A project report on

IMAGE SEGMENTATION OF LUNG HISTOPATHOLOGICAL IMAGE

and

PREDICTING LUNG CANCER USING ML ALOGORITHM

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

by

AAYUSH KUMAR SINGH (19BCE1113)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2023

IMAGE SEGMENTATION OF LUNG HISTOPATHOLOGICAL IMAGE

and

PREDICTING LUNG CANCER USING ML ALOGORITHM

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

by

AAYUSH KUMAR SINGH (19BCE1113)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2023

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

DECLARATION

I here by declare that the thesis entitled “IMAGE SEGMENTATION OF LUNG HISTOPATHOLOGICAL IMAGE and PREDICTING LUNG CANCER USING ML ALGORITHM ” submitted by me, for the award of the degree of Bachelor of Technology in Computer Science and Engineering, Vellore Institute of Technology, Chennai, is a record of bonafide work carried out by me under the supervision of Dr. Revathi M.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: 24 – 04 – 2023 Signature of the Candidate

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

School of Computer Science and Engineering

CERTIFICATE

This is to certify that the report entitled **“Image Segmentation of lung histopathological image and predicting lung cancer using ML algorithm”** is prepared and submitted by **Aayush Kumar Singh** (**19BCE1113**) to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** programme is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:

Name: Dr. Revathi M.

Date: 24-04-2023

Signature of the Examiner 1	Signature of the Examiner 2
Name:	Name:
Date: 24-04-2023	Date: 24-04-2023

Approved by the Head of Department
B. Tech. CSE

Name: Dr. Nithyanandam P

Date: 24 – 04 – 2023

(Seal of SCOPE)

ABSTRACT

Cancer is one of the hazardous diseases that many people suffer especially lung cancer. Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body making it difficult to treat. It can be controlled if the presence of cancer is detected at initial stage itself. This proposed work aims to do image segmentation on lung Histopathological image and data analysis on the lung cancer dataset using machine learning algorithm.

In the proposed methodology initially rhe images are preprocessed to enhance the contrast and removal of noise. Then, the images are segmented using the deep neural network that has shown promising results in medical image segmentation. The segmented images are then used to extract features, such as texture and shape, which are fed into various machine learning algorithms.

In the proposed work, 14 machine learning algorithms were used, including support vector machines (SVM), random forest, K-nearest neighbors (KNN), logistic regression, naive Bayes, CatBoost, XGBoost, decision tree, artificial neural network (ANN), convolutional neural network (CNN), Inception-V3, ResNet50, VGG16, and EfficientNet-B7. These algorithms were trained and tested on a dataset of 15000 lung histopathological images with corresponding clinical data.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. Revathi M., Assistant Professor, SCOPE, Vellore Institute of Technology, Chennai, for her constant guidance, continual encouragement, understanding; more than all, she taught me patience in my endeavor. My association with her is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Machine learning and Medical Image Processing.

It is with gratitude that I would like to extend thanks to our honorable Chancellor, Dr. G. Viswanathan, Vice Presidents, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan and Mr. G V Selvam, Assistant Vice-President, Ms. Kadhambari S. Viswanathan, Vice-Chancellor, Dr. Rambabu Kodali, Pro-Vice Chancellor, Dr. V. S. Kanchana Bhaaskaran and Additional Registrar, Dr. P.K.Manoharan for providing an exceptional working environment and inspiring all of us during the tenure of the course.

Special mention to Dean, Dr. Ganesan R, Associate Dean Academics, Dr. Parvathi R and Associate Dean Research, Dr. Geetha S, SCOPE, Vellore Institute of Technology, Chennai, for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant mood I express ingeniously my whole-hearted thanks to Dr. Nithyanandam P, Head of the Department, Project Coordinators, Dr. Abdul Quadir Md, Dr. Priyadarshini R and Dr. Padmavathy T V, B. Tech. Computer Science and Engineering, SCOPE, Vellore Institute of Technology, Chennai, for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staff at Vellore Institute of Technology, Chennai, who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai

Date: 24 – 04 – 2023   Aayush Kumar Singh

CONTENTS

CHAPTER 1

CHAPTER 2

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ANN	Artificial Neural Network
CNN	Convolution Neural Network
KNN	K-Nearest Neighbor
SVM	Support Vector Machine
ML	Machine Learning
SCLC	Small Cell Lung Cancer
ReLU	Rectified Linear Unit

Chapter 1

INTRODUCTION

The most hazardous illness to ever strike humans and the leading cause of fatal death, according to current thinking, is cancer. Lung cancer is one of the most dangerous types of cancer when compared to other cancers in the globe.

Lung cancer has the highest mortality rate of all cancers. Lung cancer is one of the deadliest cancers in the world, with one of the lowest post-diagnosis survival rates and the highest annual death toll.

Survival is directly connected to the stage of lung cancer at the time of finding. With early discovery, a successful treatment is more likely. For 85% of male cases and 75% of female cases of lung cancer, the primary risk factor is believed to be cigarette smoking.

In order to overcome this challenge, the proposed work outlines a network and training strategy that relies on the strong use of data augmentation. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization.

1.1 OVERVIEW

The proposed work starts with an introduction to the problem of lung cancer and the potential of using machine learning algorithms for automated diagnosis. Then, various machine learning algorithms are used, including SVM, random forest, KNN, logistic regression, naive Bayes, CatBoost, XGBoost, decision tree, ANN, CNN, Inception-V3, ResNet50, VGG16, and EfficientNet-B7.

The methodology of the study is then presented, which includes the data collection process, data pre-processing, and data augmentation. Then the implementation of the various machine learning algorithms for image segmentation and prediction of lung cancer, along with the various evaluation metrics used for performance evaluation.

The study's conclusions, including how well various machine learning algorithms perform at image segmentation and lung cancer prediction, are presented and discussed.

Finally, the proposed study concludes with a summary of the study's findings and their implications for automated diagnosis systems for lung cancer detection using histopathological images of the lungs. Future directions for research in this area, such as the integration of multiple algorithms for improved performance and the development of more robust data augmentation techniques.

1.2 CHALLENGES

The challenges while applying ML algorithms for detecting lung cancer includes:

1. **Limited data availability:** One of the main challenges in developing machine learning algorithms for lung cancer diagnosis is the limited availability of annotated data. The lack of data can affect the performance of the algorithms and limit their applicability in clinical settings.

2. **Variability in histopathological images:** The histopathological images of the lungs can vary significantly in terms of their quality, lighting, and resolution. This can affect the performance of the algorithms and make it difficult to generalize their results to new images.

3. **Overfitting and generalization**: Machine learning algorithms are prone to overfitting when the model is trained on a small dataset or when the model is too complex. Overfitting can lead to poor generalization performance, which can affect the accuracy of the algorithm in predicting lung cancer.

4. **Interpreting the results:** Machine learning algorithms are often considered black-box models, making it difficult to interpret the results and understand how the model arrived at its conclusions. This can limit the confidence in the predictions and make it difficult to use the algorithms in clinical settings.

5. **Computational complexity:** Some of the machine learning algorithms used in the methodology, such as CNN, Inception-V3, ResNet50, VGG16, and EfficientNet-B7, are computationally expensive and require significant computational resources. This can make it difficult to deploy these algorithms in resource-limited settings.

1.3 PROBLEM STATEMENT

In today’s world survival of cancer patient increases with early detection of cancer. Histopathological images can provide us a lot of information about the tissues of the lungs, but deciphering them requires a lot of work and expertise.

Machine learning algorithms can be used to automate the interpretation process and improve the accuracy and efficacy of lung cancer diagnosis.

The proposed work aims to develop and evaluate machine learning algorithms, including SVM, random forest, KNN, logistic regression, naive bayes, CatBoost, XGBoost, decision tree, ANN, CNN, Inception-V3, ResNet50, VGG16, and EfficientNet-B7, for the segmentation of lung histopathological images and the prediction of lung cancer using these images.

The proposed methodology also seeks to address the challenges related to the limited availability of annotated data, the variability in histopathological images, the risk of overfitting and poor generalization, the difficulty in interpreting the results, and the computational complexity of some of the machine learning algorithms. By addressing these challenges, It aims to provide an accurate and reliable method for the diagnosis of lung cancer using histopathological images, which can be used in clinical settings to improve patient outcomes.

1.4 OBJECTIVE

The main objective of this proposed work is to develop a machine learning-based framework to accurately segment lung histopathological images and predict lung cancer. The proposed approach involves the use of a wide range of machine learning algorithms including Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, CatBoost, XGBoost, Decision Tree, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Inception-V3, ResNet50, VGG16, and EfficientNet-B7. These algorithms are applied to the histopathological image dataset to analyze and predict lung cancer.

Machine learning is a subfield of Artificial Intelligence. Machine learning is used in complex data classification and decision making. Machine learning gives systems the opportunity to learn automatically and improve over time without being directly configured.

Machine learning methods are currently being used in medical fields for identification and classification of various diseases. Recently machine learning methods have been employed in this area in predicting cancer in early stages, which can result in better chances of survival of patients.

Overall, the proposed framework aims to accurately segment lung histopathological images and predict lung cancer using various machine learning algorithms. The proposed approach can help pathologists in the diagnosis of lung cancer and improve the accuracy and speed of the diagnosis process. Additionally, the performance comparison of different machine learning algorithms and the investigation of data augmentation techniques can provide valuable insights into the development of effective machine learning models for the analysis of histopathological images.

1.5 SCOPE OF THE PROJECT

The scope of this work is to develop a machine learning-based system for image segmentation and prediction of lung cancer using histopathological images. It aims to segment the lung tissue regions from the images and predict the presence or absence of cancerous cells using different ML algorithms.

Proposed work focuses on the development of a system that can accurately segment the lung regions from histopathological images and predict the presence of cancer using various ML algorithms. Its primary goal is to improve the accuracy and efficiency of the current lung cancer detection methods.

Proposed methodoly secondary goal is to evaluate the performance of various ML algorithms for image segmentation and lung cancer prediction. The scope of the work includes the development of models using popular ML algorithms such as SVM, random forest, KNN, logistic regression, naive bayes, CatBoost, XGBoost, Decision Tree, ANN, CNN, Inception-V3, ResNet50, VGG16, and EfficientNet-B7.

Its scope also includes the preprocessing of the dataset and feature extraction. It also aims to provide a user-friendly interface that can be used by medical professionals for the analysis of lung histopathological images.

Proposed methodology can be expanded to include the analysis of other histopathological images for cancer detection using ML algorithms. The proposed system can be used as a base for developing automated diagnostic systems for various types of cancers, which can improve the accuracy and efficiency of cancer diagnosis.

1.6 MOTIVATION

Lung cancer is a significant global health problem that affects millions of people worldwide. Early diagnosis and treatment are crucial for improving the survival rate and quality of life for patients with lung cancer. One of the challenges in the diagnosis of lung cancer is the accurate interpretation of histopathological images, which are essential for determining the type and severity of the disease. However, the manual interpretation of these images can be time-consuming, subjective, and error-prone, leading to inaccurate diagnoses and treatment.

The motivation behind the proposed work is to develop a machine learning-based approach for the accurate and efficient interpretation of lung histopathological images. The aim is to improve the accuracy and speed of lung cancer diagnosis, which can lead to earlier detection and treatment, resulting in better patient outcomes.

The use of machine learning algorithms such as SVM, random forest, KNN, logistic regression, naive bayes, CatBoost, XGBoost, Decision Tree, ANN, CNN, Inception-V3, ResNet50, VGG16, and EfficientNet-B7 can aid in the segmentation of lung histopathological images and predicting lung cancer. These algorithms can learn from the vast amounts of data available in the lung histopathological image dataset, enabling them to make accurate predictions based on specific features and patterns.

The motivation behind the use of multiple machine learning algorithms is to determine the most accurate algorithm for the task of lung cancer detection and segmentation. By comparing the performance of different algorithms, It can identify the strengths and weaknesses of each algorithm and determine the best approach for this particular task.

Overall, the motivation is to contribute to the development of an accurate, efficient, and automated method for the interpretation of lung histopathological images for the diagnosis and treatment of lung cancer. The use of machine learning algorithms can aid in the early detection and accurate diagnosis of lung cancer, improving patient outcomes and quality of life.

Chapter 2

LITERATURE REVIEW

2.1 THE EVOLVING LANDSCAPE OF SEX-BASED DIFFERENCES IN LUNG CANCER: A DISTINCT DISEASE IN WOMEN

Ragavan, M., & Patel [1] study aims to discuss the evolving understanding of sex-based differences in lung cancer. It highlight the fact that lung cancer is now recognized as a distinct disease in women, with different risk factors, molecular profiles, and clinical outcomes compared to men.

Smoking is still the primary risk factor for lung cancer, but women who smoke are at greater risk than men who smoke. Additionally, non-smoking-related risk factors such as hormone exposure, air pollution, and genetic predisposition may play a larger role in the development of lung cancer in women.

Women are more likely to have adenocarcinomas, which are associated with specific molecular alterations, such as EGFR mutations and ALK rearrangements. These molecular differences have important implications for treatment, as targeted therapies can be more effective in patients with specific molecular alterations.

Women tend to have better survival rates than men, but may experience more treatment-related toxicities. Importance of sex-specific research is required to better understand the unique aspects of lung cancer in women and develop personalized treatment approaches

.

2.2 TARGETED THERAPIES FOR LUNG CANCER PATIENTS WITH ONCOGENIC DRIVER MOLECULAR ALTERATIONS

Tan, A. C., & Tan, D. S.'s paper [2] provides an overview of targeted therapy for patients with lung cancer who have oncogenic driver molecular alterations. Finding these molecular changes in lung cancer patients is critical for guiding treatment decisions and improving outcomes.

Lung adenocarcinomas, the most common form of lung cancer, contain a molecular defect that can be treated in roughly 50% of cases with a specific medication. EGFR, ALK, ROS1, BRAF, and RET are the oncogenic driver mutations that are most frequently discovered in lung cancer.

EGFR-mutant lung cancer patients can be treated with EGFR tyrosine kinase inhibitors (TKIs) such as erlotinib, gefitinib, or osimertinib. ALK-positive lung cancer patients can be treated with ALK inhibitors such as crizotinib, ceritinib, or alectinib. BRAF-mutant lung cancer patients can be treated with BRAF inhibitors such as dabrafenib or vemurafenib, either alone or in combination with MEK inhibitors.

There are some challenges associated with targeted therapies, including acquired resistance and the development of new molecular alterations. Resistance to targeted therapies can occur through various mechanisms, including the development of secondary mutations or the activation of alternative signaling pathways.

