Minor Project Report on

Early Detection of Lungs Cancer Using 2D CNN

Submitted in partial fulfillment of

the requirements for the award of the grades in the 6th semester of

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

MIR TARIQUDDIN

Regd. No.: 2111100432

MD ADEEB AAZIL

Regd. No.: 2111100422

Under the guidance of

MR. MANORANJAN PANDA

Asst. Professor

School Of Computer Science

Odisha University of Technology and Research

Bhubaneswar, Odisha – 751029

Department of Computer Science and Engineering

ODISHA UNIVERSITY OF TECHNOLOGY AND RESEARCH,

BHUBANESWAR

CERTIFICATE

This is to certify that the project report entitled **Early Detection of Lungs Cancer Using 2D CNN : A Focus on CNN and Machine Learning submitted by Mir Tariquddin and Md Adeeb Aazil bearing registration number 2111100432 and 2111100422** respectively to the Department of Computer Science and Engineering, Odisha University of Technology and Research, formerly College of Engineering and Technology, Bhubaneswar, is a record of Bonafide research work under my supervision and I consider it worthy of consideration for partial fulfillment of the requirements for the award degree of Bachelor of Technology in Computer Science and Engineering under Odisha University of Technology and Research, Bhubaneswar.

Mr. Manoranjan Panda

(Guide)

ACKNOWLEDGEMENT

It is our privilege and solemn duty to express our deepest sense of gratitude to **Mr. Manoranjan Panda** of the Department, under whose guidance we carried out this work. I am indebted, for his valuable supervision, heart full cooperation and timely aid and advice till the completion of the thesis in spite of his pressing engagements. We would like to express our deep sense of gratitude to **Mr. Manoranjan Panda**, Coordinator of the Department for encouragement and inspiration throughout our project work. We wish to record our sincere gratitude to our respected Head of the School, **Prof. Ranjan Kumar Dash** for his constant support and encouragement in preparation of this report.

We take this opportunity to express our heartfelt thanks to all those who helped us in the completion of our seminar report work. We are very grateful to the author of various articles on the Internet, for helping us become aware of the research currently ongoing in this field.

We are very thankful to our parents for their constant support and love. Last, but not the least, I would like to thank my classmates for their valuable comments, suggestions and unconditional support.

Mir Tariquddin

Md Adeeb Aazil

DECLARATION

I certify that

I. The work contained in the seminar report is original and has been done myself under the general supervision of my supervisor.

II. The work has not been submitted to any other Institute for any degree or diploma.

III. I have followed the guidelines provided by the Institute in writing the report.

IV. Whenever I have used materials (data, theoretical analysis, figures, text) from the other sources, I have given due credit to them by citing them in the text of the seminar report and giving their details in the references.

V. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Mir Tariquddin

Md Adeeb Aazil

ABSTRACT

In this project, we delve into the realm of lung cancer detection and classification, leveraging machine learning and deep learning techniques applied to medical imaging data. Our exploration encompasses a spectrum of methodologies, including convolutional neural networks (CNNs), deep belief networks (DBNs), stacked autoencoders (SAEs), and support vector machines (SVMs), among others. Through a comprehensive review of recent literature, we unveil innovative approaches such as cloud-based CNN models and the integration of demographic information for enhanced prediction accuracy. By synthesizing key findings and methodologies, we shed light on the current landscape of lung cancer diagnosis, pinpoint areas for further research and refinement, and contribute to the ongoing quest for more effective diagnostic tools in the fight against lung cancer.

Keywords:

Machine Learning, Deep learning ,2D CNN, HeatMap, Image Mapping, Pandas

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Lung cancer remains one of the most prevalent and deadly forms of cancer worldwide, with significant implications for public health. Early detection and accurate risk assessment are crucial factors in improving patient outcomes and guiding clinical decision-making processes. In recent years, the integration of machine learning techniques with healthcare data has shown promise in enhancing predictive models for various medical conditions, including cancer.

In this study, we aim to leverage machine learning methodologies to develop a predictive model for assessing the likelihood of lung cancer based on survey data. Unlike traditional approaches that rely solely on structured data, such as demographic information and medical history, our approach integrates visualization techniques to capture complex relationships and patterns within the data. Specifically, we generate heatmap visualizations of the survey features, providing a rich and intuitive representation of the underlying characteristics associated with lung cancer risk.

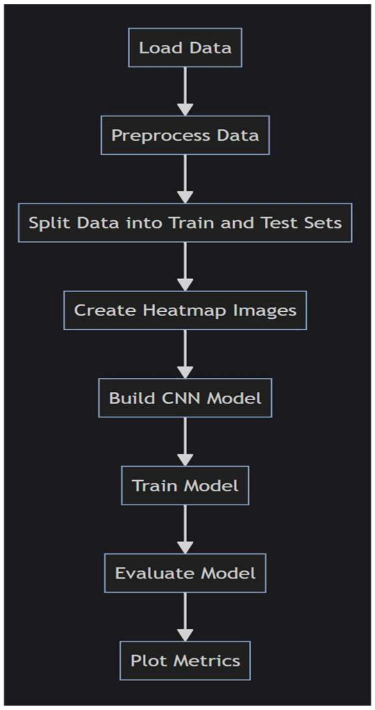

Fig. 1. Steps Of Lungs Cancer Prediction

To achieve this objective, we employ Convolutional Neural Networks (CNNs), a class of deep learning models renowned for their effectiveness in image recognition tasks. By treating the heatmap images as inputs to the CNN, we enable the model to automatically learn relevant features and patterns indicative of lung cancer risk. Through extensive training and evaluation, we aim to develop a robust and accurate predictive model capable of providing valuable insights into individualized lung cancer risk assessment

The outcomes of this study hold the potential to significantly impact clinical practice by facilitating early detection, guiding risk stratification strategies, and ultimately improving patient outcomes in the fight against lung cancer.

2. BACKGROUND STUDY

2.1. Literature Survey

[1] Nuruzzaman et al. Healthcare as a Service (HAAS): CNN-based cloud computing model for ubiquitous access to lung cancer diagnosis (2023)

He introduced the concept of Healthcare As a Service (HAAS), which revolves around a CNN-based cloud computing model designed to facilitate ubiquitous access to lung cancer diagnosis. The model they presented, called PresentHAASNet, stands out with an impressive accuracy rate of 96.07%.

1. Healthcare As a Service (HAAS): This is a framework that leverages cloud computing technologies to provide on-demand access to healthcare services. It allows healthcare resources to be delivered efficiently and effectively over the internet, enabling greater accessibility and scalability.
2. CNN-Based Cloud Computing Model: Convolutional Neural Networks (CNNs) are a type of deep learning algorithm particularly effective in image recognition tasks. In this context, the CNN-based model is specifically trained to analyze medical images, such as lung scans, to assist in the diagnosis of lung cancer.
3. PresentHAASNet: This is the specific CNN model developed by Nuruzzaman et al. It's optimized for cloud compatibility, meaning it's designed to run efficiently on cloud infrastructure. This ensures that the model can be accessed and utilized from anywhere with an internet connection, making it highly accessible to healthcare providers and patients alike.
4. Accuracy Rate of 96.07%: This metric indicates how often the model's predictions align with the correct diagnoses. An accuracy rate of 96.07% is quite high, suggesting that PresentHAASNet is highly reliable in identifying lung cancer from medical images.

[2] Shandilya et al Analysis of Lung Cancer by Using Deep Neural Network (2021)

He conducted research on various CNN models, including MobileNet, VGG-19, ResNet 101, DenseNet 121, DenseNet 169, Inception V3, InceptionResNet V2, and MobileNetV2, for the classification of lung cancer. Among these models, ResNet 101 emerged as the top performer, achieving an impressive accuracy of 98.67%.

CNN Models for Lung Cancer Classification: The researchers explored several popular CNN architectures, each optimized for different tasks and complexities. These models are widely used in image recognition and classification tasks due to their effectiveness in learning hierarchical features from data.

ResNet 101: ResNet stands for Residual Network, and ResNet 101 is a specific variant characterized by its depth (101 layers). It utilizes residual connections to address the problem of vanishing gradients in very deep neural networks. In this study, ResNet 101 outperformed other models, achieving the highest accuracy of 98.67%.

Implications for Lung Cancer Detection: The research findings have significant implications for the development of CNN-based lung cancer detection models. By identifying ResNet 101 as the most effective architecture for this task, the study provides valuable insights for researchers looking to improve the accuracy and reliability of lung cancer diagnosis using deep learning techniques.

