with neither in-text citation nor quotation marks
- **2** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5%	Internet sources
4%	Publications
10%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	**Coventry University on 2023-08-08**	**1%**
2	Submitted works	**University of Hertfordshire on 2024-01-08**	**0%**
3	Submitted works	**Liverpool John Moores University on 2023-12-18**	**0%**
4	Submitted works	**University of Huddersfield on 2024-01-16**	**0%**
5	Internet	**www.ncbi.nlm.nih.gov**	**0%**
6	Submitted works	**University of Essex on 2024-08-21**	**0%**
7	Publication	**C Kishor Kumar Reddy, P R Anisha, Samiya Khan, Marlia Mohd Hanafiah, Lavanya ...**	**0%**
8	Submitted works	**University of Malaya on 2024-06-14**	**0%**
9	Submitted works	**University of Ulster on 2023-05-12**	**0%**
10	Submitted works	**Mid-Kent College on 2024-05-06**	**0%**

11	Submitted works	
Manchester Metropolitan University on 2024-10-04		0%
12	Submitted works	
University of Asia Pacific on 2022-12-25		0%
13	Submitted works	
University of Sydney on 2022-04-09		0%
14	Internet	
www.biorxiv.org		0%
15	Submitted works	
The University of Wolverhampton on 2024-05-10		0%
16	Submitted works	
University of Shajrah on 2024-11-05		0%
17	Submitted works	
University of Technology, Sydney on 2023-11-03		0%
18	Internet	
livrosdeamor.com.br		0%
19	Internet	
www.researchsquare.com		0%
20	Submitted works	
Konya Gıda ve Tarım Üniversitesi on 2023-05-17		0%
21	Submitted works	
Liverpool John Moores University on 2024-12-21		0%
22	Submitted works	
University of Hertfordshire on 2024-12-13		0%
23	Publication	
V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challeng...		0%
24	Internet	
www.ijdiic.com		0%

25 Submitted works

Leeds Beckett University on 2024-09-06 0%

26 Submitted works

Middlesex University on 2024-10-06 0%

27 Submitted works

University of Dundee on 2024-08-22 0%

28 Internet

www.irejournals.com 0%

29 Submitted works

Metropolia Ammattikorkeakoulu on 2023-05-07 0%

30 Publication

Rohit Tanwar, Surbhi Bhatia Khan, Varun Sapra, Neelu Jyoti Ahuja. "Artificial Intel... 0%

31 Internet

discovery.researcher.life 0%

32 Internet

www.mdpi.com 0%

33 Submitted works

Higher Education Commission Pakistan on 2024-06-22 0%

34 Submitted works

Houston Community College on 2024-05-04 0%

35 Submitted works

Napier University on 2023-12-15 0%

36 Submitted works

University of Warwick on 2018-04-26 0%

37 Submitted works

Multimedia University on 2017-10-14 0%

38 Submitted works

University of Nottingham on 2024-09-05 0%

39	Internet	
hdl.handle.net		0%
40	Internet	
ijritcc.org		0%
41	Submitted works	
College of Engineering, Pune on 2024-10-11		0%
42	Internet	
www.american-cse.org		0%
43	Internet	
doaj.org		0%
44	Internet	
eeb.org		0%
45	Internet	
www.civilenvironjournal.com		0%
46	Internet	
www.researchgate.net		0%
47	Submitted works	
The University of the South Pacific on 2023-05-31		0%
48	Internet	
content.iospress.com		0%
49	Internet	
www.diva-portal.org		0%
50	Publication	
Abu-Ghoush, Layan Salah. "An Integrated Framework Using Variable Encoding-TF...		0%
51	Submitted works	
Higher Education Commission Pakistan on 2023-06-10		0%
52	Submitted works	
University of Bedfordshire on 2024-12-13		0%

53	Submitted works	
	University of North Texas on 2023-04-27	0%
54	Submitted works	
	University of North Texas on 2023-04-27	0%
55	Submitted works	
	University of Westminster on 2023-07-17	0%
56	Internet	
	par.nsf.gov	0%
57	Internet	
	pesquisa1.bvsalud.org	0%
58	Submitted works	
		0%
59	Submitted works	
	British University in Egypt on 2024-06-08	0%
60	Submitted works	
	National College of Ireland on 2019-08-11	0%
61	Submitted works	
	The Robert Gordon University on 2023-08-22	0%
62	Submitted works	
	Indian Institute of Technology, Madras on 2021-08-03	0%
63	Publication	
	Vivek S. Sharma, Shubham Mahajan, Anand Nayyar, Amit Kant Pandit. "Deep Lear...	0%
64	Submitted works	
	Liverpool John Moores University on 2020-06-01	0%
65	Submitted works	
	University of Glasgow on 2023-04-27	0%
66	Submitted works	
	University of Hertfordshire on 2023-08-21	0%

67	**Submitted works**
University of Technology, Sydney on 2023-04-14	**0%**