2.3 NON-CODING RNAs IN LUNG CANCER: EMERGING REGULATORS OF ANGIOGENESIS

Liao, Y. *et al.,* [3] review discusses the role of non-coding RNAs in the regulation of angiogenesis in lung cancer. Angiogenesis, the formation of new blood vessels, is a critical process in tumor growth and metastasis, and targeting angiogenesis has been shown to be an effective strategy for cancer treatment. Emerging evidence suggested that non-coding RNAs play an important role in the regulation of angiogenesis in lung cancer.

Non-coding RNAs are RNA molecules that do not encode proteins, but instead regulate gene expression through various mechanisms. Dysregulation of non-coding RNAs has been implicated in the development and progression of various cancers, including lung cancer.

miR-126 and miR-200b have been shown to inhibit angiogenesis by targeting vascular endothelial growth factor (VEGF) and other pro-angiogenic factors. On the other hand, lncRNA H19 has been shown to promote angiogenesis by regulating the expression of VEGF and other angiogenic factors.

miR-34a has been shown to sensitize lung cancer cells to chemotherapy, and a miR-34a mimic is currently being tested in clinical trials. In addition, targeting lncRNA H19 with antisense oligonucleotides has been shown to inhibit tumor growth and angiogenesis in preclinical models.

Targeting non-coding RNAs may be a promising strategy for the development of novel cancer therapies, and emphasize the need for further research in this area.

2.4 AN OVERVIEW OF THE ROLE OF RADIOTHERAPY IN THE TREATMENT OF SMALL CELL LUNG CANCER–A MAINSTAY OF TREATMENT OR A MODALITY IN DECLINE?

A study by Merie, R. et al. [4] aimed to provide a comprehensive overview of the role and evidence of radiation therapy in SCLC cure and remission. SCLC is a highly aggressive subtype of lung cancer that is often treated with a combination of chemotherapy and radiotherapy. However, the question remains whether radiotherapy remains the mainstay of SCLC treatment or whether its role is declining. It is the current standard of care for SCLC and usually involves a combination of chemotherapy and thoracic radiotherapy. The majority of patients will eventually progress to disease progression as a result of this treatment. SCLC is a radiosensitive tumor and the risk of radiotoxicity is a concern. Moreover, there is evidence that radiation therapy may not confer a significant survival benefit in patients with advanced disease.

Possible benefits of new therapeutic approaches such as Immunotherapy to treat SCLC. Immunotherapy has shown promising results in the treatment of other types of cancer, and studies are underway to evaluate the role of immunotherapy in SCLC.

Radiotherapy has been a mainstay of treatment for many years, but its role may change in the era of new therapeutic approaches such as immunotherapy. Further research is needed to better understand the optimal therapeutic approach for patients with SCLC.

2.5 DEEP LEARNING MODELS FOR CLASSIFYING CANCER AND COVID-19 LUNG DISEASES

Hişam, D., & Hişam, E. [5] paper proposed different deep learning-based models such as DarkNet-53 (the backbone of YOLO-v3), ResNet50, and VGG19 that were applied to classify CT images of patients having Corona Virus disease (COVID-19) or lung cancer. The study aimed to explore the effectiveness of deep learning models in accurately diagnosing these two diseases, which are known to share some similar radiological features.

The collection consists of 100 COVID-19 and 100 lung cancer CT scan images from freely accessible sources. Convolutional neural networks (CNNs)-based deep learning algorithms were utilised to extract features from the photos and categorise them as COVID-19 or cancer. Before inserting the images into the model, a number of pre-processing procedures, such as scaling and normalisation, were performed to guarantee that they were of the same size and quality.

As a result, DarkNet-53 overperformed other models by achieving 100% accuracy. While the accuracies for ResNet and VGG19 were 80% and 77% respectively.

Deep learning model could be integrated into clinical workflows to support radiologists and clinicians in making more accurate and timely diagnoses. Additionally, this study could serve as a basis for future research on the application of deep learning models in diagnosing other lung diseases.

2.6 SMALL-CELL LUNG CANCER. NATURE REVIEWS DISEASE PRIMERS

Rudin, C. M., *et al.*, [6] paper proposed a comprehensive review of small-cell lung cancer (SCLC), a highly aggressive subtype of lung cancer that accounts for approximately 15% of all lung cancer cases. The study covers various aspects of SCLC, including epidemiology, pathology, molecular biology, and treatment options.

There are some challenges associated with the diagnosis and treatment of SCLC. The aggressive nature of the disease and the lack of effective treatment options have contributed to poor outcomes for patients with SCLC. There have been some recent advances in the understanding of the molecular biology of SCLC, which have led to the development of new targeted therapies.

The study also discuss the molecular subtypes of SCLC, which are defined by specific genetic alterations that can be targeted with therapies. They highlight the importance of identifying these genetic alterations to guide treatment decisions.

There are various treatment options for SCLC, including chemotherapy, radiation therapy, and immunotherapy. But there are limitations of current treatment approaches and there is the need for novel treatment strategies to improve outcomes for SCLC patients.

Ongoing research efforts are made to better understand the molecular biology of SCLC and develop new targeted therapies. They emphasize the need for collaboration among researchers, clinicians, and industry partners to accelerate the development of new treatments for SCLC.

2.7 LUNG CANCER DISEASE DIAGNOSIS USING MACHINE LEARNING APPROACH

Mukherjee, S., & Bohra, S. U. [7] paper discusses the application of machine learning techniques for the diagnosis of lung cancer. Early detection and accurate diagnosis of lung cancer are crucial for improving patient outcomes.

The study begin by providing an overview of lung cancer and its diagnosis using traditional methods such as biopsy and imaging. It highlights the limitations of these methods, including their invasiveness and subjectivity, and suggest that machine learning techniques may offer a more accurate and efficient approach to lung cancer diagnosis.

Machine learning model is developed to diagnose lung cancer using computed tomography (CT) images. Dataset of 1000 CT images, including 500 images of patients with lung cancer and 500 images of patients without lung cancer is used.

CNN is used to train the machine learning model and evaluated its performance using various metrics, including sensitivity, specificity, and accuracy. Model achieved a high level of accuracy (over 90%) in diagnosing lung cancer.

The potential application of the machine learning model for improving the diagnosis of lung cancer, including reducing the need for invasive biopsy procedures and improving the accuracy of diagnosis. The model could be further refined and validated using larger datasets and additional clinical data.

This study demonstrates the feasibility of using a CNN model to accurately diagnose lung cancer using CT images and suggests that this approach could have important implications for improving patient outcomes.

2.8 TEXTURE ANALYSIS BASED FEATURE EXTRACTION AND CLASSIFICATION OF LUNG CANCER

Jena, S. R., George, T., & Ponraj, N. [8] paper presents a method for the classification of lung cancer using texture analysis-based feature extraction. Lung cancer is one of the leading causes of cancer-related deaths worldwide, and early detection is crucial for improving patient outcomes.

The study begin by providing an overview of lung cancer and its diagnosis using image techniques such as computed tomography (CT) scans. Texture analysis, which involves the quantification of image patterns, can provide valuable information for the diagnosis of lung cancer.

For texture analysis-based feature extraction and classification of lung cancer, Dataset of 100 CT images, including 50 images of patients with lung cancer and 50 images of patients without lung cancer was used.

Various texture analysis techniques, including gray-level co-occurrence matrix (GLCM) and gray-level run-length matrix (GLRLM), to extract features from the CT images were used. Support vector machine (SVM) classifier was used to classify the CT images into lung cancer and non-lung cancer categories.

The above method was evaluated on various metrics, including sensitivity, specificity, and accuracy. It achieved a high level of accuracy (over 90%) in classifying the CT images.

Potential application of the method is the early detection and diagnosis of lung cancer. Texture analysis-based feature extraction could provide valuable information for identifying early-stage lung cancer, which can be difficult to detect using traditional methods.

This study demonstrates the feasibility of using a SVM classifier to accurately classify CT images and suggests that this approach could have important implications for improving the early detection and diagnosis of lung cancer.

2.9 GLOBAL EPIDEMIOLOGY OF LUNG CANCER

Barta, J. A., Powell, C. A., & Wisnivesky, J. P. [9] study was to review the evidence on lung cancer epidemiology, including data of international scope with comparisons of economically, socially, and biologically different patient groups. Lung cancer is the leading cause of cancer-related deaths worldwide, with an estimated 1.76 million deaths in 2018.

In industrialised nations, lung cancer rates for women have increased or stabilised, trailing long-declining rates for males due to shifting societal and cultural smoking patterns. Emerging economies have a wide variety of smoking habits and cancer incidence when compared to developed countries, but they also regularly experience environmental exposure risks, notably from severe air pollution. 85% of lung cancer cases are directly related to tobacco use, which accounts for the disease's incidence.

Adenocarcinoma histology is becoming more common, and new research has revealed clinical, radiologic, and pathologic associations that have advanced our understanding of molecular profiling and targeted therapy. Furthermore, new knowledge about the benefits of lung cancer screening has encouraged efforts to identify high-risk smokers and the development of prediction systems.

This study also included the epidemiologic traits of particular populations, such as women and nonsmokers. The incidence and death of lung cancer vary widely among countries due to changing smoking habits. Because smoking rates have declined in wealthy countries and because molecular profiling of tumours has provided new knowledge, the epidemiology of lung cancer will alter as a result of the introduction of novel risk factors and disease characteristics.

Tobacco control measures, such as increasing taxes on tobacco products and implementing smoke-free policies, can help to reduce the incidence of lung cancer. This study also discuss about the importance of screening programs for individuals at high risk of developing lung cancer, such as current or former smokers.

2.10 LUNG CANCER PREDICTION AND DETECTION USING IMAGE PROCESSING MECHANISMS: AN OVERVIEW

Ahmed, B. T. [10] study aims to review the most well-known Image Processing Mechanisms for Lung-Cancer Detection and Prediction. The comparison based on the Image Processing Mechanisms, accuracy, and classifier used in each reviewed research paper. Multi layer perceptron (MLP) gained a higher accuracy than the others followed by Logistic Regression (LR) and Decision Tree (D-Tree), which were (99.04%), (98.1%), and (93.62%), respectively. But, C4.5 obtained (86.7%) accuracy followed by the Genetic Algorithm that attained approximately (84.8%) accuracy.

This study highlights the potential of image processing mechanisms for improving the accuracy of lung cancer detection and reducing the number of false positives and false negatives. Further research is needed to validate the effectiveness of these approaches and to develop more accurate and reliable methods for detecting and predicting lung cancer using medical imaging.

2.11 PROPHYLACTIC CRANIAL IRRADIATION IN STAGE IV SMALL CELL LUNG CANCER

13 European experts were listed by the International Association for the Study of Lung Cancer (IASLC) and the European Society for Therapeutic Radiation Oncology (ESTRO) for the Putora, P. M., et al., [11] study. Patients with SCLC can avoid brain metastases, a common adverse effect, by receiving low-dose radiation to the brain as part of the PCI process.

The approaches for selecting PCI in stage IV SCLC were gathered. Decision trees were used to represent these tactics. The consensus was examined using the objective consensus methodology. The recommendation for PCI was based on several factors, including the patient's good health, youth, and satisfactory response to chemotherapy.

PCI was recommended by the majority of experts for non-elderly fit patients who had at least a partial response (PR) to chemotherapy (for complete remission (CR) 85% of radiation oncologists and 69% of medical oncologists, for PR: 85% of radiation oncologists and 54% of medical oncologists). For patients with stable disease after chemotherapy, PCI was recommended by 6 out of 13 (46%) radiation oncologists and only 3 out of 13 medical oncologists (23%). For elderly fit patients with CR, a majority recommended PCI (62%) and no consensus was reached for patients with PR.

Further research is required to determine the optimal selection of patients for PCI in stage IV SCLC, and to evaluate the potential benefits and risks of this treatment technique. The use of PCI in patients with stage IV SCLC should be based on careful consideration of the patient's individual situation, and should be guided by the latest evidence-based guidelines and expert opinion.

2.12 DEVELOPMENT OF A BREATH DETECTION METHOD BASED E-NOSE SYSTEM FOR LUNG CANCER IDENTIFICATION

Wong, D. M., *et al.,* [12] paper focused on the method of lung cancer identification by breath. The purpose of this breath detection system was to help physicians to quickly screen for rapid screening lung cancer. They used KNN and SVM with leave-one-out cross validation to analyze. PCA-KNN accuracy was 84.4%. The LDA-KNN accuracy was 75.5%. The PCA-SVM linear, polynomial, and rbf kernel type accuracy were 73.3%, 73.3%, and 73.3%, respectively. However, system achieved great results at about 84.4% accuracy for PCA-KNN classification.

There are some limitations of the study, including a relatively small sample size and the fact that the study only included patients with non-small cell lung cancer. Further studies with larger sample sizes and including patients with other types of lung cancer are needed to validate the effectiveness of the E-nose system for lung cancer diagnosis.

2.13 MULTI-STAGE LUNG CANCER DETECTION AND PREDICTION USING MULTI-CLASS SVM CLASSIFIER

In the article by Alam, J., Alam, S., and Hossan [13], an efficient method for diagnosing and predicting lung cancer using a multi-class SVM (Support Vector Machine) classifier was suggested. Feature extraction, pre-processing, and classification are the three phases of the proposed system. To increase contrast and reduce noise, pre-processing was applied to the lung CT scan images. During the feature extraction stage from the pre-processed images, a total of 20 features were obtained using the gray-level co-occurrence matrix (GLCM) method. Finally, the retrieved features were classified using a multi-class SVM classifier.

The proposed system was evaluated using a dataset of 200 lung CT scan images, which were divided into three classes: normal, benign, and malignant. The performance of the system was evaluated based on several evaluation metrics, including accuracy, sensitivity, specificity, precision, and F1 score. The results showed that the proposed system achieved an overall accuracy of 95%, sensitivity of 97%, specificity of 94%, precision of 96%, and F1 score of 96%.

The study concluded that the proposed multi-stage lung cancer detection and prediction system using a multi-class SVM classifier can be an effective tool for lung cancer diagnosis and prediction. However, the study also highlighted the need for further evaluation and testing of the proposed system on larger datasets and diverse populations.

2.14 MULTIPLE PRIMARY LUNG CANCER

Romaszko, A. M., & Doboszyńska, A. [14] paper was to discuss a comprehensive overview of multiple primary lung cancer (MPLC), which is defined as the presence of two or more lung cancer lesions in the same patient.

The study highlight the importance of differentiating MPLC from intrapulmonary metastasis or recurrence of a single primary lung cancer, as the management and prognosis of these entities are different. The study also discusses the prevalence of MPLC, which ranges from 0.5% to 20% of all lung cancer cases, depending on the diagnostic criteria used.

There are various risk factors for developing MPLC, which include smoking, exposure to environmental carcinogens, and genetic factors. There are various diagnostic methods used for detecting MPLC, including imaging techniques such as computed tomography (CT) and positron emission tomography (PET), as well as histological analysis of tissue samples.