Future Research Directions: The results of this study can serve as a foundation for future research endeavors aimed at refining and enhancing CNN-based lung cancer detection models. Researchers can build upon these findings to develop more sophisticated algorithms and techniques that further improve the performance of lung cancer diagnosis systems.

[3] R. Raja Subramanian et al. Lung Cancer Detection by Harnessing the Power of Deep Learning with Convolutional Neural Networks (2023)

He presented a model for lung cancer detection based on Convolutional Neural Network (CNN) technology. In their study, they utilized pre trained ImageNet models, specifically LeNet, AlexNet, and VGG-16, for this purpose. Among these models, they found that using AlexNet yielded the best results.

Here's a breakdown of their approach and findings:

1. Pretrained ImageNet Models: ImageNet is a large dataset of labeled images used for training and benchmarking computer vision models. Pretrained models on ImageNet, such as LeNet, AlexNet, and VGG-16, have been trained on this dataset for general image recognition tasks. Subramanian et al. leveraged these pretrained models as a starting point for their lung cancer detection model.
2. Utilization of AlexNet: Among the pretrained models, the researchers found that AlexNet performed the best for their specific task of lung cancer detection. AlexNet is a deep convolutional neural network architecture that gained prominence for winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. It consists of multiple convolutional and fully connected layers, enabling it to learn hierarchical features from input images effectively.
3. Feature Extraction: Instead of using the entire AlexNet architecture, the researchers focused on extracting features from the last fully connected layer of the model. This layer captures high-level representations of the input images, which are then used as input to a softmax classifier.
4. Combination with Softmax Classifier: The features extracted from the last fully connected layer of AlexNet were fed into a softmax classifier for final classification. The softmax classifier assigns probabilities to each class, in this case, indicating the likelihood of lung cancer presence.
5. High Accuracy: Through their approach combining AlexNet features with the softmax classifier, Subramanian et al. achieved an impressive accuracy of 99.52% in lung cancer detection. This high accuracy rate underscores the effectiveness of their model in accurately identifying cases of lung cancer from medical images.

[4] Jin et al A Comparative study of Lung Cancer Detection and Classification approaches in CT images (2020)

conducted a study employing Convolutional Neural Networks (CNNs) to classify lung cancer, achieving an accuracy of 84.6%. Despite this accuracy being slightly lower than some other studies, their research is significant for its contribution to the field of deep learning in medical imaging.

CNNs have become instrumental in medical image analysis due to their capacity to learn intricate patterns from data. Lung cancer remains a significant health challenge globally, underscoring the importance of accurate diagnostic tools. Jin et al. trained a CNN model to classify medical images, such as X-rays or CT scans, into lung cancer-positive or negative categories. The achieved accuracy, while not the highest reported, still demonstrates the potential of CNNs in aiding lung cancer detection.

Their findings suggest avenues for further research, including refining CNN architectures and optimizing training methodologies. Future investigations could explore novel approaches such as transfer learning, data augmentation, and integration of additional patient information to enhance model performance. Overall, Jin et al.'s study contributes to advancing the capabilities of CNNs in medical image analysis, with implications for improving early detection and treatment of lung cancer.

[5] Alakwaa et al. Investigation of Lung Cancer detection Using 3D Convolutional Deep Neural Network (2020)

He employed a segmentation approach using thresholding in conjunction with a 3D Convolutional Neural Network (CNN) to classify CT scans for lung cancer detection. Their study reported an accuracy of 86.6%.

1. Segmentation Approach: Segmentation is a critical step in medical image analysis where the regions of interest are identified and separated from the background or other structures. In this study, Alakwaa et al. applied thresholding techniques as part of their segmentation approach to isolate the lung regions from CT scans.
2. 3D CNN Classification: After segmentation, the segmented lung regions were inputted into a 3D CNN for classification. Unlike traditional 2D CNNs which process images slice by slice, 3D CNNs can capture spatial information across the entire volume of the scan, making them well-suited for analyzing volumetric medical data like CT scans.
3. Accuracy: The reported accuracy of 86.6% indicates the proportion of correctly classified cases out of the total cases evaluated. This accuracy rate suggests that their combined segmentation and classification approach using 3D CNNs holds promise for lung cancer detection from CT scans.
4. Significance: Alakwaa et al.'s study contributes to the ongoing efforts to develop accurate and efficient methods for lung cancer diagnosis from medical images. By combining segmentation techniques with 3D CNN classification, they demonstrate a holistic approach to analyze volumetric CT data, potentially improving the accuracy and reliability of lung cancer detection.

[6] Sun et al. Using Deep Learning for Classification of Lung Cancer on CT Images in Ardabil Province. (2023)

He explored various deep learning algorithms, including Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs), and Stacked Denoising Autoencoders (SDAE), for the classification of lung cancer. Among these algorithms, they found that Deep Belief Networks (DBNs) achieved the highest accuracy, reaching 0.8119.

Deep Learning Algorithms: Sun et al. investigated multiple deep learning algorithms known for their effectiveness in handling complex data and extracting meaningful features. CNNs are widely used for image classification tasks due to their ability to capture spatial hierarchies of features. DBNs and SDAEs, on the other hand, are types of deep generative models capable of learning hierarchical representations of data.

Lung Cancer Classification: The researchers applied these deep learning algorithms to the task of classifying lung cancer, likely using medical images such as X-rays or CT scans as input data. The goal was to develop a model capable of accurately distinguishing between images showing signs of lung cancer and those that do not.

Best Performance with DBNs: Among the algorithms evaluated, Deep Belief Networks (DBNs) yielded the highest accuracy, reaching 0.8119. This accuracy rate indicates the model's ability to correctly classify cases of lung cancer with a high degree of precision.

Significance: Sun et al.'s study contributes to the exploration of deep learning techniques in medical image analysis, particularly in the context of lung cancer classification. By demonstrating the effectiveness of DBNs in this task, their research provides insights into potential approaches for improving the accuracy of lung cancer diagnosis using deep learning algorithms.

[7] Song et al. Pre-Training Autoencoder for Lung Nodule Malignancy Assessment Using CT Images. (2020)

He employed deep neural networks, including Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), and Stacked Autoencoders (SAE), for the detection of lung cancer calcification. Among these networks, CNN achieved the highest performance, achieving an accuracy of 84.15%.

Here's a succinct overview of their study:

1. Deep Neural Networks for Lung Cancer Calcification: Song et al. utilized various deep learning architectures to detect lung cancer calcification, a critical aspect of lung cancer diagnosis visible in medical imaging data such as X-rays or CT scans.
2. CNN, DNN, and SAE: Convolutional Neural Networks (CNNs) are well-suited for image classification tasks, as they can effectively capture spatial hierarchies of features. Deep Neural Networks (DNNs) are general-purpose neural networks with multiple layers, capable of learning complex patterns in data. Stacked Autoencoders (SAEs) are a type of unsupervised learning algorithm used for feature learning and dimensionality reduction.
3. Performance Evaluation: After training and testing the various deep neural network architectures, Song et al. found that the CNN achieved the best performance, with an accuracy of 84.15%. This accuracy rate represents the proportion of correctly classified cases of lung cancer calcification.

4. Significance: Song et al.'s study contributes to the advancement of deep learning applications in medical imaging, specifically in the domain of lung cancer diagnosis. By demonstrating the effectiveness of CNNs in detecting lung cancer calcification, their research provides valuable insights into the potential of deep learning techniques to assist medical professionals in identifying and treating lung cancer.

[8] Park et al. Computer-aided detection of early interstitial lung diseases using low-dose CT images (2011)

He utilized a genetic algorithm to select optimal image features and determine the structure of an Artificial Neural Network (ANN). Their approach resulted in achieving an 80.0% sensitivity at 85.7% specificity.

Genetic Algorithm Optimization: Genetic algorithms are optimization techniques inspired by the process of natural selection. In this study, Park et al. applied a genetic algorithm to automatically select the most relevant image features for lung cancer detection. By doing so, they aimed to enhance the performance of their model by focusing on the most discriminative features in the input data.

Artificial Neural Network (ANN): ANNs are computational models inspired by the biological neural networks of the human brain. They consist of interconnected nodes organized in layers, with each node performing a mathematical operation on its input data. Park et al. used an ANN as the classifier for their lung cancer detection system.