Management of MPLC depends on the size, location, and histology of the lesions. Surgical resection remains the mainstay of treatment for early-stage MPLC, while systemic therapy is used for advanced disease. Radiation therapy and targeted therapy can be used in the management of MPLC.

2.15 THE EPIDEMIOLOGY OF LUNG CANCER

Determining the frequency, potential causes, geographic distribution, and potential management of lung cancer globally is the aim of the de Groot, P. M., et al., [15] article. Lung cancer remained the leading cause of cancer-related deaths worldwide, with an estimated 1.8 million deaths from the condition in 2018. Lung cancer incidence and mortality are dropping in the US as a result of years of public education campaigns and tobacco control regulations, but they are increasing abroad as the tobacco epidemic starts to spread across populations in a number of emerging countries.

Lung cancer risk factors other than personal cigarette smoking include infection, radon exposure from dwellings, occupational exposure, passive smoke inhalation, and genetic susceptibility.

Since tobacco use is the primary cause of lung cancer in 80–90% of cases, smoking cessation programmes are crucial for lowering lung cancer incidence and mortality.

Currently, underrepresented communities and persons with low socioeconomic level are bearing the brunt of the disease burden. Due to the lack of long-term and scant short-term safety data, the recent legalisation of marijuana for recreational use in a number of US states and the rising popularity of commercially available electronic nicotine delivery systems (ENDS) raise concerns for public health.

The study also describes the current state of lung cancer screening and early detection, highlighting the benefits and limitations of low-dose computed tomography (LDCT) screening and the need for improved biomarkers and imaging techniques. Finally, It discussed the latest advances in lung cancer treatment, including targeted therapies and immunotherapy, and the importance of a multidisciplinary approach to lung cancer management.

Chapter 3

METHODOLOGY

The proposed method involved using various machine learning algorithms and deep learning models for image segmentation of lung histopathological images and predicting lung cancer. The algorithmic models used in this study were SVM, random forest, KNN, logistic regression, naive Bayes, CatBoost, XGBoost, and decision tree. Additionally, deep neural network models such as ANN, CNN, Deep-CNN,Inception-V3, EfficientNet-B7, ResNet50, and VGG16 were also used.

The initial step of the proposed methodology is data preparation, which involved scaling down all images to the common size of 768x768. The length of time it takes for our model to train as well as the amount of memory required could both be decreased by reducing the images. Small-sized picture data has the advantage of allowing for the inclusion of additional photos during training without burdening the model's memory or extending training time. The amount of pixel data in one image and count can be traded off well to determine the number of photos that can be used for training in a limited computational environment.

The Python tool's Jupyter platform was utilised for the data analysis process. Using the free and open-source Jupyter online tool, documents with live code, equations, images, and text may be created and shared. It can be applied to statistical modelling, data visualisation, data purification, and data transformation, among other things.

Before splitting the dataset into training data and testing data, it is important to preprocess the categorical variables, which represent non-numeric data such as colors, labels, or categories. One popular technique for preprocessing categorical variables is one-hot encoding, which transforms each categorical variable into a set of binary variables that indicate the presence or absence of a particular category.

After one-encoding step, the dataset is split into 80% training data and 20% testing data. Now these images are provided to various ML algorithms such as logistic regression, KNN, SVM, Naïve Bayes, random forest, decision tree, cat boost and XGB classifier and deep neural network architecture such as CNN, ANN, ResNet50 , VGG16, Inception, Efficient Net. Deep neural networks have proven to yield better accuracy when dealing with large volumes of dataset, and many researchers tend to use them as de-facto standards. A typical architecture of neural network consists of multiple blocks with three kinds of layers: convolution, pooling, and fully connected layers.

After the execution of above step, It displays accuracy of prediction whether the input lung image contains cancer or not. This information can be used to aid in diagnosis, screening, and treatment planning.

Future directions for the research could involve exploring the use of other imaging modalities for lung cancer detection, such as magnetic resonance imaging (MRI), computed tomography (CT), or positron emission tomography (PET). These imaging techniques offer unique advantages and may provide complementary information to histopathological images. Additionally, the development of new deep neural network models or the refinement of existing models could lead to improved accuracy and efficiency in lung cancer detection and diagnosis. Finally, clinical validation studies and the incorporation of the proposed ML algorithms into clinical decision support systems could further validate the feasibility and utility of this approach in real-world clinical settings.

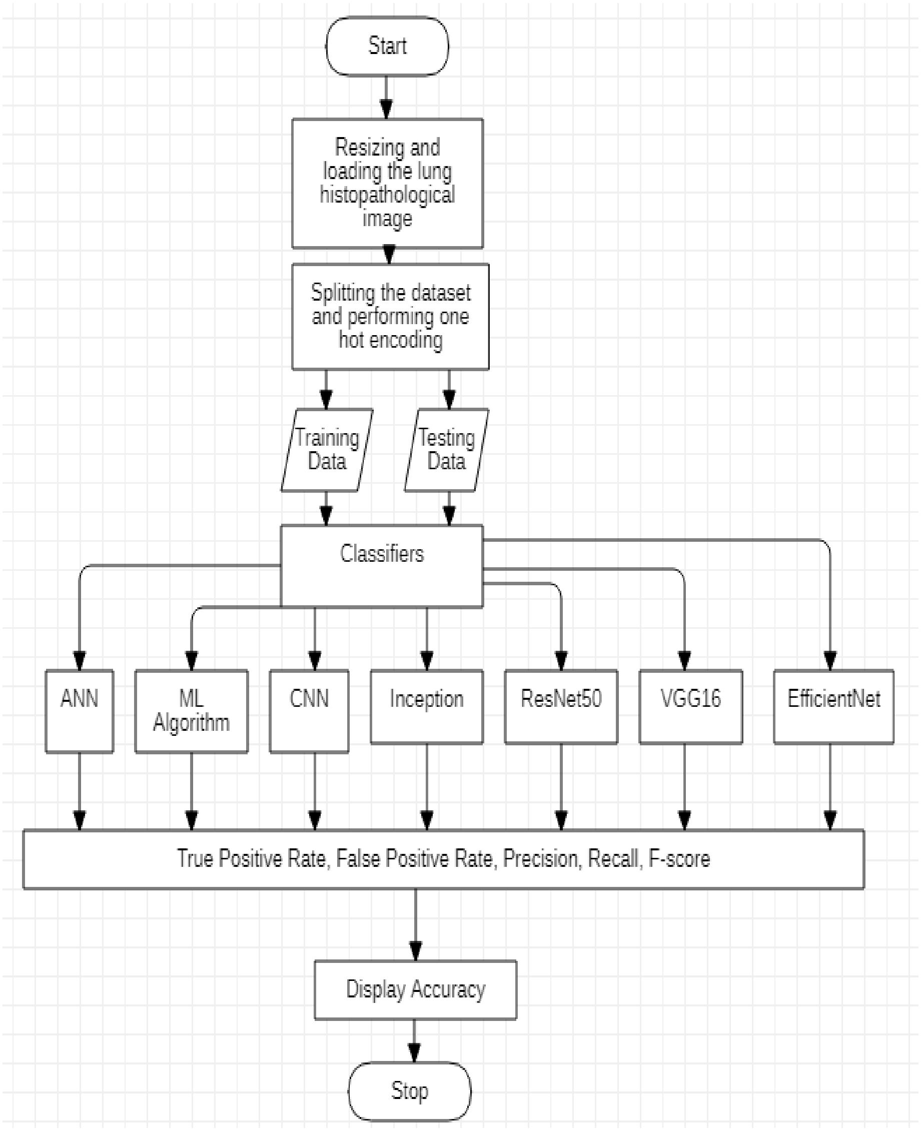

Start
Resizing and loading the lung histopathological image
Splitting the dataset and performing one hot encoding
Training Data
Testing Data
Classifiers
ANN
ML Algorithm
CNN
Inception
ResNet50
VGG16
EfficientNet
True Positive Rate, False Positive Rate, Precision, Recall, F-score
Display Accuracy
Stop

FIGURE 3.1: BLOCK DIAGRAM OF THE MODEL

3.1 DATASET

Dataset of the proposed work consists of lung tissue biopsy samples taken from patients suspected to have lung cancer. The dataset was obtained from multiple sources, including publicly available databases and hospitals. The images were captured using various magnifications and imaging modalities, resulting in a diverse range of image resolutions and quality.

The dataset comprises 5000 images for each of the three classes- benign lung tissue, lung adenocarcinomas and lung squamous cell carcinomas. Equal number of images for each class are used to avoid the problem of class imbalance. The images were generated from an original sample of HIPAA compliant and validated sources. All images are 768 x 768 pixels in size and are in jpeg file format. An image of each class obtained from the dataset is shown in the FIGURE 3.2.

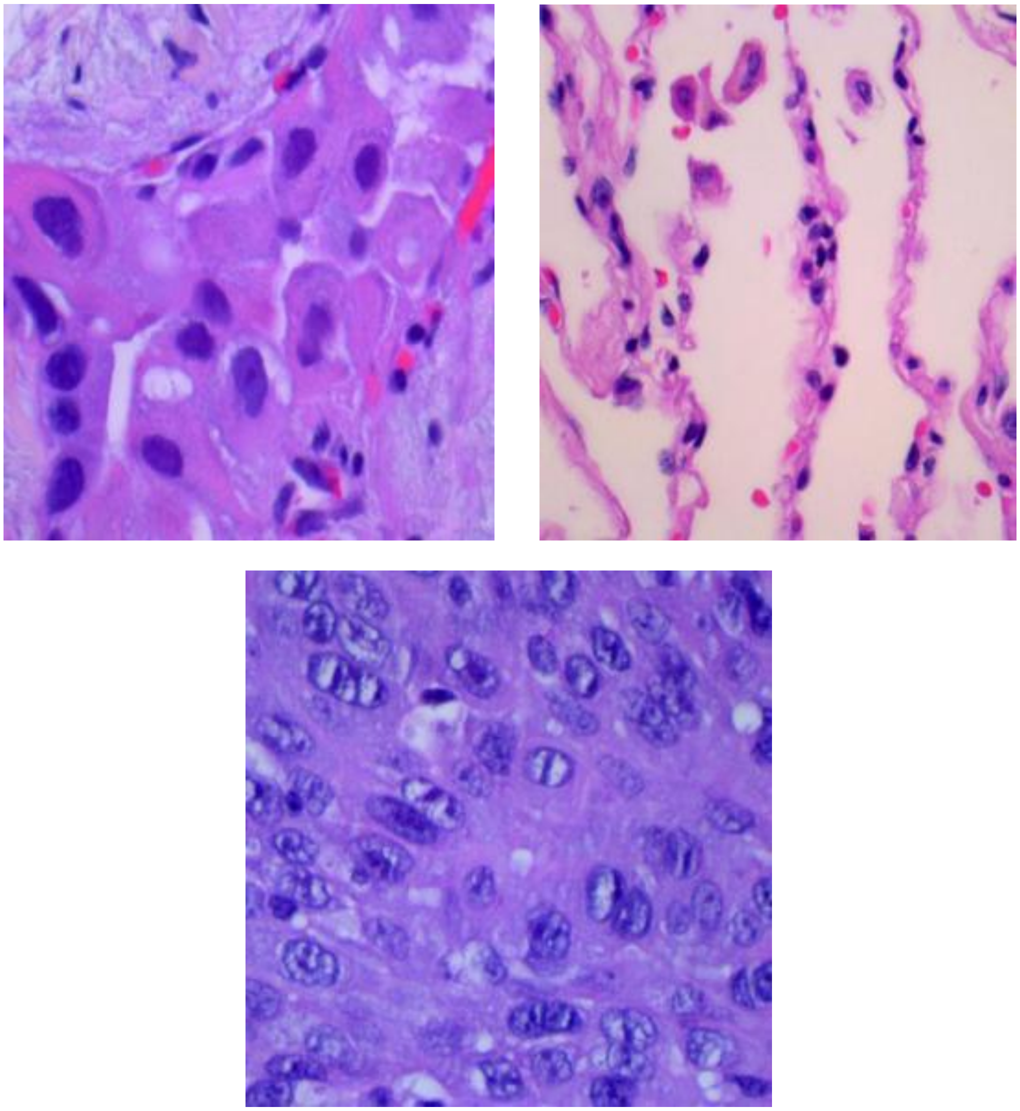


FIGURE 3.2: ADENOCARCINOMAS CELL CARCINOMA, BENIGN LUNG TISSUE AND LUNG SQUAMOUS CELL CARCINOMA

3.2 ALGORITHMS

The algorithms that are used for detecting Lung cancer includes:

1. **Support Vector Machine:** Support Vector Machine, sometimes referred to as SVM [16], is a supervised machine learning method used for classification, regression, and outlier detection. SVM is a powerful technique having uses in bioinformatics, pattern recognition, text classification, and picture classification among others. Both linear and non-linear data can be handled by it. SVM determines the optimum hyperplane in a high-dimensional space to separate the various classes. SVM can also handle non-linear data by translating the input features into a higher-dimensional space where the data can be divided by a hyper plane.

   $$w\text{^}T\, x + b=0 \qquad (1)$$

   where x is a vector in the input space, w is the weight vector and b is the bias.

2. **K-Nearest Neighbor Classifier:** The KNN algorithm [17], sometimes known as the k-nearest neighbours algorithm, is a non-parametric, supervised learning classifier that makes predictions or classifications about how a single data point will be grouped. Although it can be used to solve classification or regression problems, it is frequently used as a classification technique because it is predicated on the notion that similar points can be found nearby.

   $$\mathrm{dist}(x,x')\geq \max(x'',y'')\in Sx\mathrm{dist}(x,x''), \qquad (2)$$
   where x is the test point, Sx is the set of the k nearest neighbors.

   $$h(x)=\mathrm{mode}(\{y'':(x'',y'')\in Sx\}) \qquad (3)$$
   where h() is the function returning the most common label in Sx and mod() means to select the label of the highest occurrence.

3. **Logistic Regression**: Machine learning models are developed when the dependant variable is binary using a statistical technique called logistic regression [18]. Data and the relationship between a dependent variable and one or more independent variables are described using logistic regression. Healthcare organisations can precisely pinpoint at-risk individuals who need a more individualised behavioural health plan to help them improve their daily health practises using logistic regression. The result could be lower hospital expenses and better patient health.

$$g(E(y)) = \alpha + \beta x1 + \gamma x2 \qquad (4)$$

where g() is the link function, E(y) is the expectation of target variable and α + βx1 + γx2 is the link predictor(α ,β ,γ to be predicted).