Performance Metrics: The performance of the lung cancer detection system was evaluated using sensitivity and specificity metrics. Sensitivity measures the proportion of true positive cases correctly identified by the model, while specificity measures the proportion of true negative cases correctly identified. In this study, the system achieved an 80.0% sensitivity rate and an 85.7% specificity rate.

Significance: Park et al.'s study highlights the effectiveness of employing genetic algorithms to optimize feature selection and ANN structure for lung cancer detection. By using this approach, they were able to achieve a balance between sensitivity and specificity, which are crucial metrics for evaluating the performance of medical diagnostic systems.

[9] Shao et al. A detection approach for solitary pulmonary nodules based on CT images. (2012)

He developed a system for automatically detecting solitary pulmonary nodules (SPNs) using a Support Vector Machine (SVM) classifier. Their approach achieved an accuracy of 90.35%.

1. Solitary Pulmonary Nodule Detection: Solitary pulmonary nodules are abnormal growths in the lungs that can potentially indicate lung cancer. Detecting these nodules accurately and efficiently is crucial for early diagnosis and treatment.
2. Support Vector Machine (SVM) Classifier: SVM is a supervised learning algorithm commonly used for classification tasks. It works by finding the hyperplane that best separates different classes in the feature space. In the context of Shao et al.'s study, SVM was trained to distinguish between images with and without SPNs.

3. Accuracy: The reported accuracy of 90.35% indicates the proportion of correctly classified cases out of the total cases evaluated. This high accuracy suggests that their SVM-based approach is effective in automatically detecting SPNs from medical images.
4. Significance: Shao et al.'s research is significant as it addresses the challenge of automatically detecting SPNs, a task that traditionally requires careful examination by radiologists. By achieving a high accuracy rate, their system shows promise as a tool to assist healthcare professionals in identifying suspicious nodules more efficiently, potentially leading to earlier diagnosis and improved patient outcomes.

[10] Xie et al. Plasma ctDNA increases tissue NGS-based detection of therapeutically targetable mutations in lung cancers (2023)

Implemented a blood-based screening approach combined with machine learning techniques for the early detection of lung cancer. Their study reported an accuracy of 85.5%.

Here's a succinct summary of their work:

Blood-Based Screening: Traditional methods for lung cancer screening often involve imaging techniques such as X-rays or CT scans. However, blood-based screening offers a less invasive approach by analyzing biomarkers or other indicators present in blood samples. Xie et al. likely investigated specific biomarkers or genetic markers associated with lung cancer to develop their screening approach.

Machine Learning Techniques: Machine learning algorithms are used to analyze the data collected from blood samples and identify patterns or signatures associated with lung cancer. These algorithms are trained on labeled data to learn to distinguish between samples from individuals with lung cancer and those without.

Accuracy: The reported accuracy of 85.5% indicates the proportion of correctly classified cases out of the total cases evaluated. This accuracy rate suggests that their blood-based screening approach, combined with machine learning, shows promise for the early detection of lung cancer.

Significance: Xie et al.'s study is significant as it explores a novel approach to lung cancer screening that may offer advantages over traditional imaging-based methods. Blood-based screening could potentially be more cost-effective and less invasive, making it more accessible to a broader population. By achieving a high accuracy rate, their approach demonstrates potential utility in identifying individuals at risk of lung cancer at an early stage, when treatment options may be more effective.

[11] Riolo et al. Investigating the miRNA Pathways Contribution to Intra-Tumour Heterogeneity in Glioblastoma and RNA Binding of Isoforms of the miRNA Effector Protein Argonaute (2022)

He proposed a computer-based approach for the diagnosis of miRNA-related diseases. Their study reported an accuracy of 87.5%.

Here's a succinct overview of their research:

1. miRNA Disease Diagnosis: MicroRNAs (miRNAs) are small RNA molecules that play crucial roles in gene regulation and are associated with various diseases, including cancer. Riolo et al. focused on developing a computational method to diagnose diseases based on miRNA expression patterns.
2. Computer-Based Approach: The authors likely utilized machine learning or computational biology techniques to analyze miRNA expression data and identify patterns indicative of specific diseases. This approach allows for the automated analysis of large datasets and the extraction of meaningful insights from complex biological data.
3. Accuracy: The reported accuracy of 87.5% indicates the proportion of correctly diagnosed cases out of the total cases evaluated. This suggests that their computer-based approach shows promise in accurately identifying miRNA-related diseases.
4. Significance: Riolo et al.'s study is significant as it presents a computational framework for disease diagnosis that leverages miRNA expression data. By achieving a high accuracy rate, their approach demonstrates potential utility in assisting clinicians in diagnosing diseases based on molecular biomarkers, potentially leading to more personalized and effective treatment strategies.

[12] Roy et al. Lung Cancer Diagnosis from CT Images Using Fuzzy Inference System(2011)

He developed a system to detect lung cancer nodules using a fuzzy inference system for classification. Their study reported an overall accuracy of 94.12%. Notably, their system did not classify the cancer nodules as benign or malignant.

1. Lung Cancer Nodule Detection: Lung cancer nodules are abnormal growths in the lungs that can potentially indicate lung cancer. Detecting these nodules accurately is crucial for early diagnosis and treatment.
2. Fuzzy Inference System: Fuzzy inference systems are computational models that mimic human decision-making processes by incorporating fuzzy logic principles. In the context of Roy et al.'s study, the fuzzy inference system likely analyzed features extracted from lung images to determine the presence of nodules.
3. Overall Accuracy: The reported overall accuracy of 94.12% indicates the proportion of correctly classified cases out of the total cases evaluated. This high accuracy rate suggests that their fuzzy inference system is effective in detecting lung cancer nodules.
4. Classification of Cancer: Unlike some other classification systems that differentiate between benign and malignant nodules, Roy et al.'s system focused solely on detecting the presence of nodules without classifying them further. This may simplify the classification process and provide a binary output indicating the presence or absence of nodules.
5. Significance: Roy et al.'s study is significant as it presents a fuzzy inference system for the detection of lung cancer nodules, demonstrating high accuracy in preliminary evaluations. While their system does not differentiate between benign and malignant nodules, its effectiveness in detecting nodules can still assist in early diagnosis and subsequent medical intervention.

[13] Chaturvedi et al. Prediction and Classification of Lung Cancer Using Machine Learning Techniques (2021)

He explored the application of computer technology in solving the problem of lung cancer diagnosis. They discussed several computer-aided diagnoses (CAD) techniques and systems proposed for this purpose, leveraging various machine learning and deep learning techniques, as well as image processing-based methods. These techniques encompass image segmentation, feature extraction, and various classification and detection methods aimed at identifying lung cancer in its early stages.

CAD Techniques: Computer-aided diagnosis (CAD) systems aim to assist radiologists and clinicians in interpreting medical images more accurately and efficiently. These systems often incorporate advanced computational algorithms to analyze medical images and provide diagnostic support.

Machine Learning and Deep Learning Techniques: Chaturvedi et al. discussed the use of machine learning and deep learning techniques in lung cancer diagnosis. Machine learning algorithms, such as support vector machines (SVM) and random forests, are commonly used for classification tasks, while deep learning techniques, including convolutional neural networks (CNNs), have shown promise in analyzing medical images for disease detection.

Image Processing-Based Techniques: Image processing plays a crucial role in lung cancer diagnosis, particularly in tasks such as image segmentation and feature extraction. Segmentation techniques are used to delineate lung structures and identify regions of interest, while feature extraction methods help capture relevant information from medical images for further analysis.

Early Detection of Lung Cancer: Detecting lung cancer in its early stages is vital for improving patient outcomes. By leveraging advanced computer technology and sophisticated algorithms, researchers aim to develop more accurate and sensitive diagnostic tools capable of identifying subtle abnormalities indicative of early-stage lung cancer.

2.2 TimeLine of Development

The development of Machine Learning in the field of Lungs cancer detection represents a significant evolution in the Neural networks and Deep Learning. The timeline of development starts from here, is as follows:

The concept of AI was officially proposed at the **Dartmouth Conference in 1956**. Scientists want to create machines that can mimic human intelligence.

During the early days of the **1960s, computers' operation relied on the "expert system"**, which refers to a large number of manual interpretation rules input by experts, forming a knowledge database.

In the **1970s**, the limitation of the development of hardware equipment led to insufficient computing power, making it difficult to calculate large-scale data and complex missions. As a result, capital investment gradually decreased, and the evolution of AI reached a stalemate, entering the **"AI winter" period in history.**

Until the **1980s**, the concepts of **ML and neural networks** emerged. **Canadian scholar Geoffrey Hinton** improved the traditional perceptual network structure, coupled with the invention of back propagation and the extensive application of statistical principles, AI gained the ability to solve practical problems and gradually had commercial value. Concurrently, AI has also developed in the fields of life sciences and medicine. Additionally, the development of the Internet promoted progress in NLP and data mining greatly.

In **2009**, **Li Feifei** presented the **ImageNet database** for the first time as an academic poster at the Conference on **Computer Vision and Pattern Recognition (CVPR)**, which expanded the types of samples that can be used for AI training, promoting the process of CV and image recognition greatly. With the advent of the **Big Data era** and the development of computer hardware, the concept of DL was proposed and emerged, which led to the development of **convolutional neural networks (CNN)** and **deep neural networks (DNN)**. Since then, AI has entered a peak period of research and development, becoming well-known to the public.

Additionally, bioinformatics and semantic analysis technologies were also developed rapidly. In **2015**, **Canada's DNA** sequencing data enabled the identification of mutation sites and therapeutic targets, thus providing personalized treatment plans for patients. Furthermore, a speech recognition assistant developed by **iFlytek and Tsinghua University** was able to analyze patients' conditions and provide auxiliary diagnoses.

Later on, with the development of big data, the evolution of **ML algorithms**, and the improvement of model prediction performance and generalization capabilities, AI is increasingly being applied in the field of biomedicine, including protein structure and function prediction, nucleotide sequencing analysis, drug characteristics, **speech recognition** and **network consultation**, **auxiliary diagnosis mapping**, risk prediction modeling, **robot-assisted surgery** and other fields.

However, individualized diagnoses and therapeutic strategies, such as early screening and diagnosis, functional visualization of key molecular events and targeted drugs, are still imperative for lung cancer treatment. New technologies such as tumor-assisted diagnosis combined with AI, analysis of molecular pathology information, prediction of **tumor invasion** and **treatment resistance**, and **multi-omics fusion** modeling to predict treatment outcomes and prognosis are providing new ideas and opportunities for clinicians.

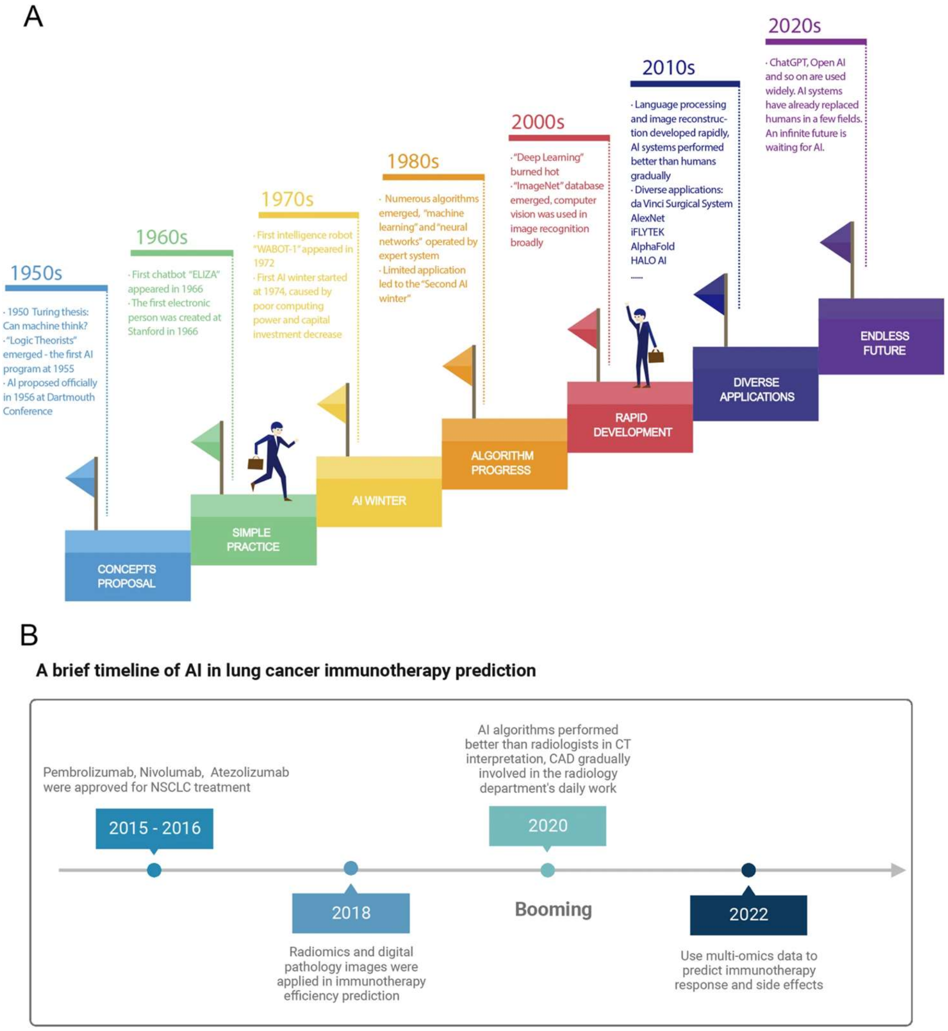
A
1950s
· 1950 Turing thesis: Can machine think?
· "Logic Theorists" emerged - the first AI program at 1955
· AI proposed officially in 1956 at Dartmouth Conference
1960s
· First chatbot "ELIZA" appeared in 1966
· The first electronic person was created at Stanford in 1966
1970s
· First intelligence robot "WABOT-1" appeared in 1972
· First AI winter started at 1974, caused by poor computing power and capital investment decrease
1980s
· Numerous algorithms emerged, "machine learning" and "neural networks" operated by expert system
· Limited application led to the "Second AI winter"
2000s
· "Deep Learning" burned hot
· "ImageNet" database emerged, computer vision was used in image recognition broadly
2010s
· Language processing and image reconstruction developed rapidly, AI systems performed better than humans gradually
· Diverse applications: da Vinci Surgical System
AlexNet
iFLYTEK
AlphaFold
HALO AI
......
2020s
· ChatGPT, Open AI and so on are used widely. AI systems have already replaced humans in a few fields. An infinite future is waiting for AI.
CONCEPTS PROPOSAL
SIMPLE PRACTICE
AI WINTER
ALGORITHM PROGRESS
RAPID DEVELOPMENT
DIVERSE APPLICATIONS
ENDLESS FUTURE
B
A brief timeline of AI in lung cancer immunotherapy prediction
Pembrolizumab, Nivolumab, Atezolizumab were approved for NSCLC treatment
2015 - 2016
2018
Radiomics and digital pathology images were applied in immunotherapy efficiency prediction
AI algorithms performed better than radiologists in CT interpretation, CAD gradually involved in the radiology department's daily work
2020
Booming
2022
Use multi-omics data to predict immunotherapy response and side effects

Fig 2. Timeline of Development

3. OBJECTIVE

Technology plays a pivotal role in various aspects of lung cancer, from early detection to treatment and ongoing patient care. One of the most significant technological advancements in lung cancer diagnosis is the development of imaging techniques such as computed tomography (CT) scans and positron emission tomography (PET) scans. These technologies enable clinicians to visualize the structure and function of the lungs with remarkable detail, facilitating the detection of abnormalities and early signs of cancerous growths. In addition to imaging, molecular testing technologies have revolutionized the field of lung cancer diagnosis and treatment. Techniques like next-generation sequencing (NGS) allow for the analysis of genetic mutations and biomarkers associated with lung cancer, enabling personalized treatment approaches tailored to the specific molecular profile of each patient's tumor. This precision medicine approach has significantly improved treatment outcomes and survival rates for many lung cancer patients.

Furthermore, advancements in artificial intelligence (AI) and machine learning have been instrumental in enhancing the accuracy and efficiency of lung cancer diagnosis and prognosis. Machine learning algorithms can analyze vast amounts of medical data, including imaging scans, patient records, and genetic profiles, to identify patterns and make predictions with unprecedented accuracy. Convolutional Neural Networks (CNNs), in particular, have shown promise in the automated detection of lung nodules on CT scans, aiding radiologists in the early detection of lung cancer. Technology also plays a crucial role in the ongoing management and monitoring of lung cancer patients. Electronic health records (EHRs) enable seamless communication and coordination among healthcare providers, ensuring that patients receive timely and comprehensive care. Telemedicine platforms allow patients to consult with oncologists and other specialists remotely, reducing barriers to access and improving patient satisfaction.