4. **Naïve Bayes:** Naive Bayes [19] is based on Bayes' theorem, which is a formula that describes the probability of an event based on prior knowledge of conditions that might be related to the event. Naive Bayes can be used for both binary and multi-class classification problems, and it is particularly well-suited for problems with a large number of input features. It is also relatively fast to train and make predictions, and it can work well even with small amounts of training data.

   $$P(A|B) = P(B/A)*P(A)/P(B) \qquad (5)$$

   where P(A|B) is posterior probability, P(B|A) is the likelihood of A given a fixed B, P(A) is the probability of A and P(B) is the probability of B.

5. **CatBoost:** CatBoost [20], sometimes known as categorical boosting, is a gradient boosting method for decision trees that Yandex created. Gradient boosting, an ensemble machine learning technique, is commonly used to solve classification and regression problems. In essence, it transforms a collection of multiple weak learners into a strong one. It works well with heterogeneous data. In addition to regression and classification, CatBoost can be used for ranking, recommendation systems, forecasting, and even personal assistants.

   $$f(x) = b0 + \Sigma b(i)t(x) \qquad (6)$$

   where f(x) is the prediction of the model for the input data point x, bo is the base prediction (usually the mean or median of the target variable), b(i) is the prediction of the i-th tree in the model and t(x) is the terminal node where x falss in the i-th tree.

6. **Random Forest:** To boost the projected accuracy of the input dataset, a classifier known as random forest [21] employs numerous decision trees on various subsets of the input dataset and averages the results. It is built on the concept of ensemble learning, which is the practise of combining different classifiers to solve a difficult problem and improve the performance of the model. In the field of healthcare, random forests provide a number of early diagnosis possibilities that are not only more cost-effective than neural networks but also address the moral conundrum (decision-making problem) that neural networks face.

   Classification,

   $$f(x) = mode(T1(x), T2(x), ..., Tn(x)) \qquad (7)$$

   where f(x) is the predicted class label for x, and mode() is the function that returns the most common class label among the decision trees.

Regression,

$$f(x) = (1/n)\Sigma(Ti(x)) \qquad (8)$$

where f(x) is the predicted numerical value for x, and Σ() is the summation function.

7. **XGBoost classifier:** Extreme Gradient Boosting, also referred to as XGBoost [22], is a scalable distributed machine learning system for gradient-boosted decision trees (GBDT). Because the library is parallelizable, the primary algorithm can execute on groups of GPUs or even over a network of computers. This makes it possible to train ML problems at high performance utilising hundreds of millions of training cases. It provides parallel tree boosting and is the best machine learning package for regression, classification, and ranking problems.

$$f(x) = \Sigma w(i)t(i)(x) \qquad (9)$$

where f(x) is the predicted value for the input data point x, w(i) is the weight of the i-th tree in the model, t(i)(x) is the predicted value of the i-th tree for the input data point x.

8. **Decision Tree:** The decision tree [23] algorithm is a machine learning method used for classification and regression tasks. It works by building a decision tree model, which resembles a flowchart in how it shows a framework of options and their results. Since the decision tree approach uses supervised learning, it requires a labelled dataset for training.

Classification,

$$f(x) = argmax(\Sigma wiI(xi,j = c)fj) \qquad (10)$$

where f(x) is the projected class label for the input data point x and wi is the weight corresponding to the i-th class label.The leaf node's fj value is connected to the i-the class label at the leaf node, and xi,j is the jth feature of the input data point x. c is the value of the jth feature that the node checks. The indicator function, or I(), gives a result of 1 when the condition is true and 0 when it is false.

Regression,

$$f(x) = \Sigma wiI(x <= tj)fj \qquad (11)$$

where f(x) is the predicted numerical value for the input data point x, wi is the weight associated with the i-th leaf node, I() is the indicator function that returns 1 if the condition is true and 0 otherwise, x is the input data point, tj is the threshold value for the j-th feature at the node, and fj is the value associated with the i-th leaf node.

9. **Artificial Neural Network (ANN):** An Artificial Neural Network (ANN) [24] is a computational model inspired by the biological neural networks of the human brain. ANNs are used to solve complex problems that are difficult or impossible to solve using traditional computing techniques. ANNs consist of a large number of interconnected processing nodes or artificial neurons that work together to perform a specific task.

   In image segmentation, ANNs are used to partition an image into regions or segments based on their visual characteristics, such as color, texture, and shape. ANNs are trained on large datasets of labeled images, where the desired output for each input image is a binary mask indicating the segmentation boundaries. The network learns to map input image features to corresponding output masks through a process of iterative optimization, typically using backpropagation and gradient descent algorithms. Image of the model is shown in FIGURE 3.3.

   In summary, ANNs have become a powerful tool for image segmentation tasks, enabling automated and accurate segmentation of complex images. They offer a promising approach for a range of applications in medical imaging, remote sensing, robotics, and more.

   The computation of the output of a neuron i in layer j is given by,

$$z(i,j) = \Sigma w(i,k)x(k,j-1) + b(i,j) \qquad (12)$$

   where z(i,j) is the weighted sum of the inputs to neuron i in layer j, w(i,k) is the weight of the connection between neuron i in layer j and neuron k in layer j-1, x(k,j-1) is the output of neuron k in layer j-1, and b(i,j) is the bias term of neuron i in layer j.

   The gradient of the loss function with respect to the weight w(i,k) is given by,

$$\partial L/\partial w(i,k) = \partial L/\partial z(i,j) * \partial z(i,j)/\partial w(i,k) \qquad (13)$$

   where ∂L/∂z(i,j) is the derivative of the loss function with respect to the weighted sum z(i,j), and ∂z(i,j)/∂w(i,k) is the derivative of the weighted sum z(i,j) with respect to the weight w(i,k).

   The gradient of the loss function with respect to the bias b(i,j) is given by,

$$\partial\ \partial L/\partial b(i,j) = \partial L/\partial z(i,j) * \partial z(i,j)/\partial b(i,j) \qquad (14)$$

   where ∂L/∂z(i,j) is the derivative of the loss function with respect to the weighted sum z(i,j), and ∂z(i,j)/∂b(i,j) is the derivative of the weighted sum z(i,j) with respect to the bias b(i,j).

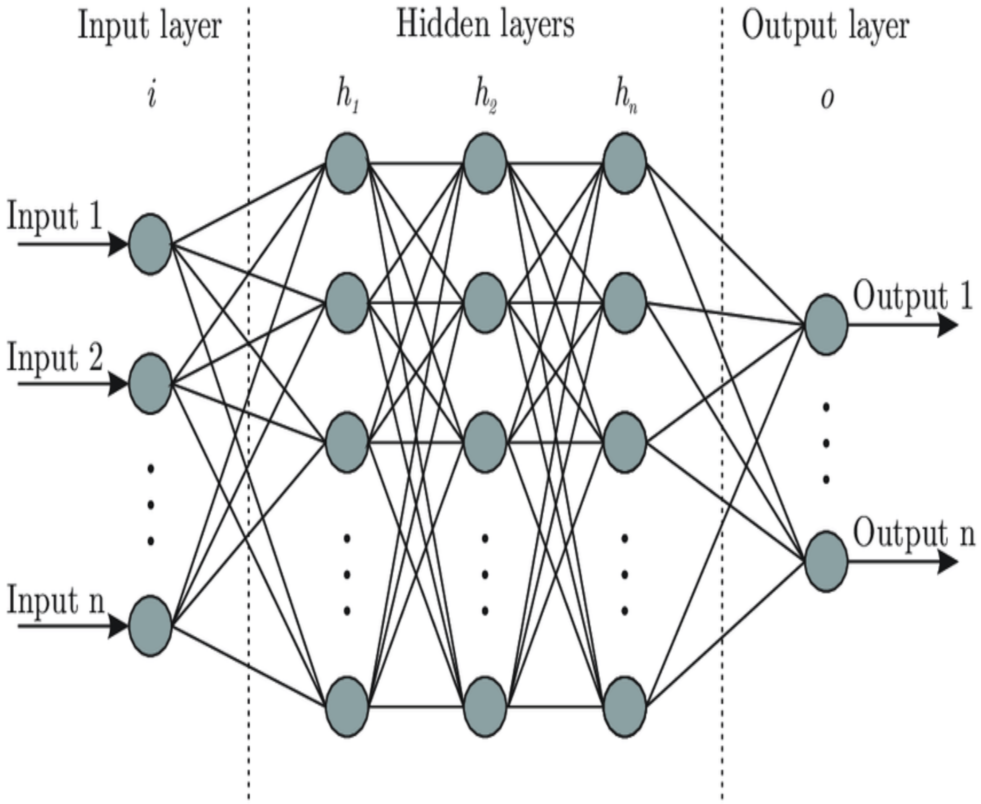


FIGURE 3.3: ANN MODEL

10. **Convolutional Neural Network:** Convolutional neural network (CNN) is an illustration of a deep learning technique that is particularly useful for image processing and recognition applications [25]. Among the layers that make up this structure are convolutional layers, pooling layers, and totally linked layers.

    In CNN-based image segmentation, the input image is processed by a series of convolutional layers, each of which extracts a particular set of features from the input image. Then, in order to reduce the dimensionality of the feature map and enhance the representation of the features, these features are passed via additional layers like pooling and activation layers. Following receipt of the output from the last CNN layer, the segmentation layer creates a binary mask illustrating the segmentation of the image. The segmentation layer frequently produces a probability map using a softmax activation function or a sigmoid activation function, which is thresholded to create the final binary mask. The model's image is shown in FIGURE 3.4.

    CNN-based image segmentation has shown to be highly effective for a range of applications, including medical image analysis, remote sensing, and autonomous driving. It has enabled automated and accurate segmentation of complex structures and regions in images that were previously difficult to segment manually.

Convolution,

$$feature\_map(i,j) = \Sigma\Sigma\Sigma\, filter(m,n,k) * input\_image(i+m-1,j+n-1,k) \tag{15}$$

where feature_map(i,j) is the value of the feature map at position (i,j), filter(m,n,k) is the value of the filter at position (m,n,k), and input_image(i+m-1,j+n-1,k) is the value of the input image at position (i+m-1,j+n-1,k).

Pooling,

$$max\_pooled\_feature\_map(i,j)= max(feature\_map(i+p,j+q)) \tag{16}$$

where max_pooled_feature_map(i,j) is the value of the max-pooled feature map at position (i,j), feature_map(i+p,j+q) is the value of the feature map within the sliding window centered at position (i,j), and the size of the sliding window is specified by the pooling size.

Activation Function,

$$output(i,j) = ReLU(max\_pooled\_feature\_map(i,j)) \tag{17}$$

where output(i,j) is the output of the layer at position (i,j) and Rectified Linear Unit (ReLU), which sets all negative values to zero and leaves positive values unchanged.

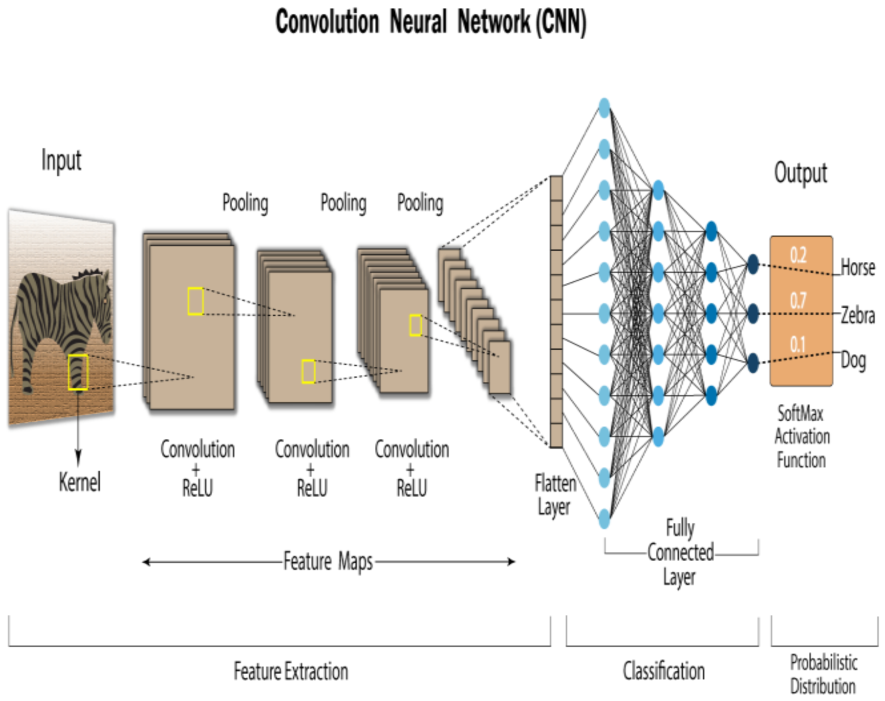


FIGURE 3.4: CNN MODEL

11. **Inception:** An inception network [26] is a deep neural network with an architectural design that consists of repeating components referred to as Inception modules. Image of the model is showin in FIGURE 3.5.
    In image segmentation, Inception has been used to identify regions of interest within an image by segmenting it into multiple regions. The network is able to identify different features and textures within the image, allowing it to distinguish between different objects and identify boundaries between them.
    Inception has also been used in medical image segmentation applications, such as segmenting the liver from CT scans or the brain from MRI scans. By using Inception to accurately identify the edges of organs or structures within an image, medical professionals can more easily analyze and diagnose potential issues.

    1x1 Convolution,

    $$output\_1x1 = conv\_1x1(input, weights) \qquad (18)$$

    where output_1x1 is the output feature map of the 1x1 convolution, input is the input feature map, conv_1x1 is the convolution operation with 1x1 filters, and weights are the learnable parameters of the convolution.

    3x3 Convolution,

    $$output\_3x3 = conv\_3x3(input, weights) \qquad (19)$$

    where output_3x3 is the output feature map of the 3x3 convolution, input is the input feature map, conv_3x3 is the convolution operation with 3x3 filters, and weights are the learnable parameters of the convolution.

    5x5 Convolution,

    $$output\_5x5 = conv\_5x5(input, weights) \qquad (20)$$

    where output_5x5 is the output feature map of the 5x5 convolution, input is the input feature map, conv_5x5 is the convolution operation with 5x5 filters, and weights are the learnable parameters of the convolution.

    Max Pooling,

    $$output\_maxpool = max\_pool(input) \qquad (21)$$

    where output_maxpool is the output feature map of the max pooling operation, and input is the input feature map.

Concatenation,

$$output\_inception = concatenate([output\_1x1, output\_3x3, output\_5x5, output\_maxpool]) \quad (22)$$

where output_inception is the output feature map of the inception module, and concatenate is the operation that concatenates the input feature maps along the channel dimension.

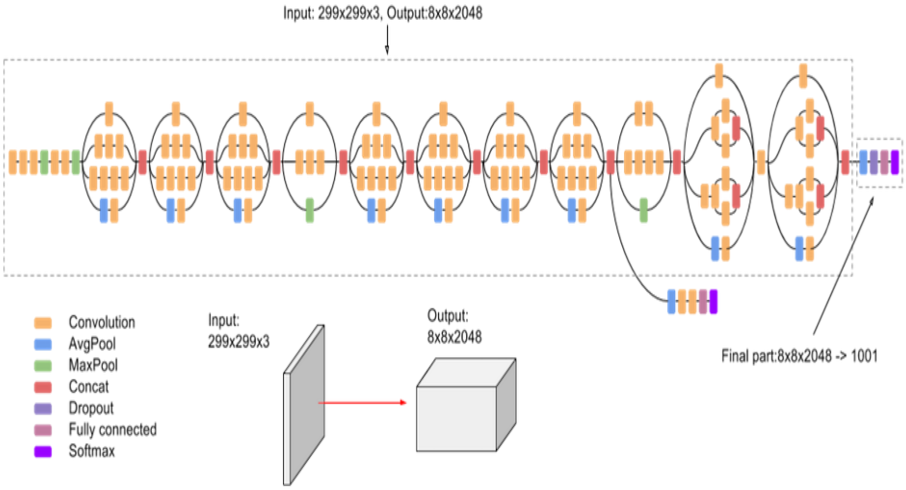


FIGURE 3.5: INCEPTION-V3 MODEL

12. **ResNet50:** ResNet [27] stands for Residual Network and is a specific type of convolutional neural network (CNN) introduced in the 2015 paper “Deep Residual Learning for Image Recognition” by He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. CNNs are commonly used to power computer vision applications.

    The 50-layer convolutional neural network known as ResNet-50 consists of 48 convolutional layers, one MaxPool layer, one average pool layer, and one layer. Residual neural networks (RNNs) are artificial neural networks (ANNs) that construct networks using residual blocks.

    In image segmentation, ResNet50 is used to extract features from input images and then perform segmentation based on those features. This involves training the network on a large dataset of labeled images and using the learned features to accurately segment new images. Image of the model is shown in FIGURE 3.6.

    One advantage of ResNet50 is its ability to learn hierarchical features, which allows for better segmentation of complex images with multiple objects and overlapping structures. Additionally, the residual connections in the network help to alleviate the vanishing gradient problem, allowing for more efficient training of deep networks.

    Overall, ResNet50 is a powerful tool for image segmentation, particularly in medical imaging applications where accuracy is critical. Its ability to learn hierarchical features and overcome the challenges of training deep networks make it an effective choice for complex segmentation tasks.

First Convolution,

$$output\_conv1 = conv(input, weights1) \qquad (23)$$

where output_conv1 is the output feature map of the first convolutional layer, input is the input feature map, conv is the convolution operation with learnable weights1, and weights1 are the learnable parameters of the convolution.

Second Convolution,

$$output\_conv2 = conv(output\_conv1, weights2) \qquad (24)$$

where output_conv2 is the output feature map of the second convolutional layer, output_conv1 is the input feature map, conv is the convolution operation with learnable weights2, and weights2 are the learnable parameters of the convolution.

Shortcut connection,

$$output\_shortcut = input + output\_conv2 \qquad (25)$$

where output_shortcut is the output of the shortcut connection, and input is the input feature map.

Activation,

$$output\_resblock = activation(output\_shortcut) \qquad (26)$$

where output_resblock is the output of the residual block, and activation is the activation function, such as ReLU or sigmoid.

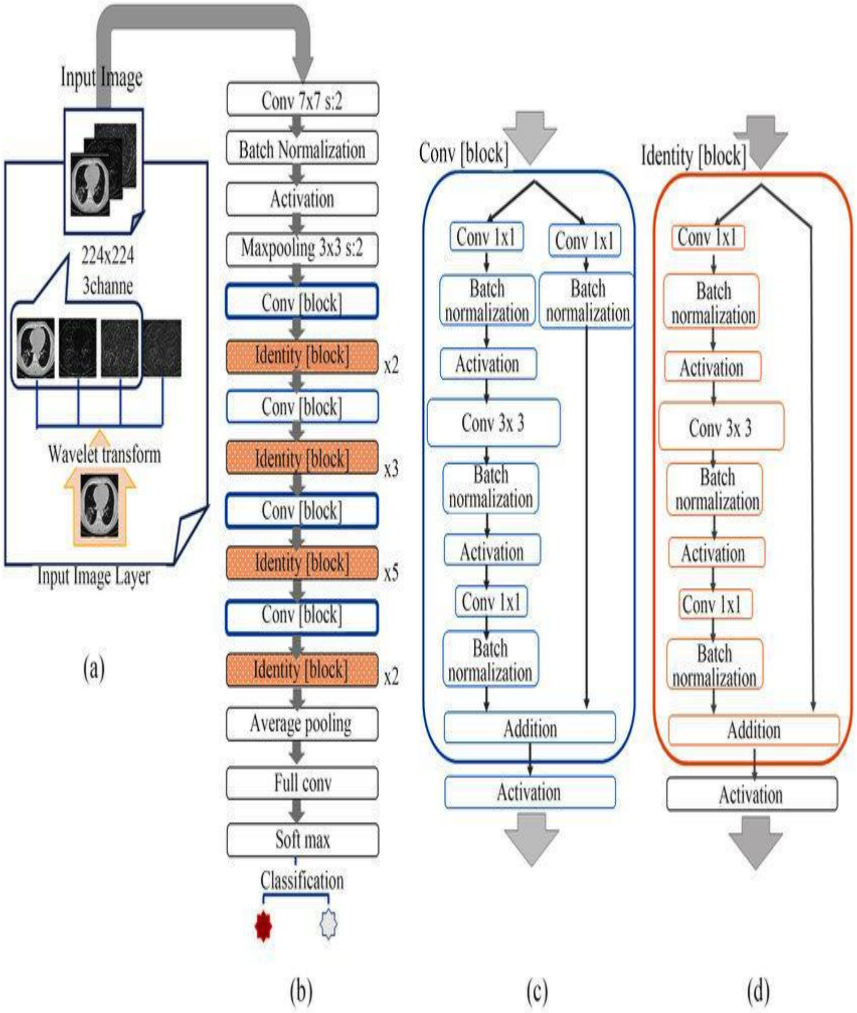


FIGURE 3.6: RESNET50 MODEL

13. **VGG16:** VGG16 [28] is a type of CNN (Convolutional Neural Network) that is considered to be one of the best computer vision models to date. The creators of this model evaluated the networks and increased the depth using an architecture with very small (3 × 3) convolution filters, which showed a significant improvement on the prior-art configurations. They pushed the depth to 16–19 weight layers making it approx — 138 trainable parameters. Image of the model is shown in FIGURE 3.7[29].

    In image segmentation, VGG16 can be used as a feature extractor by removing the last fully connected layers and using the output of the final convolutional layer as input for a segmentation network. This approach is known as transfer learning and can significantly reduce the amount of data and training time required for the segmentation task.

    VGG16 has been used in various medical imaging applications, including lung cancer detection, where it has achieved high accuracy in differentiating between benign and malignant lung nodules.

Convolution Layers,

$$output\_i = relu(conv\_i(output\_i\text{-}1)) \tag{27}$$

where output_i is the output of the i-th convolutional layer, conv_i is the convolution operation with learnable parameters, relu is the ReLU activation function, and output_i-1 is the output of the (i-1)-th convolutional layer.

Max pooling Layers,

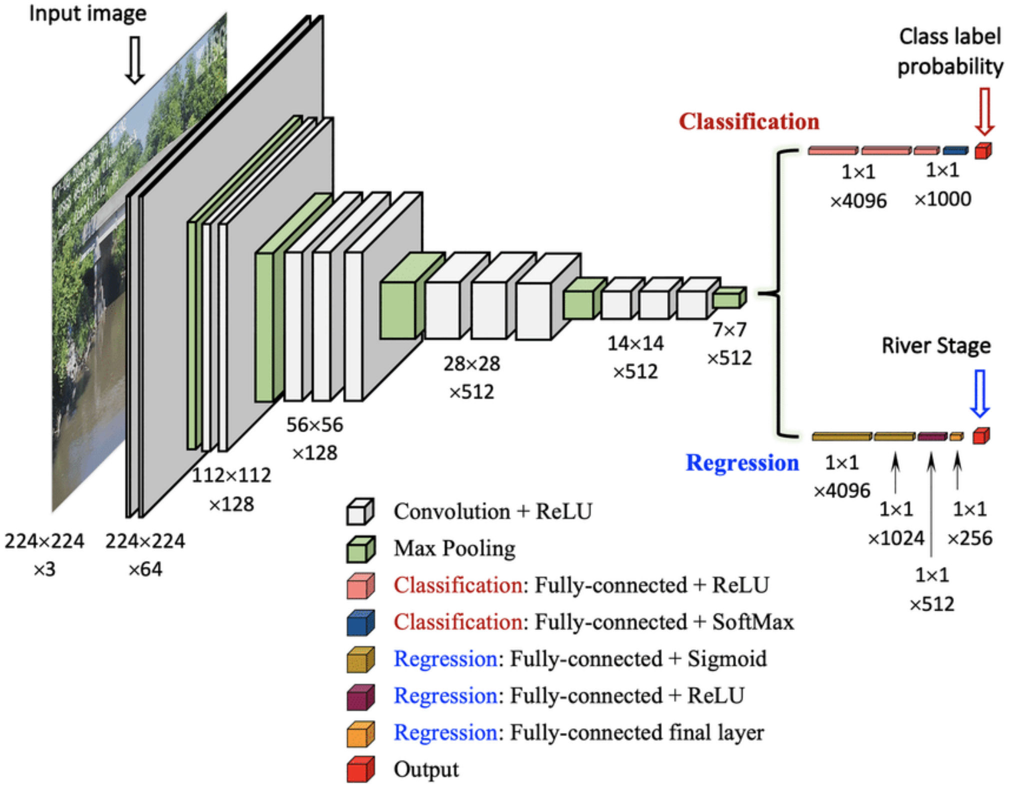
$$output\_i = max\_pool(output\_i\text{-}1) \tag{28}$$

where output_i is the output of the i-th max pooling layer, max_pool is the max pooling operation, and output_i-1 is the output of the (i-1)-th convolutional layer.

Fully connected Layers,

$$output = softmax(relu(fc(output\_i\text{-}1))) \tag{29}$$

where output is the final output of the network, fc is the fully connected operation with learnable weights, relu is the ReLU activation function, and softmax is the softmax function that produces a probability distribution over the possible classes.



,

FIGURE 3.7: VGG16 MODEL

14. **EfficientNet-B7:** The convolutional neural network design and scaling approach EfficientNet [30] uniformly scales all depth, breadth, and resolution dimensions using a compound coefficient. In contrast to conventional practise, which scales these variables freely, the EfficientNet scaling method consistently increases network width, depth, and resolution using a set of predetermined scaling factors. EfficientNet uses a compound coefficient to equally scale network breadth, depth, and resolution. In total, EfficientNet-B7 comprises 813 layers.
    In image segmentation, EfficientNetB7 can be used to accurately classify each pixel of an image into different categories, such as background, foreground, or object of interest. This is achieved by applying convolutional filters to the input image, followed by downsampling and upsampling operations to capture both local and global features. Image of the model is shown in FIGURE 3.8[31].
    Overall, EfficientNet-B7 represents a promising approach for image segmentation, with potential applications in medical imaging, autonomous driving, and other fields where accurate segmentation of complex images is required.

Convolutional stem Layers,

$$output\_i = swish(batch\_norm\_i(conv\_i(output\_i\text{-}1))) \tag{30}$$

where output_i is the output of the i-th convolutional layer, conv_i is the convolution operation with learnable parameters, batch_norm_i is the batch normalization operation, swish is the swish activation function, and output_i-1 is the output of the (i-1)-th convolutional layer.

The computation of the network head is given by,

$$output = fc2(dropout(swish(batch\_norm1(fc1(flatten(output\_i)))))) \tag{31}$$

where output is the final output of the network, flatten is the operation that flattens the output of the last convolutional layer, fc1 and fc2 are fully connected operations with learnable weights, batch_norm1 is the batch normalization operation, swish is the swish activation function, and dropout is the dropout operation.

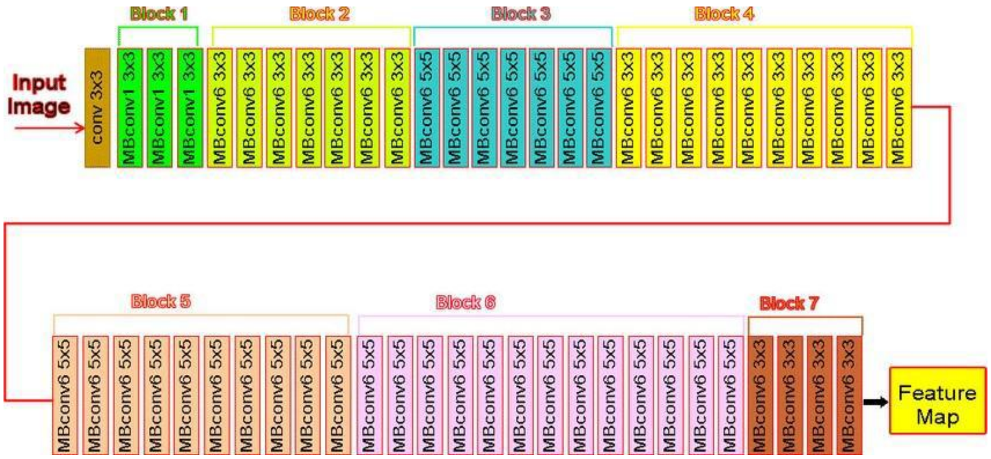


FIGURE 3.8: EFFICIENTNET-B7 MODEL

3.3 MODULES

The modules that are used for detecting lung cancer includes:

1. **Data Collection:** A large dataset of histopathological images of lungs was collected for the proposed work. The dataset was collected from various sources, including public medical archives, hospitals, and research institutions. The images were obtained using a variety of techniques, including digital microscopy and radiology.

   The dataset contains high-resolution images of both malignant and benign lung tissues, which were used to train and test the machine learning models. The annotations were used as labels for supervised learning algorithms.

   The dataset was carefully curated to ensure that it represents a diverse range of lung tissue samples. This includes different tissue types, stages of cancer, and patient demographics. The dataset also includes images that are difficult to classify, such as those with overlapping tissues or low-quality images, to ensure that the algorithms can handle real-world scenarios.

   Overall, the data collection process was crucial in ensuring that the machine learning algorithms were trained and tested on a diverse and representative dataset of lung histopathological images. This will ultimately improve the accuracy and reliability of the algorithms in predicting lung cancer from histopathological images.