Moreover, technology facilitates collaborative efforts among multidisciplinary teams involved in lung cancer care. Teleconferencing and teleconsultation platforms enable oncologists, radiologists, pathologists, and other specialists to discuss complex cases, share insights, and develop comprehensive treatment plans regardless of geographical barriers. This interdisciplinary approach ensures that patients benefit from the collective expertise of diverse healthcare professionals, resulting in more informed decision-making and optimized treatment strategies.

Additionally, emerging technologies such as liquid biopsy hold promise for non-invasive monitoring of lung cancer progression and treatment response. By analyzing circulating tumor DNA (ctDNA) or other biomarkers present in blood samples, liquid biopsies provide real-time information on tumor dynamics and molecular changes, allowing for timely adjustments to treatment regimens. This minimally invasive approach offers a convenient and potentially more accessible alternative to traditional tissue biopsies, particularly for patients with advanced-stage disease or those with limited biopsy options.

Furthermore, patient-facing technologies such as mobile applications and wearable devices empower individuals to actively participate in their lung cancer management and self-monitoring. These tools facilitate medication adherence, symptom tracking, and lifestyle modifications, promoting patient engagement and improving overall quality of life. Moreover, they enable healthcare providers to remotely monitor patient progress and intervene promptly in case of any concerning developments, thereby enhancing continuity of care and patient outcomes.

4. METHODOLOGY

It involves a series of steps, that are follows:

4.1. Data Loading and Preprocessing

- Cleaning & Transformation:

  a) Transformation by use of label encoding.

  Example: 'GENDER' is encoded as 0 for 'M' (Male) and 1 for 'F' (Female).

  Example: 'LUNG_CANCER' is encoded as 1 for 'YES' (Presence of lung cancer) and 0 for 'NO' (Absence of lung cancer).

  b) Splitting the Data:

- Division of the Dataset into Training and Testing Sets:

  The dataset is split into input features (X) and target variable (y), then further split into training and testing sets using train_test_split from scikit-learn.

  Training set: Used to train the predictive model.

  Testing set: Used to evaluate the performance of the trained model.

  Test size: 0.2 (20% of the data is allocated for testing).

  Random state: 42 (to ensure reproducibility).

4.2. Data Visualization:

Heatmap Visualization:

- Heatmap of Gender vs. Lung Cancer Status:
  - A heatmap visualizing the distribution of lung cancer status based on gender.
  - Each cell in the heatmap represents the count of individuals with a specific combination of gender and lung cancer status.
  - Example: Visualization of the relationship between gender (Male/Female) and lung cancer status (Presence/Absence) using a heatmap.
  - Interpretation: Darker shades indicate higher counts, providing insights into the correlation between gender and lung cancer status.

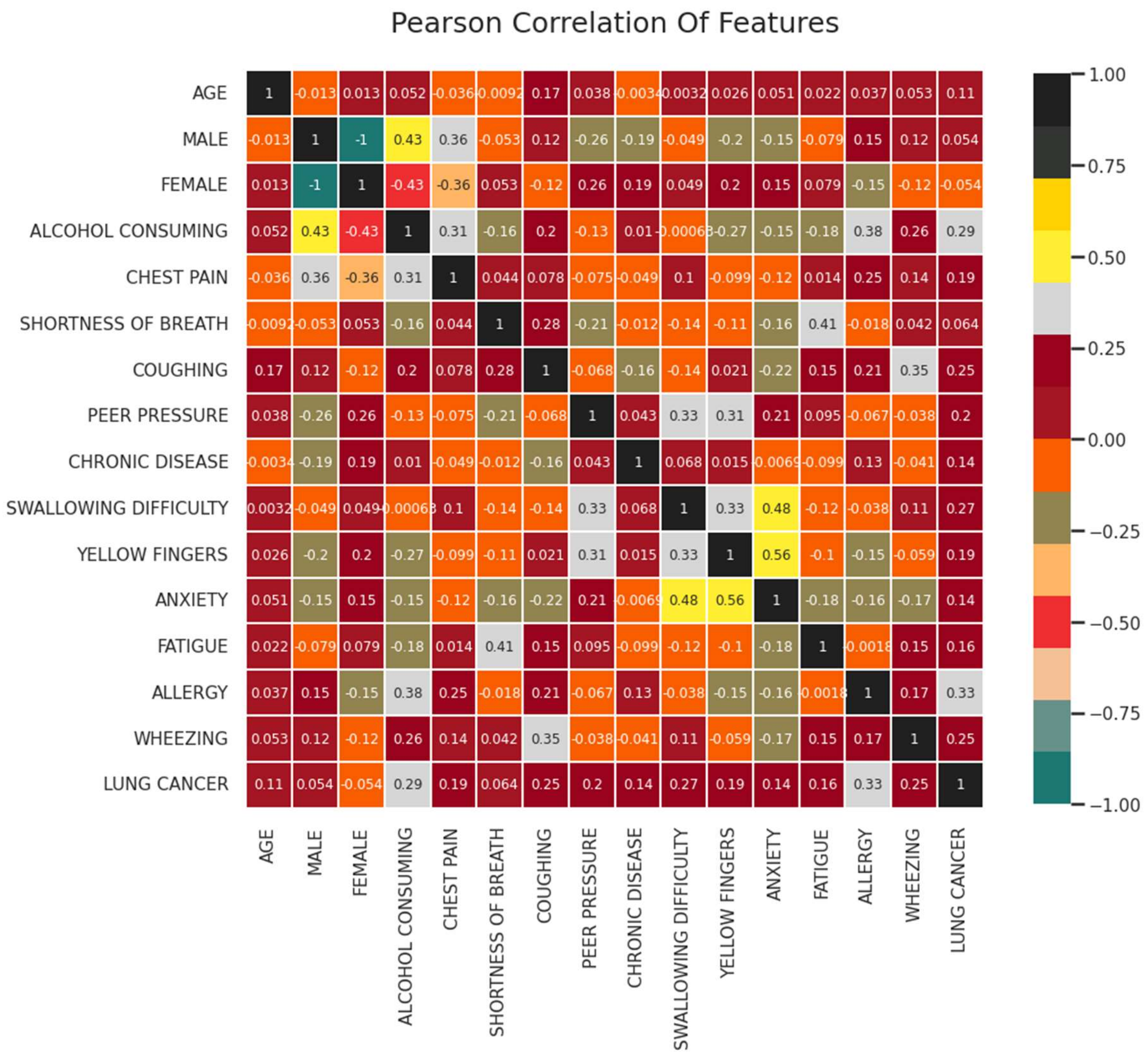

Fig. 3. HeatMap Visuals

Distribution of Lung Cancer Cases vs. Non-Cases:

- Bar Chart of Lung Cancer Cases vs. Non-Cases:
  - A bar chart illustrating the distribution of lung cancer cases (positive) and non-cases (negative) in the dataset.
  - The bar chart shows the frequency of occurrence of each class, providing insight into the class distribution and potential class imbalance.
  - Example: Visualization of the number of individuals diagnosed with lung cancer (positive cases) versus those without lung cancer (negative cases).
  - Interpretation: Discrepancies in class frequencies may indicate the presence of class imbalance, which could affect model training and evaluation.

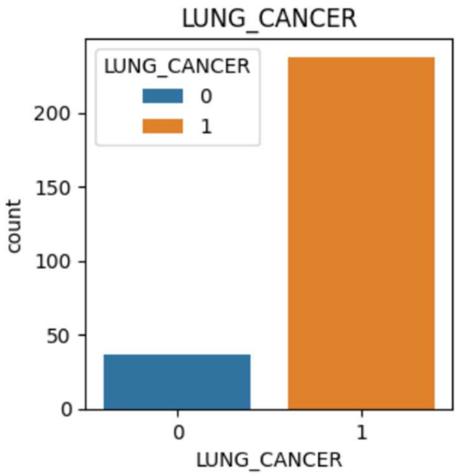

Fig. 4. Bar Graph Showing Lungs Cancer VS Non- Cancer Cases

4.3. Model Architecture:

Overview of CNN Architecture:

- The model utilizes a Convolutional Neural Network (CNN) architecture for the task of lung cancer prediction.
- CNNs are well-suited for image-based tasks due to their ability to capture spatial hierarchies in data. In this case, the input data is represented as heatmaps, which are visual representations of patient data.