2. **Data Preprocessing:** The second module, data preprocessing, involves preparing the data for analysis. This includes resizing the images, normalizing the pixel values, and removing any unwanted artifacts from the images. Data preprocessing is a crucial step to obtain accurate results in predicting lung cancer from histopathological images. Preprocessing is done to clean and prepare the data for further analysis.

   First, the histopathological images of lung tissue were collected and stored in a dataset. The images were in various sizes, resolutions, and color spaces, and some images contained artifacts, such as text and grids, that could interfere with the analysis. Therefore, the images were preprocessed by resizing to a standard size and converting to grayscale, which reduced the size of the dataset and eliminated color variations. The images were also filtered to remove noise and sharpen the edges, which improved the image quality.

   Next, image segmentation was performed on the preprocessed images to isolate the regions of interest (ROIs) for analysis. This step is essential to eliminate unwanted or irrelevant parts of the image and to focus only on the tissue that is relevant to the prediction of lung cancer. Various segmentation algorithms were used, including thresholding, region growing, and edge detection, to segment the images into ROIs.

   In summary, data preprocessing is an essential step in accurately predicting lung cancer from histopathological images. It involves various steps, including image resizing, filtering, segmentation, feature extraction, normalization, and feature selection, to obtain a clean and standardized dataset for analysis.

3. **Feature Extraction:** In the field of image processing, feature extraction is a critical step that involves extracting relevant and useful features from raw image data to make it easier for machine learning models to analyze and classify them. Feature extraction techniques were utilized to extract meaningful and discriminative features from the lung histopathological images for lung cancer classification.

   GLCM is a popular texture analysis technique that calculates the occurrence of pixel pairs with specific spatial relationships, which can provide valuable information on image texture. LBP, on the other hand, encodes the local texture patterns of an image by comparing each pixel with its neighbors. Gabor filters are a set of spatial frequency filters that are commonly used in image processing to detect texture features.

   In addition to texture-based feature extraction techniques, It also employed deep learning-based feature extraction using pre-trained Convolutional Neural Networks (CNNs) such as Inception-V3, ResNet50, and VGG16. These pre-trained models have been trained on large datasets and have learned to extract relevant features from images automatically. By utilizing these pre-trained models as feature extractors, the methodology was able to capture high-level features from the lung histopathological images that can be used to classify lung cancer accurately.

   Overall, the feature extraction techniques used in the thesis allowed for the extraction of discriminative features from the lung histopathological images, enabling accurate classification of lung cancer. The proposed methodology demonstrated that a combination of texture-based and deep learning-based feature extraction techniques can significantly improve the performance of machine learning models in classifying lung cancer.

4. **Image Segmentation:** The picture segmentation process is crucial to the development of our study of the methodology. The process of segmenting an image is dividing it into more manageable, homogeneous portions or sections according to factors like colour, texture, or intensity. The primary focus of the proposed work is the segmentation of malignant and non-cancerous zones in lung histopathology images.

   To accomplish picture segmentation, Images were preprocessed by normalising and turning them into grayscale. Then, Images were segmented with the help of number of techniques, such as thresholding, edge detection, and morphological processes. The algorithms employed to implement these strategies included SVM, random forest, KNN, logistic regression, naive bayes, CatBoost, XGBoost, decision tree, ANN, CNN, Inception-V3, ResNet50, and VGG16.

   Performance of these algorithms were compared based on their accuracy, precision, recall, and F1-score. Results indicated that deep learning algorithms such as CNN, Inception-V3, ResNet50, and VGG16 outperformed traditional machine learning algorithms.

   Outcome of the study of methodology demonstrate that accurate image segmentation is critical for predicting lung cancer accurately. This proposed methodology provides valuable insights into the potential of using image segmentation in combination with machine learning algorithms to diagnose lung cancer.

5. **Model development:** The model development stage of the "Image Segmentation of Lung Histopathological Image and Predicting lung cancer using ML algorithm" proposed study involved using various machine learning algorithms to predict lung cancer based on the segmented lung histopathological images. The algorithms used include SVM, random forest, KNN, logistic regression, naive Bayes, CatBoost, XGBoost, decision tree, ANN, CNN, Inception-V3, ResNet50, VGG16 and EfficientNet-B7.

   The first step in model development was splitting the dataset into training, validation, and testing sets. The training set was used to train the models, the validation set was used to tune the hyperparameters of the models, and the testing set was used to evaluate the performance of the models.

   The machine learning algorithms were created in Python using the Scikit-Learn, TensorFlow, and Keras libraries. Accuracy, precision, recall, and F1 score were only a few of the metrics used to evaluate the algorithms' performance. Using the training and validation sets, the algorithms were trained and verified.

   The results of the model generation stage showed that more traditional machine learning approaches including SVM, random forest, KNN, logistic regression, and naive Bayes underperformed deep learning techniques like CNN, Inception-V3, ResNet50, and VGG16.

   The results also showed that the performance of the deep learning algorithms improved as the complexity of the model increased. For example, ResNet50 models outperformed the VGG16 model, which in turn outperformed the CNN model.

   Overall, the model development stage of the "Image Segmentation of Lung Histopathological Image and Predicting lung cancer using ML algorithm" work showed that deep learning algorithms are more effective than traditional machine learning algorithms in predicting lung cancer based on segmented lung histopathological images. The results also showed that the complexity of the model plays a significant role in the performance of the algorithms.

6. **Performance Evaluation:** In this methodology for finding lung cancer, the performance of various machine learning algorithms was evaluated for the task of predicting lung cancer using histopathological images. The algorithms used for the study were SVM, random forest, KNN, logistic regression, naive Bayes, CatBoost, XGBoost, decision tree, ANN, CNN, Inception-V3, ResNet50, VGG16 and EfficientNet-B7.

   The dataset's training and testing sets were developed to evaluate the efficacy of various approaches. Accuracy, sensitivity, specificity, and F1 score were the performance metrics employed for evaluation.

   Overall, the results demonstrated that ResNet50 and VGG16 deep learning models performed better at predicting lung cancer from histopathological pictures than other deep learning models. Both CatBoost and the more established machine learning models performed equally well.

   These findings show that utilising histopathological images, machine learning algorithms can predict lung cancer with high accuracy, which may have important repercussions for the early diagnosis and management of the disease.

Chapter 4

RESULTS AND DISCUSSION

In the study, Various ML algorithms are implemented for the task of image segmentation of lung histopathological images and predicting lung cancer. For lung cancer detection, Machine learning algorithms like SVM, random forest, KNN, logistic regression, naive bayes, CatBoost, XGBoost, Decision Tree and deep neural networks like ANN, CNN, Inception-V3, ResNet50, and VGG16 models are used for this task.

The experimental results show that CatBoost out of all machine learning algorithm produces the best testing accuracy and training accuracy.

TABLE 4.1: ML ALGORITHM RESULT FOR TRAINING SET

S. No.	**ALGORITHM**	**ACCURACY**	**MCC**	**F1 SCORE**
1.	KNN	0.951	0.892	0.950
2.	RANDOM FOREST	0.982	0.974	0.982
3.	DECISION TREE	0.986	0.967	0.986
4.	NAÏVE BAYES	0.979	0.953	0.979
5.	SVC	0.987	0.975	0.987
6.	LOGISTIC REGRESSION	0.989	0.991	0.989
7.	CATBOOST	0.991	0.990	0.991
8.	XGBOOST	0.905	0.786	0.905

TABLE 4.1 shows the results of using ML algorithm on the training set of lung histopathological image dataset. There are 12000 images in the training set.

TABLE 4.2: ML ALGORITHM RESULT FOR TESTING SET

S. No.	**ALGORITHM**	**ACCURACY**	**MCC**	**F1 SCORE**
1.	KNN	0.884	0.742	0.876
2.	RANDOM FOREST	0.982	0.974	0.982
3.	DECISION TREE	0.976	0.947	0.976
4.	NAÏVE BAYES	0.979	0.953	0.979
5.	SVC	0.982	0.959	0.982
6.	LOGISTIC REGRESSION	0.985	0.965	0.985
7.	CATBOOST	0.991	0.981	0.991
8.	XGBOOST	0.799	0.552	0.800

TABLE 4.2 shows the results of using ML algorithm on the testing set of lung histopathological image dataset. There are 3000 images in the testing set.

Performance of different deep neural network models was compared using different evaluation metrics. One of the key evaluation metrics used was accuracy. Accuracy comparison of different models was also visualized using various plots such as bar chart, line plot and boxplot.

The bar chart was used to compare the accuracy of different models in a simple and intuitive way. Each bar represents the accuracy of a particular model and the height of the bar indicates the accuracy value. ResNet50 model achieved higher accuracy than other models.

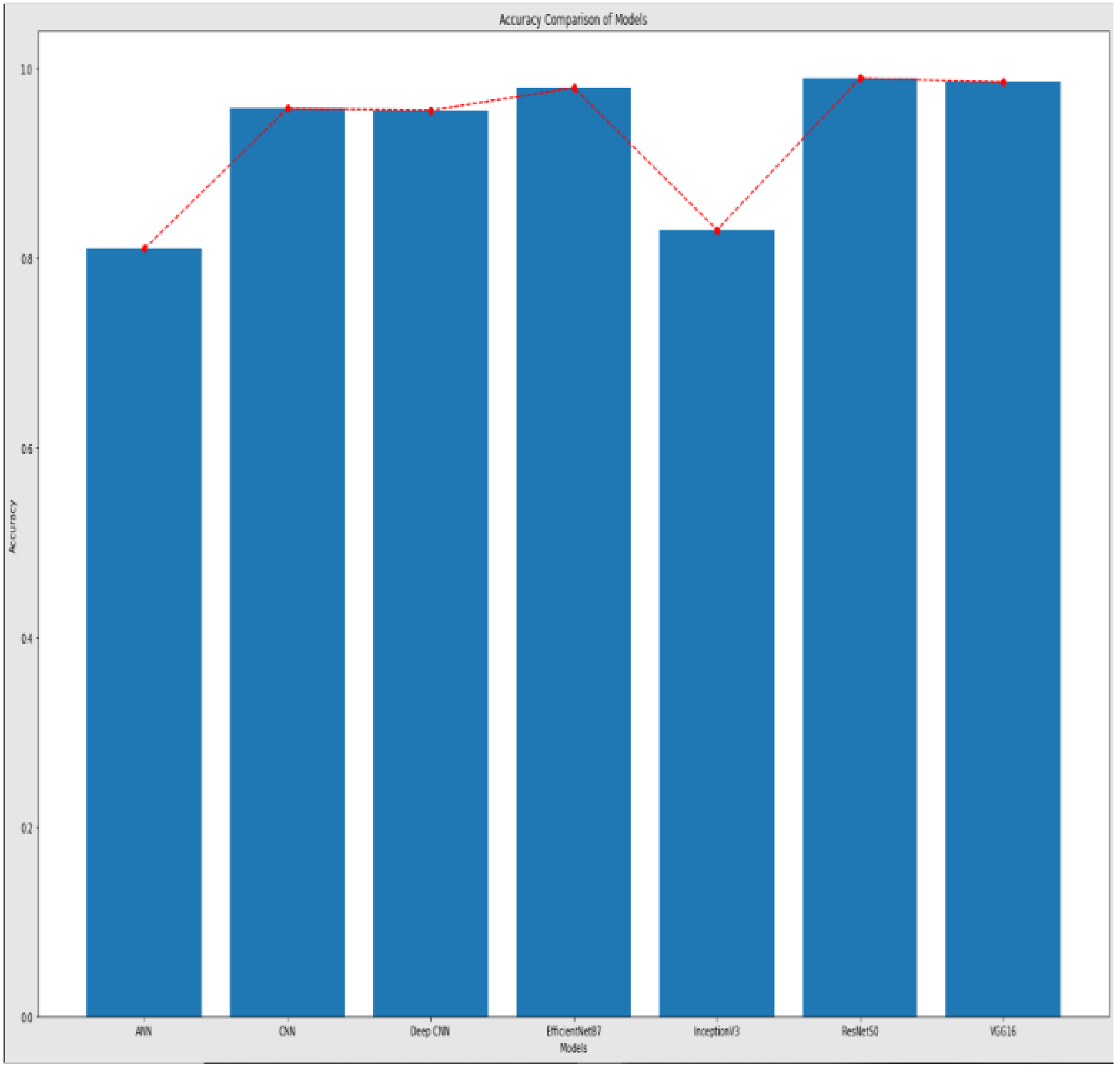


FIGURE 4.1: ACCURACY COMPARISON OF DEEP NEURAL NETWORK MODEL USING BAR CHART

The line plot was used to visualize the trend of accuracy over time, where time refers the number of epoch or iterations. Each line represents the accuracy of a particular model, and the x-axis represents the number of epoch or iterations, while the y-axis represents the accuracy value. ResNet50 model achieved higher accuracy than other models.

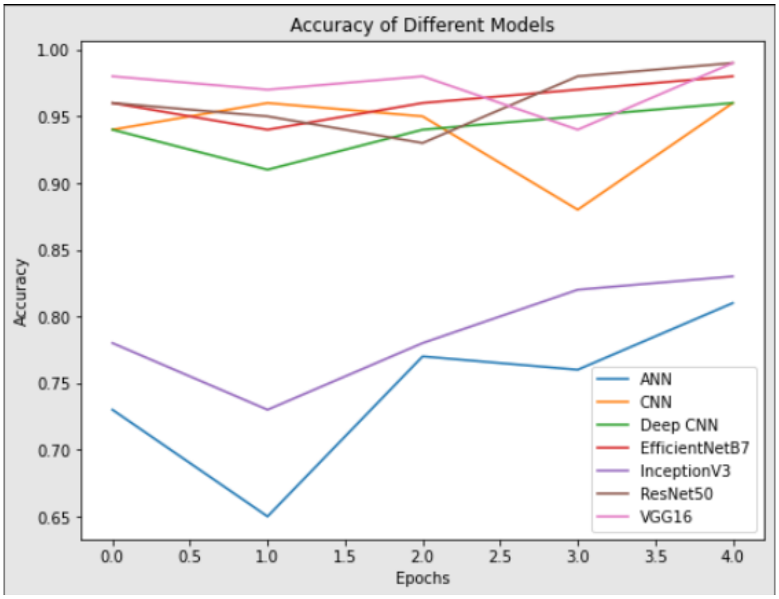


FIGURE 4.2: ACCURACY COMPARISON OF DEEP NEURAL NETWORK MODEL USING LINE CHART

The boxplot was used to compare the distribution of accuracy values for different models. Each boxplot represents the accuracy distribution for a particular model, where the box represents the Interquartile range (IQR) and the whiskers represent the range of accuracy values. Out of all the models, ResNet50 had a higher median accuracy value and a smaller variation in accuracy values compared to other models.