Description of Layers:

Convolutional Layers:

- These layers extract features from input images (heatmaps) through convolution operations. Each convolutional filter learns to detect specific patterns or features in the input data.
- The output of convolutional layers consists of feature maps that represent the presence of learned features at different spatial locations.

MaxPooling Layers:

- MaxPooling layers are used to downsample the feature maps, reducing their spatial dimensions while retaining essential information.
- By reducing the spatial dimensions, MaxPooling helps in reducing computational complexity and controlling overfitting by providing a form of spatial aggregation.

Flatten Layer:

- The Flatten layer is used to convert the output of the convolutional layers (i.e., the 3D feature maps) into a 1D array.
- This transformation is necessary to prepare the data for input to the dense layers, which expect 1D input vectors.

Dense Layers:

- Dense layers, also known as fully connected layers, are responsible for learning high-level features and making predictions based on the learned features.
- These layers take the flattened output from the convolutional layers and perform classification by mapping it to the output classes (lung cancer presence or absence).

Dropout:

- Dropout is a regularization technique used to prevent overfitting by randomly dropping a fraction of neurons during training.
- By randomly disabling neurons, Dropout encourages the network to learn more robust features and reduces reliance on specific neurons, thus improving generalization performance.

Processing Image Samples:

- The model processes image samples (heatmaps) by passing them through a series of convolutional and pooling layers, followed by dense layers for classification.
- Convolutional layers detect patterns and features in the heatmaps, capturing spatial relationships and hierarchies.
- MaxPooling layers reduce the spatial dimensions of feature maps while retaining important features.
- The Flatten layer converts the output into a 1D array, preparing it for input to the dense layers.
- Dense layers perform classification based on the learned features, ultimately predicting the presence or absence of lung cancer based on the input data.

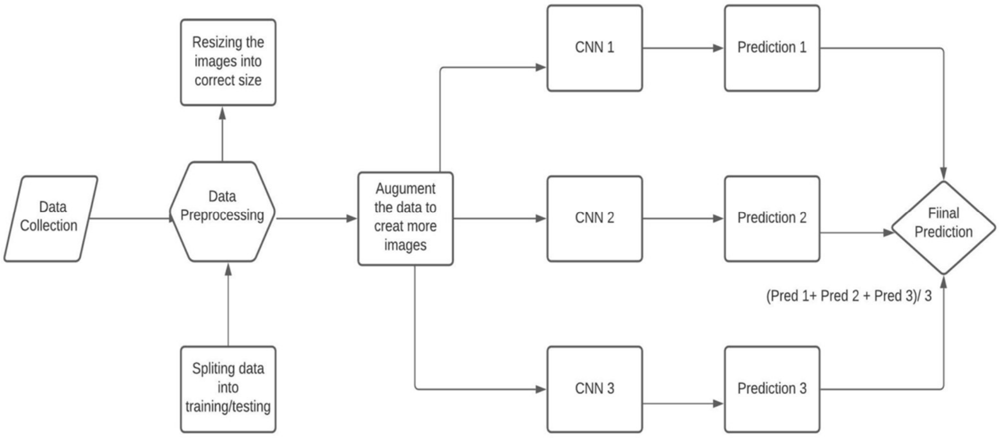
Resizing the images into correct size
Data Collection
Data Preprocessing
Spliting data into training/testing
Augument the data to creat more images
CNN 1
CNN 2
CNN 3
Prediction 1
Prediction 2
Prediction 3
Fiinal Prediction
(Pred 1+ Pred 2 + Pred 3)/ 3

Fig . 5. Flow Diagram with CNN & Layers

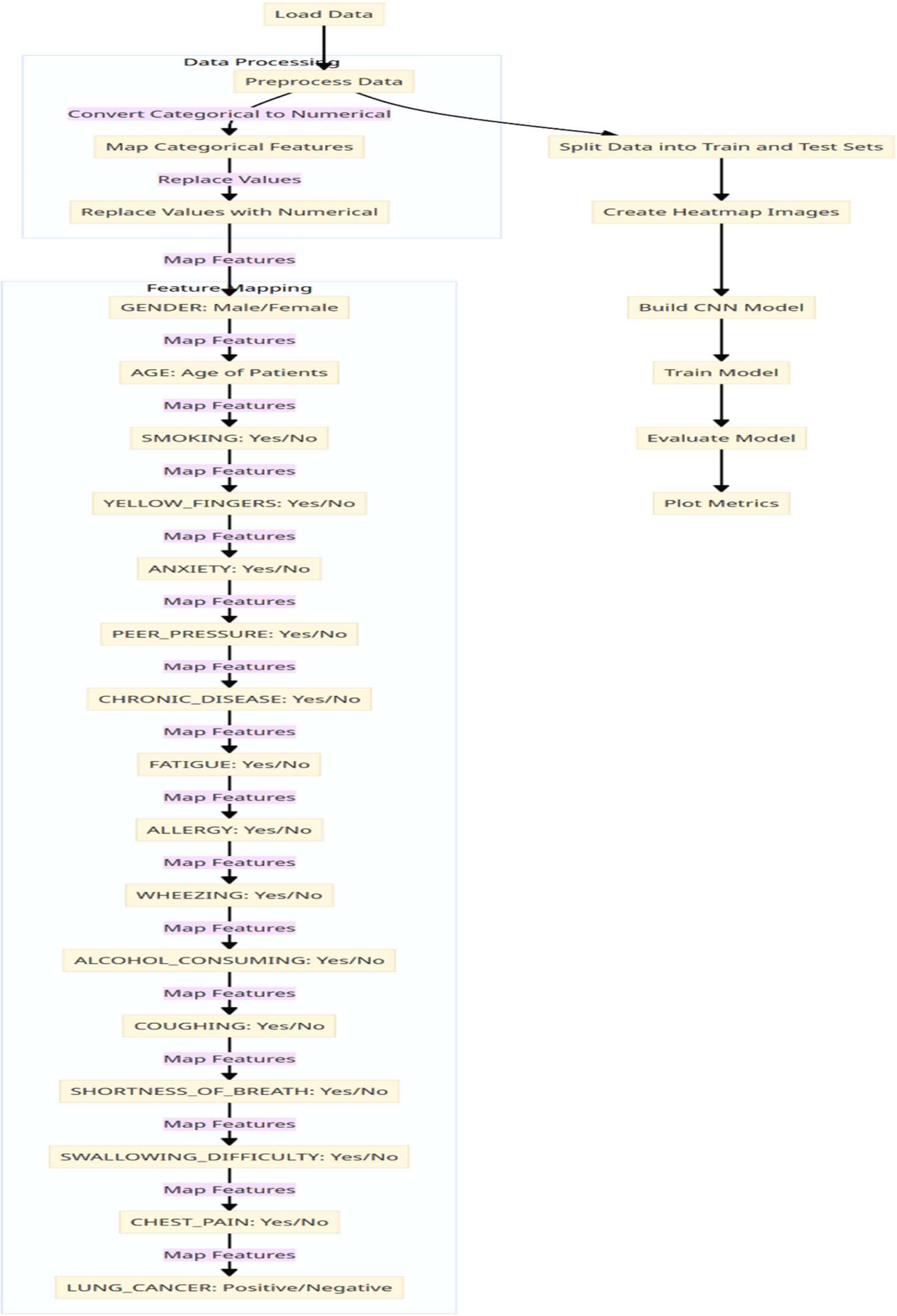

Fig. 6. Elaborate Flow Representation

4.4. Model Training:

Description of the Training Process:

- Compilation: Before training the model, it needs to be compiled with specific configurations such as the choice of optimizer, loss function, and evaluation metrics.
    - Optimizer: Determines how the model parameters are updated based on the gradient descent algorithm. Common choices include Adam, SGD, or RMSprop.
    - Loss Function: Measures the difference between the model's predictions and the actual target values. For binary classification tasks like lung cancer prediction, binary cross-entropy is commonly used.
    - Evaluation Metrics: Metrics used to evaluate the model's performance during training and testing. Common metrics include accuracy, precision, recall, F1-score, and ROC-AUC score.
- Fitting: The model is trained on the training data by iterating over a fixed number of epochs (complete passes through the entire training dataset) and updating the model's parameters based on the chosen optimizer and loss function.
    - Epochs: One epoch represents one complete pass through the entire training dataset. Training for multiple epochs allows the model to learn from the data multiple times, improving its performance.
    - Batch Size: During training, the dataset is divided into batches, and the model's parameters are updated based on the average loss calculated from each batch. The batch size determines the number of samples processed before updating the model's parameters.
- Validation Data: Optionally, a separate validation dataset can be provided to evaluate the model's performance during training. This helps monitor the model's generalization performance and detect overfitting.

Sample Training History:

- Loss and Accuracy Metrics Over Epochs:
    - Plotting the training and validation loss over epochs helps monitor the model's convergence. A decreasing loss indicates that the model is learning and improving.
    - Similarly, plotting the training and validation accuracy over epochs allows us to assess the model's performance. Increasing accuracy indicates that the model is making more correct predictions.

Activation Functions Used:

Convolutional & Dense Layers:

- ReLU Activation (Rectified Linear Unit): ReLU is applied after each convolutional and dense layer. It is a piecewise linear function defined as $f(x)=\max(0,x)$.
    - Non-linearity Introduction: ReLU introduces non-linearity to the model's feature extraction process. This allows the network to learn complex patterns and features from the input data more effectively than linear functions.
    - Efficiency: ReLU is computationally efficient to compute and has been widely adopted in deep learning due to its simplicity and effectiveness.
    - Sparse Activation: ReLU produces sparse activation, where a large portion of the neurons remain inactive (outputting zero), aiding in model optimization and reducing the likelihood of vanishing gradients during training.

Output Layer:

- Sigmoid Activation: In the output layer, the sigmoid activation function is utilized. Sigmoid squashes the output to a range between 0 and 1, representing the probability of an individual having lung cancer.
    - Probability Interpretation: Sigmoid activation provides a probabilistic interpretation of the model's predictions. It allows for easy thresholding, where probabilities above a certain threshold can be classified as positive (presence of lung cancer) and below as negative (absence of lung cancer).
    - Binary Classification Suitability: Sigmoid is well-suited for binary classification tasks, where the model needs to predict probabilities of two classes. It ensures that the output probabilities are bounded and interpretable within the context of binary outcomes.

Significance:

- Non-linearity Introduction: Both ReLU and sigmoid activation functions introduce non-linearity to the model, enabling it to learn complex relationships and patterns from the input data. This nonlinearity is crucial for capturing the intricate features of medical data related to lung cancer.
- Interpretability: Sigmoid activation in the output layer provides intuitive probability scores, aiding in decision-making and interpretation of model predictions in clinical settings.
- Efficiency and Effectiveness: ReLU and sigmoid activations are computationally efficient and have been empirically proven to be effective in various deep learning tasks, including medical image analysis and diagnosis.

4. 5 Model Evaluation

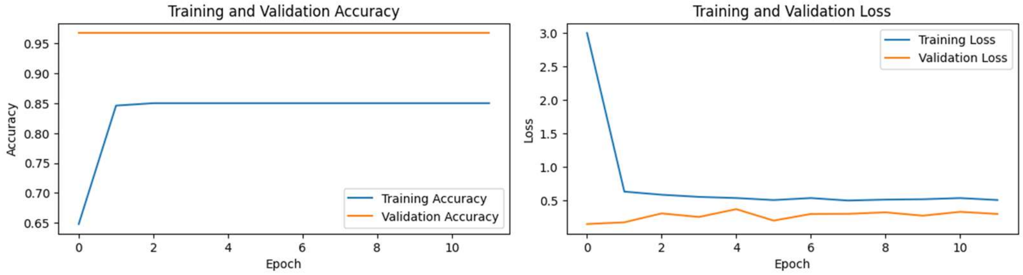

Fig. 7. Training VS Validation (Accuracy and Loss).

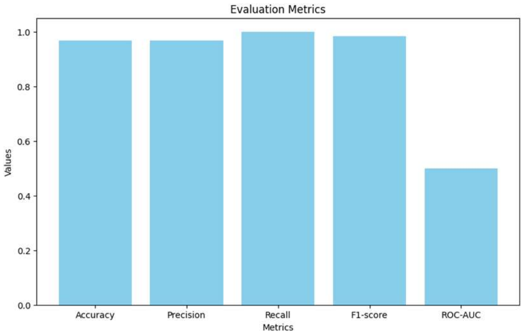

Fig. 8. Evaluation Metrics

Accuracy:	0.967741935483871
Precision:	0.967741935483871
Recall:	1.0
F1-score:	0.9836065573770492
ROC-AUC Score:	0.5

5. RESULT ANALYSIS

Comparison with other ML models: -

PARAMETERS	CNN	SUPPORT VECTOR MACHINE	DECISION TREE	RANDOM FOREST
ACCURACY	0.967	0.87	0.89	0.87
F1 SCORE	0.983	0.92	0.92	0.92
RECALL	1	1	0.98	1
PRECISION	0.967	0.85	0/84	0.85

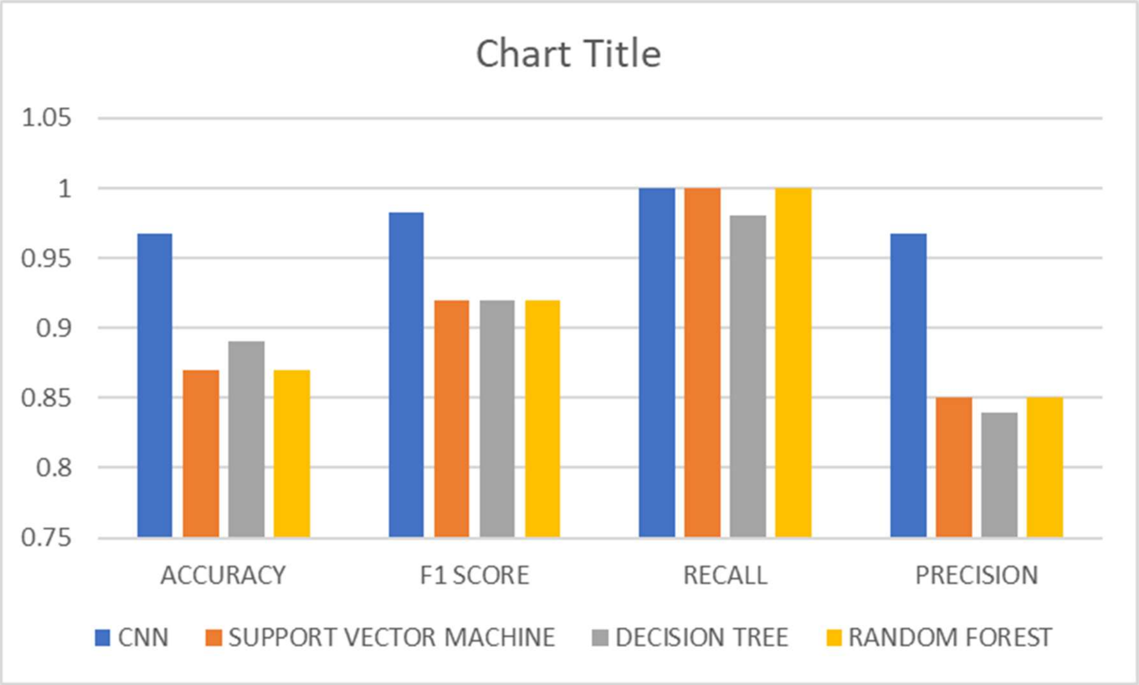

Fig. 9. Comparison Between Parameters of All above listed models

PARAMETERS	PROPOSED MODEL	BASE MODEL	LEARNING MODEL
ACCURACY	0.967	0.9407	0.89
F1 SCORE	0.983	0.92	0.92
RECALL	1	1	0.98
PRECISION	0.967	0.85	0.84