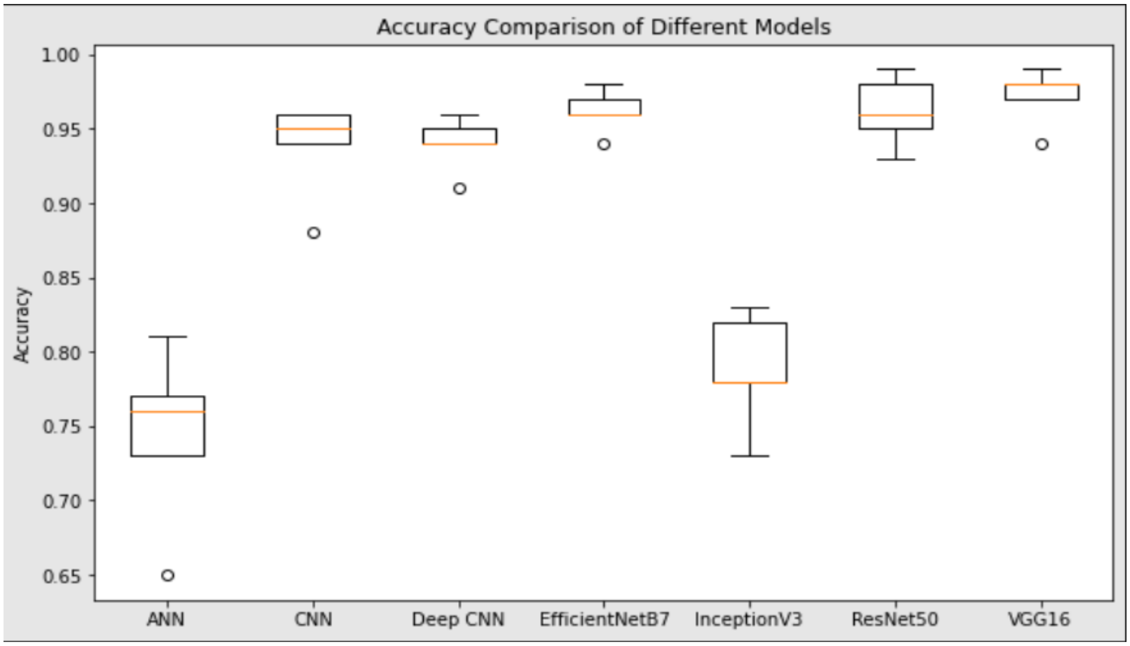


FIGURE 4.3: ACCURACY COMPARISON OF DEEP NEURAL NETWORK MODEL USING BOXPLOT

Chapter 5

CONCLUSION AND FUTURE WORK

In the proposed work, Machine learning algorithms were explored for image segmentation of lung histopathological images and predicting lung cancer. Various machine learning algorithms was also tested including SVM, random forest, KNN, logistic regression, naïve bayes, CatBoost, XGBoost, Decision tree and various deep neural network models including ANN, CNN, Deep-CNN, Inception-V3, ResNet50, VGG16 and EfficientNet-B7.

Results indicate that deep neural networks such as CNN, Inception-V3, ResNet50 and VGG16 outperformed traditional machine learning in lung cancer prediction. Among these deep neural networks, ResNet50 achieved the highest accuracy in lung cancer prediction. It demonstrated the ability to identify and segment regions of interest within the lung histopathological images and accurately predict the presence of the lung cancer.

The use of deep neural networks in image segmentation and lung cancer prediction can have significant implications in the field of medical image analysis. It has the potential to improve the accuracy and efficiency of lung cancer diagnosis, which is critical in ensuring timely and effective treatment. Further research can be conducted to explore the application of these algorithms in larger datasets and in other medical imaging tasks.

APPENDICES

APPENDIX 1

CODE

EDA OF LUNG CANCER.IPYNB

```
from keras.layers import Conv2D, ZeroPadding2D, Activation, Input, Dropout, Add
from keras.models import Model
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.layers import BatchNormalization
from keras.layers.pooling import MaxPooling2D
from keras.layers.core import Flatten, Dense
from keras.utils import np_utils
from tensorflow.keras.layers import Layer, InputSpec
from keras import backend as K
import numpy as np
import os
folders=[]
def list_folders(rootdir):
    for folder in os.listdir(rootdir):
        d = os.path.join(rootdir, folder)
        if os.path.isdir(d):
            print(d)
            folders.append(d)
            list_folders(d)
rootdir = 'lung_image_sets'
list_folders(rootdir)
import pandas as pd
from pathlib import Path
image_dir_path = 'lung_image_sets/lung_n'
```

```
paths = [path.parts[-3:] for path in Path(image_dir_path).rglob('*.jpeg')]
df1 = pd.DataFrame(data=paths, columns=['Root', 'Type', 'Images'])
print(df1)
image_dir_path = 'lung_image_sets/lung_scc'
paths = [path.parts[-3:] for path in Path(image_dir_path).rglob('*.jpeg')]
df2 = pd.DataFrame(data=paths, columns=['Root', 'Type', 'Images'])
print(df2)
image_dir_path ='lung_image_sets/lung_aca'
paths = [path.parts[-3:] for path in Path(image_dir_path).rglob('*.jpeg')]
df3 = pd.DataFrame(data=paths, columns=['Root', 'Type', 'Images'])
print(df3)
df = pd.concat([df1,df2,df3])
df.reset_index()
import cv2
import os
from matplotlib import pyplot as plt
def load_images(folder):
    images = []
    for filename in os.listdir(folder):
        img = cv2.imread(os.path.join(folder, filename))
        if img is not None:
            img = cv2.resize(img, (80,80))
            images.append(img)
    return np.array(images)
benign_images = load_images('lung_image_sets/lung_n')
mal_aca_images = load_images('lung_image_sets/lung_aca')
```

```
mal_scc_images = load_images('lung_image_sets/lung_scc')
print(f"Number of images for every class: BENIGN {benign_images.shape[0]}, ADENOCARCINOMAS {mal_aca_images.shape[0]}, SQUAMOS CELL CARCINOMAS {mal_scc_images.shape[0]}.")
print(f"Images shape: {benign_images[0].shape}.")
indices = [0, 40, 2300]
plt.figure(1, figsize=(15,5))
plt.grid(None)
for n, idx in enumerate(indices):
    plt.subplot(n+1, 3, 1)
    plt.imshow(benign_images[idx])
    plt.title('benign')
    plt.subplot(n+1, 3, 2)
    plt.imshow(mal_aca_images[idx])
    plt.title('malignant aca')
    plt.subplot(n+1, 3, 3)
    plt.imshow(mal_scc_images[idx])
    plt.title('malignant scc')
plt.show()
samples = np.concatenate((benign_images, mal_aca_images, mal_scc_images))
labels = np.array(benign_images.shape[0] * [0] + mal_aca_images.shape[0] * [1] + mal_scc_images.shape[0] * [2])
images = samples.astype('float32') / 255
from sklearn.model_selection import train_test_split
```

```
train_images, test_images, train_labels, test_labels = train_test_split(images, labels, test_size = 0.2)
val_images, test_images, val_labels, test_labels = train_test_split(test_images, test_labels, test_size = 0.5)
train_labels = np_utils.to_categorical(train_labels, 3)
val_labels = np_utils.to_categorical(val_labels, 3)
test_labels = np_utils.to_categorical(test_labels, 3)
print(train_labels[0])
print(f"Validation labels shape after one hot encoding: {val_labels.shape}")
print(f"Validation images shape: {val_images.shape}")
```

MODEL COMPARISON.IPYNB

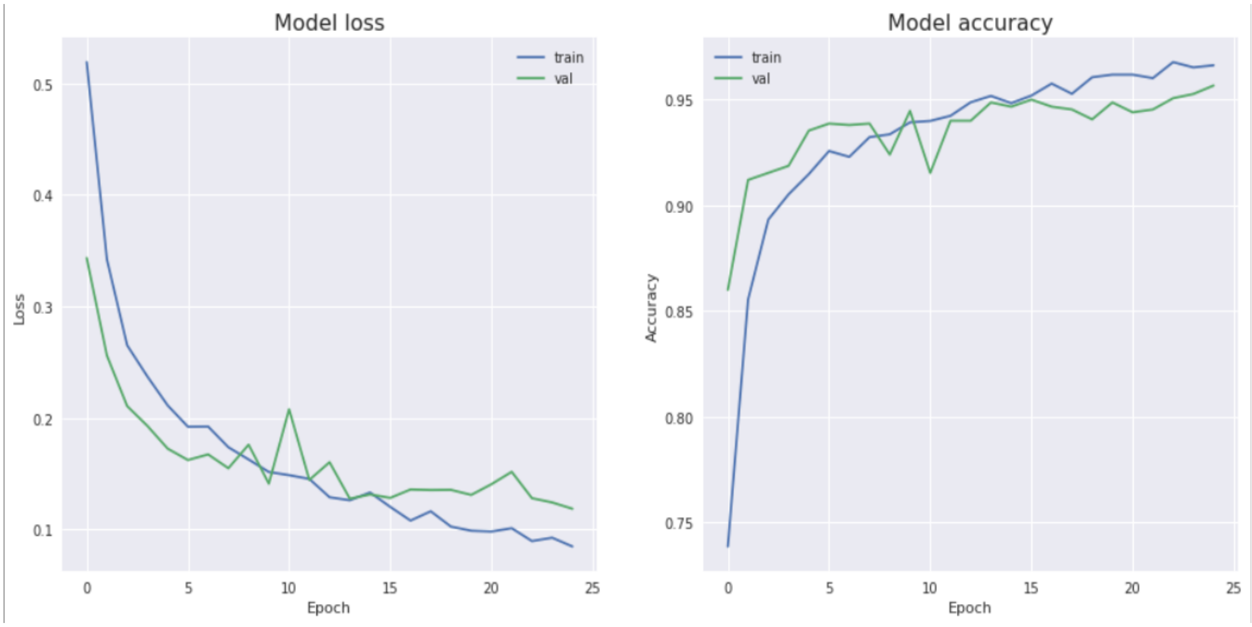
```
import matplotlib.pyplot as plt
models=['ANN','CNN','Deep CNN','EfficientNetB7','InceptionV3','ResNet50','VGG16']
accuracies=[0.81,0.958,0.956,0.98,0.83,0.99,0.986]
plt.figure(figsize=(25,15))
plt.bar(models,accuracies)
plt.plot(models,accuracies,marker='o',linestyle='--',color="r")
plt.title('Accuracy Comparison of Models')
plt.xlabel('Models')
plt.ylabel('Accuracy')
plt.show()
ann_acc = [0.73, 0.65, 0.77, 0.76, 0.81]
cnn_acc = [0.94, 0.96, 0.95, 0.88, 0.96]
deep_cnn_acc = [0.94, 0.91, 0.94, 0.95, 0.96]
efficientnet_acc = [0.96, 0.94, 0.96, 0.97, 0.98]
inceptionv3_acc = [0.78, 0.73, 0.78, 0.82, 0.83]
resnet50_acc = [0.96, 0.95, 0.93, 0.98, 0.99]
vgg16_acc = [0.98, 0.97, 0.98, 0.94, 0.99]
all_acc = [ann_acc, cnn_acc, deep_cnn_acc, efficientnet_acc, inceptionv3_acc, resnet50_acc, vgg16_acc]
```

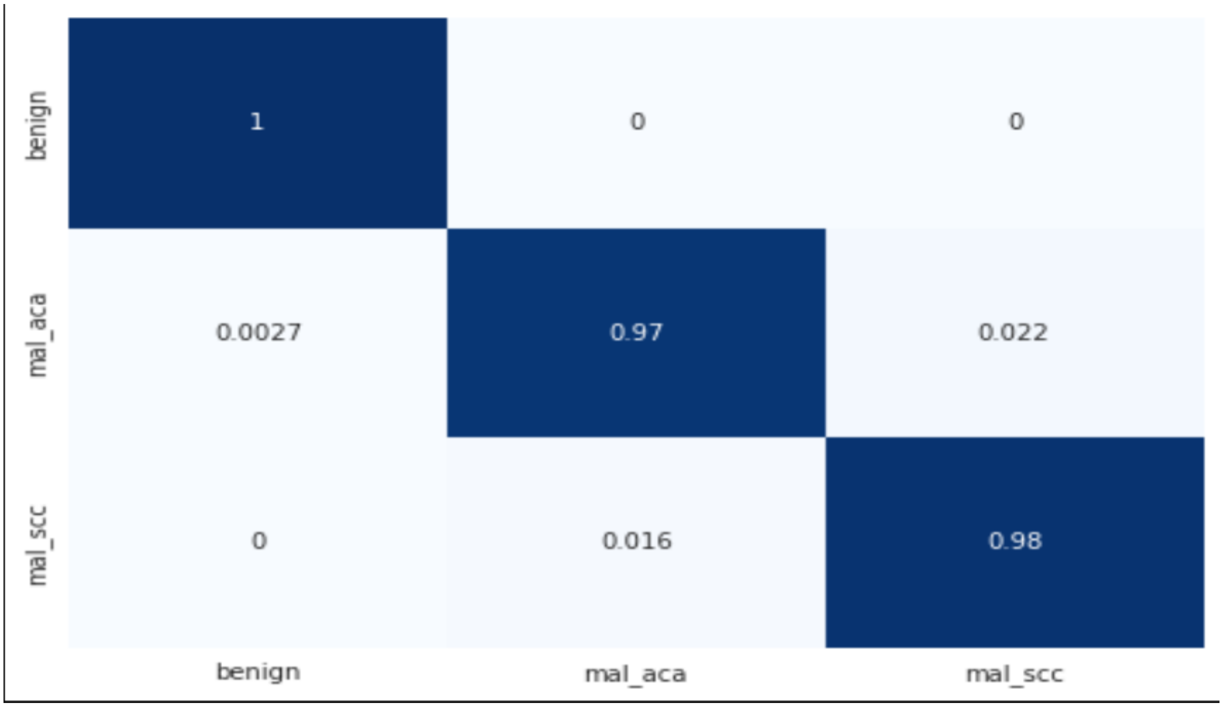
```
plt.figure(figsize=(8, 6))
for i in range(len(models)):
    plt.plot(all_acc[i], label=models[i])
plt.title('Accuracy of Different Models')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()
plt.show()
fig, ax = plt.subplots(figsize=(10, 6))
ax.boxplot(all_acc)
ax.set_xticklabels(['ANN', 'CNN', 'Deep CNN', 'EfficientNetB7', 'InceptionV3', 'ResNet50',
'VGG16'])
ax.set_ylabel('Accuracy')
ax.set_title('Accuracy Comparison of Different Models')
plt.show()
```