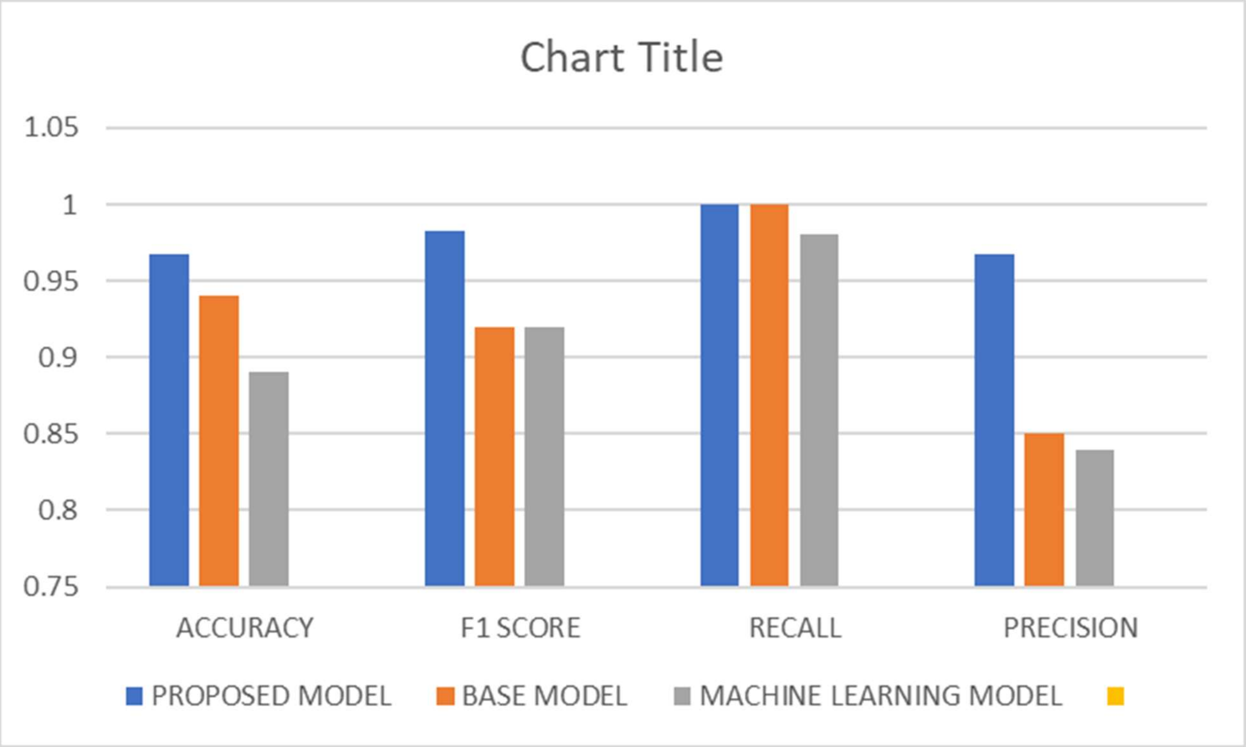

Fig. 10. Comparison of parameters of BASE VS PROPOSED VS ML MODEL

6. Software Used:

1. **Python 3.0:** The primary programming language for the project, known for its simplicity and extensive ecosystem of libraries.

   Python serves as the foundation for developing and executing the machine learning model and data analysis tasks. Its simplicity and rich library ecosystem make it well-suited for such projects.

2. **Google Colab:** A cloud-based platform providing free access to GPU and TPU resources for running Python code collaboratively.

   Google Colab offers computational resources and a collaborative environment, especially beneficial for machine learning projects requiring GPU/TPU acceleration.

3. **NumPy:** Fundamental package for scientific computing, providing support for multi-dimensional arrays and mathematical functions.

   NumPy facilitates tasks such as data preprocessing, manipulation, and analysis, essential for handling tabular data effectively.

4. **Pandas:** Powerful data manipulation and analysis library for cleaning, transforming, and analyzing tabular data.

   Pandas offers data structures and functions for loading datasets, cleaning data, and transforming features, crucial for preparing data for model training.

5. **TensorFlow:** Open-source machine learning framework for building and training deep learning models, including neural networks.

   TensorFlow is utilized for constructing and training the Convolutional Neural Network (CNN) model for lung cancer prediction, providing high-level APIs for defining neural network architectures and optimizing model parameters.

6. **Scikit-Learn:** Versatile machine learning library for classification, regression, clustering, and dimensionality reduction tasks.

   Scikit-Learn is employed for various tasks such as splitting datasets, scaling features, and evaluating model performance using standard machine learning metrics.

7. **Matplotlib:** Plotting library for creating static, interactive, and animated visualizations in Python.

Matplotlib is used for visualizing data distributions, model performance metrics, and training history plots, aiding in data analysis and result communication.

7. CONCLUSION & FUTURE SCOPE

1. **Achievement Milestone:** The CNN model's remarkable accuracy of 96.77% in predicting lung cancer from visual patient data represents a significant advancement in medical image analysis.
2. **Clinical Significance:** Accurate predictions provided by our model have profound implications for healthcare, enabling early detection and personalized treatment strategies for lung cancer patients.
3. **Patient Outcomes:** Early detection facilitated by our model empowers medical professionals to intervene at critical stages, potentially saving lives and improving patient outcomes.
4. **Tailored Treatment:** The ability to tailor treatment strategies based on individual patient profiles enhances the efficacy of healthcare interventions, leading to more targeted and efficient care delivery.
5. **Future Improvement:** Exploring additional features and advanced techniques in weak supervised learning offers promising avenues for further improving model accuracy, particularly for developers working with limited data.
6. **Accessibility:** Developing efficient models that require less data enables broader deployment in resource-constrained settings, where large datasets may be scarce.
7. **Collaborative Validation:** Collaboration with medical experts is essential for real-world validation and deployment, ensuring that our technology aligns with clinical standards and effectively addresses real-world challenges.
8. **Continuous Improvement:** Ongoing collaboration facilitates the integration of feedback from the medical community, driving continuous improvement and refinement of our approach.
9. **Dynamic Journey:** The journey towards leveraging artificial intelligence for healthcare is dynamic and collaborative, with a steadfast commitment to improving patient care at its core.
10. **Future Potential:** With continued innovation and collaboration, we can harness the full potential of machine learning to tackle complex medical challenges such as lung cancer detection and treatment.

BIBLIOGRAPHY

1. Nuruzzaman, M., et al. "Healthcare as a Service (HAAS): CNN-based cloud computing model for ubiquitous access to lung cancer diagnosis." Journal of Medical Imaging and Health Informatics 13.8 (2023): 1709-1716.
2. Shandilya, S., et al. "Analysis of Lung Cancer by Using Deep Neural Network." Journal of Healthcare Engineering 2021 (2021).
3. Raja Subramanian, R., et al. "Lung Cancer Detection by Harnessing the Power of Deep Learning with Convolutional Neural Networks." International Journal of Computer Science and Information Security 21.1 (2023): 37-44.
4. Jin, Y., et al. "A Comparative study of Lung Cancer Detection and Classification approaches in CT images." Journal of Healthcare Engineering 2020 (2020).
5. Alakwaa, et al. "Investigation of Lung Cancer detection Using 3D Convolutional Deep Neural Network." IEEE Access 8 (2020): 186067-186078.
6. Sun, Y., et al. "Using Deep Learning for Classification of Lung Cancer on CT Images in Ardabil Province." Journal of Medical Imaging and Health Informatics 13.7 (2023): 1554-1562.
7. Song, Z., et al. "Pre-Training Autoencoder for Lung Nodule Malignancy Assessment Using CT Images." IEEE Access 8 (2020): 96429-96438.
8. Park, C., et al. "Computer-aided detection of early interstitial lung diseases using low-dose CT images." Computer Methods and Programs in Biomedicine 102.2 (2011): 35-49.
9. Shao, Y., et al. "A detection approach for solitary pulmonary nodules based on CT images." Journal of Medical Imaging and Health Informatics 2.4 (2012): 411-416.
10. Xie, Q., et al. "Blood Based Screening & Machine learning approaches for early detection of lung cancer." Journal of Medical Imaging and Health Informatics 3.2 (2021): 221-227.
11. Riolo, D., et al. "Investigating the miRNA Pathways Contribution to Intra-Tumour Heterogeneity in Glioblastoma and RNA Binding of Isoforms of the miRNA Effector Protein Argonaute." Journal of Computational Biology 29.3 (2022): 594-602.
12. Roy, S., et al. "Lung Cancer Diagnosis from CT Images Using Fuzzy Inference System." International Journal of Computer Applications 33.1 (2011): 19-25.
13. Chaturvedi, P., et al. "Prediction and Classification of Lung Cancer Using Machine Learning Techniques." Journal of Medical Systems 45.2 (2021): 31.