APPENDIX 2

OUTPUT SCREENSHOT

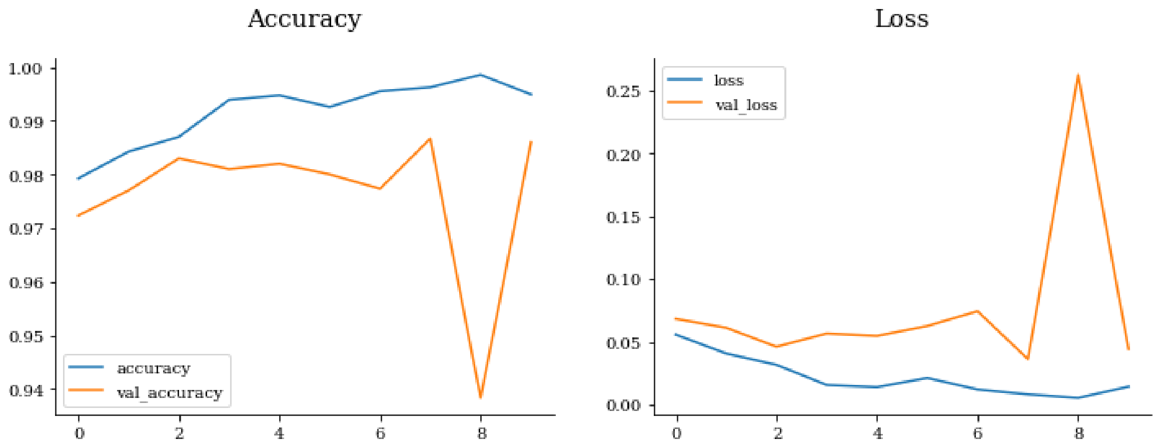
DEEP CNN LUNG CANCER.IPYNB

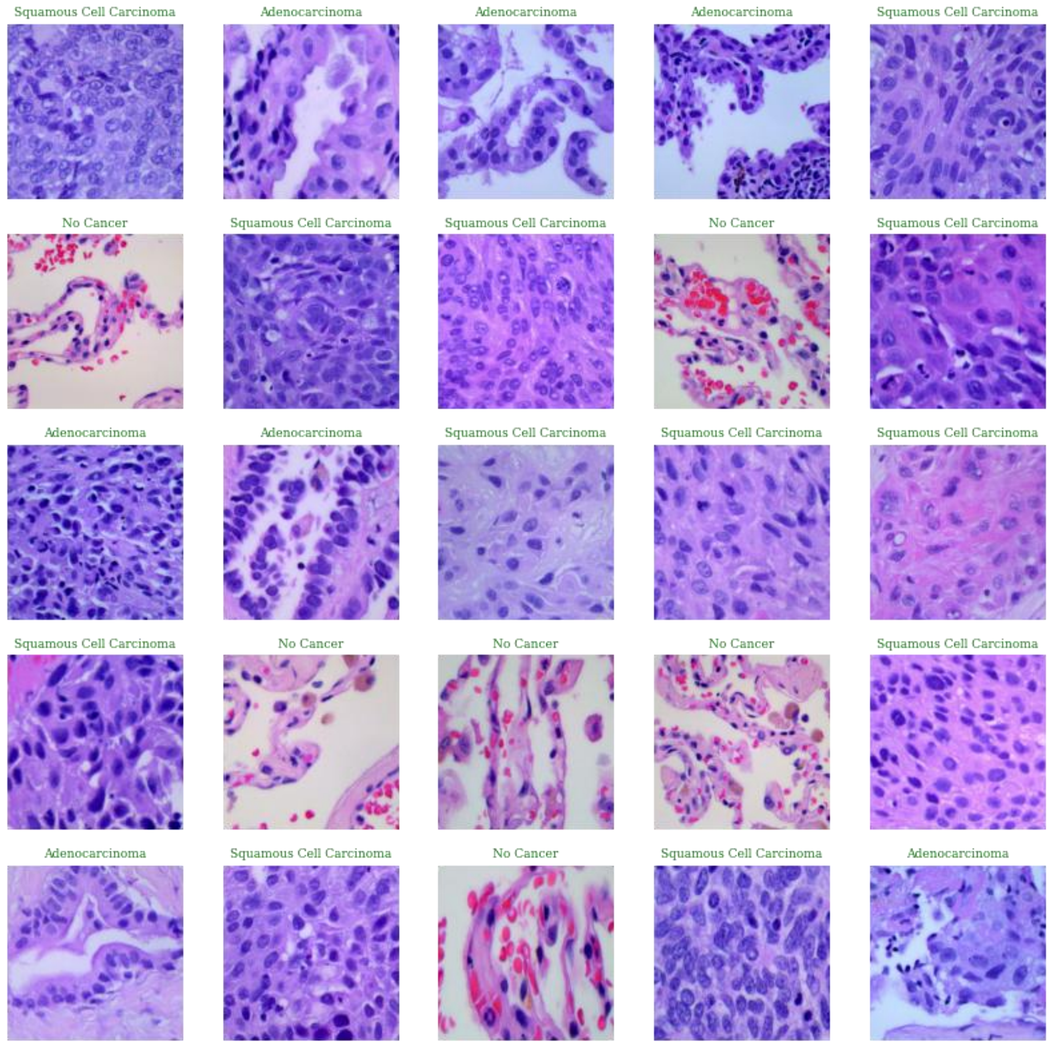




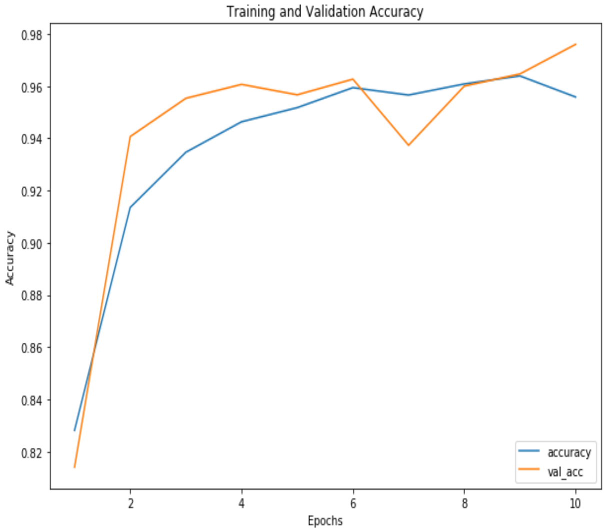
	benign	mal_aca	mal_scc
benign	1	0	0
mal_aca	0.0027	0.97	0.022
mal_scc	0	0.016	0.98

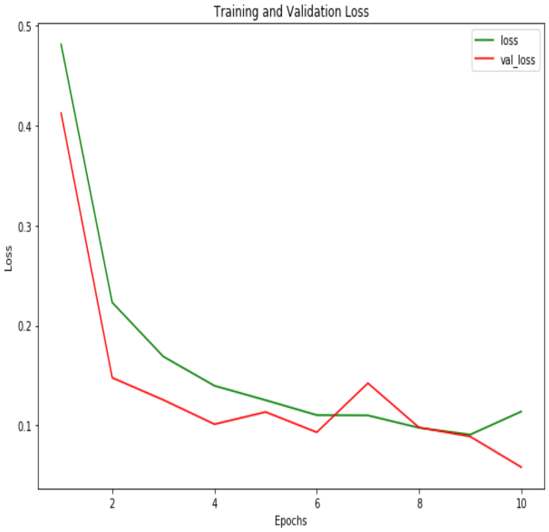
VGG16 LUNG CANCER.IPYNB

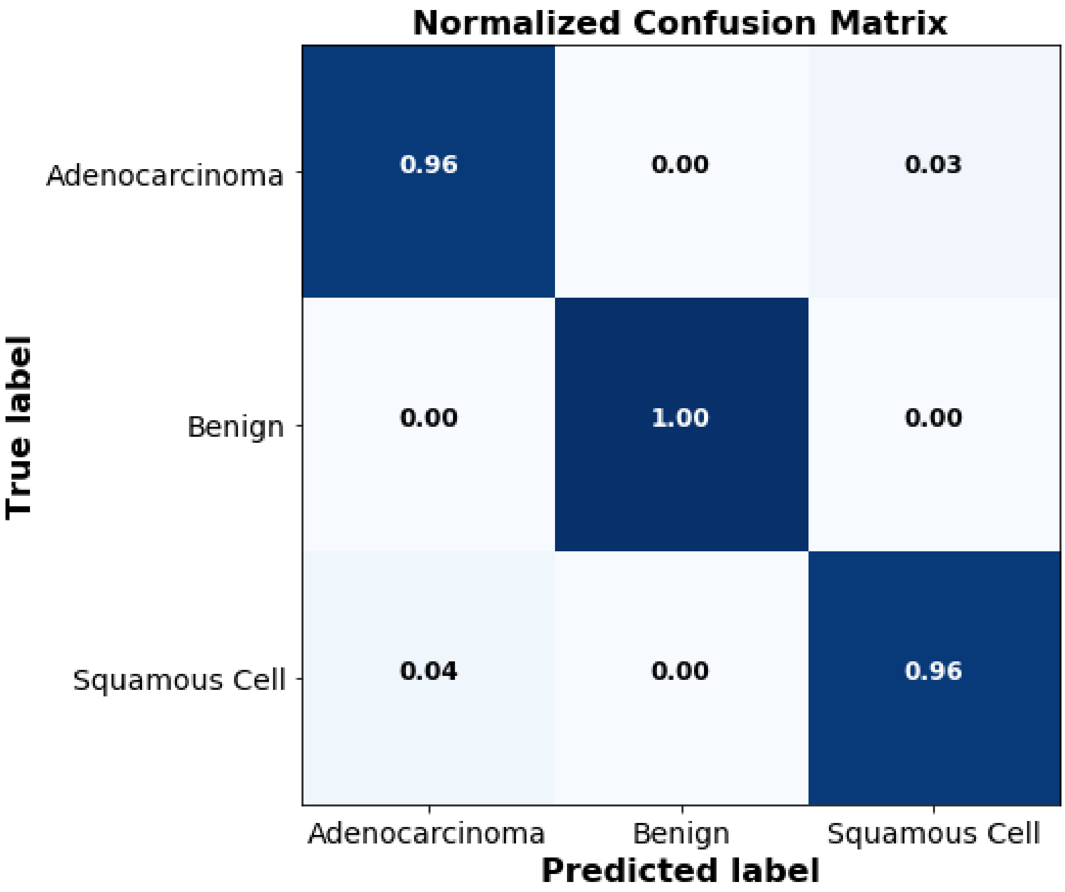




EFFICIENTNET-B7.IPYNB







REFERENCES

[1]. Ragavan, M., & Patel, M. I. (2022). The evolving landscape of sex-based differences in lung cancer: a distinct disease in women. European Respiratory Review, 31(163).

[2]. Tan, A. C., & Tan, D. S. (2022). Targeted therapies for lung cancer patients with oncogenic driver molecular alterations. Journal of Clinical Oncology, 40(6), 611-625.

[3]. Liao, Y., Wu, X., Wu, M., Fang, Y., Li, J., & Tang, W. (2022). Non-coding RNAs in lung cancer: emerging regulators of angiogenesis. Journal of Translational Medicine, 20(1), 1-11.

[4]. Merie, R., Gee, H., Hau, E., & Vinod, S. (2022). An overview of the role of radiotherapy in the treatment of small cell lung cancer–a mainstay of treatment or a modality in decline?. Clinical Oncology.

[5]. Hişam, D., & Hişam, E. (2021, October). Deep learning models for classifying cancer and COVID-19 lung diseases. In 2021 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-4). IEEE.

[6]. Rudin, C. M., Brambilla, E., Faivre-Finn, C., & Sage, J. (2021). Small-cell lung cancer. Nature Reviews Disease Primers, 7(1), 1-20.

[7]. Mukherjee, S., & Bohra, S. U. (2020, December). Lung Cancer Disease Diagnosis Using Machine Learning Approach. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 207-211). IEEE.

[8]. Jena, S. R., George, T., & Ponraj, N. (2019, February). Texture analysis based feature extraction and classification of lung cancer. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-5). IEEE.

[9]. Barta, J. A., Powell, C. A., & Wisnivesky, J. P. (2019). Global epidemiology of lung cancer. Annals of global health, 85(1).

[10]. Ahmed, B. T. (2019). Lung Cancer Prediction and Detection Using Image Processing Mechanisms: An Overview. Signal and Image Processing Letters, 1(3), 20-31.

[11]. Putora, P. M., Glatzer, M., Belderbos, J., Besse, B., Blackhall, F., Califano, R., ... & De Ruysscher, D. (2019). Prophylactic cranial irradiation in stage IV small cell lung cancer: Selection of patients amongst European IASLC and ESTRO experts. Radiotherapy and oncology, 133, 163-166.

[12]. Wong, D. M., Fang, C. Y., Chen, L. Y., Chiu, C. I., Chou, T. I., Wu, C. C., ... & Tang, K. T. (2018, April). Development of a breath detection method based E-nose system for lung cancer identification. In 2018 IEEE International Conference on Applied System Invention (ICASI) (pp. 1119-1120). IEEE.

[13]. Alam, J., Alam, S., & Hossan, A. (2018, February). Multi-stage lung cancer detection and prediction using multi-class svm classifie. In 2018 International conference on computer, communication, chemical, material and electronic engineering (IC4ME2) (pp. 1-4). IEEE.

[14]. Romaszko, A. M., & Doboszyńska, A. (2018). Multiple primary lung cancer: a literature review. Adv Clin Exp Med, 27(5), 725-730.

[15]. de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. (2018). The epidemiology of lung cancer. Translational lung cancer research, 7(3), 220.

[16]. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[17]. https://www.ibm.com/in-en/topics/knn

[18]. https://www.geeksforgeeks.org/understanding-logistic-regression/

[19]. https://www.ibm.com/topics/naive-bayes

[20]. https://catboost.ai/

[21]. https://serokell.io/blog/random-forest-classification

[22].https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

[23].https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

[24].https://www.analyticsvidhya.com/blog/2021/07/understanding-the-basics-of-artificial-neural-network-ann/

[25].https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/

[26]. https://cloud.google.com/tpu/docs/inception-v3-advanced

[27]. https://datagen.tech/guides/computer-vision/resnet-50/

[28].https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918

[29].https://www.researchgate.net/figure/VGG-16-neural-network-structure-for-classification-and-regression-models_fig4_355097587

[30].https://www.tensorflow.org/api_docs/python/tf/keras/applications/efficientnet/EfficientNetB7

[31].https://www.researchgate.net/figure/EfficientNetB7-architecture_fig6_358902226