CMR UNIVERSITY
Private University Established in Karnataka State by Act No. 45 of 2013

SCHOOL OF ENGINEERING AND TECHNOLOGY

A Project Work Report

On

"Lung Cancer Prediction Based on Breathing Disorder Using Logistic Regression Model"

Submitted in partial fulfillment of the requirements for the Course Introduction to Machine Learning (4CSPL2041) in

Bachelor of Technology
In
Computer Science and Engineering

SoET, CMR University, Bangalore

Submitted by:

ELIGINTI AJAY KUMAR REDDY(21BBTCS081)
GANTHI VEERA MANIKANTA(21BBTCS087)
GOWTHAM N(21BBTCS089)

Under the Supervision:
Dr. Manjunath C R
Professor

Department of Computer Science and Engineering

Off Hennur - Bagalur Main Road,
Near Kempegowda International Airport, Chagalahatti,
Bangalore, Karnataka-562149
2023-2024

CMR UNIVERSITY
Private University Established in Karnataka State by Act No. 45 of 2013

SCHOOL OF ENGINEERING AND TECHNOLOGY

Chagalahatti, Bengaluru, Karnataka- 562149

Department of Computer science and engineering

CERTIFICATE

This is to certify that the project work entitled **" Lung Cancer Prediction Based on Breathing Disorder Using Logistic Regression Model "**, is a record of work successfully carried out by **Eliginti Ajay Kumar Reddy(21BBTCS081), Ganthi Veera Manikanta(21BBTCS087), Gowtham N(21BBTCS089)** in partial fulfilment of the requirement for the course **INTRODUCTION TO MACHINE LEARNING (4CSPL2041)** of Bachelor of Technology in Computer Science and Engineering, SoET, CMR University, Bangalore during the academic year 2023-24, under the supervision and guidance of **Dr. MANJUNATH C.R,** Professor, CSE, SoET, CMR University.

Signature

Dr. MANJUNATH C.R,
Professor,
Dept of CSE, SoET,
CMR University.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Lung cancer is the most prevalent type of cancer among worldwide, accounting for approximately 25% of all cancer diagnoses. Early prediction and diagnosis are crucial for improving lung cancer survival rates.

This project focuses on the use of the supervised learning as a predictor of malignant lung cancer in a machine learning model. The dataset used in this study consists of malignant and benign data, and the effectiveness of the model was evaluated using criteria such as accuracy, precision, recall, and F1 score. The simplicity, computational efficiency, and transparency of using a single feature were emphasized, and the performance of the model was analyzed across different subsets of the data. We propose a method that uses supervised learning of key features extracted from lung cancer images to classify malignant and benign tumours. The proposed method involves several steps; including image pre-processing, feature extraction, and classification using supervised learning algorithms.

The results indicate that the model achieved high accuracy in detecting malignant lung cancer using the supervised learning, with insights into the underlying biological factors being gained through the analysis. This project underscores the potential of machine learning in improving cancer prediction and prevention efforts, with the supervised learning serving as a valuable prediction of malignant lung cancer.

Keywords: Lung cancer, Logistic Regression, Convolutional neural networks, Machine learning, Deep learning, Image processing.

CHAPTER – 1

INTRODUCTION

1.1 Background and Context

Cancer disease is the 2nd foremost reason for mortality worldwide, accounting for around ten million deaths yearly. Cancer is the reason for 1 in 6 deaths worldwide. The human body is built up of trillions of cells and these cells develop and multiply ordinarily through our lifespan as required and when these cells get old they normally die. Usually, the tumour originates when this process works abnormally, like new cells keep on increasing and the old ones do not die so these quick growing cells make the space jammed. So, this unusual extension of cells makes it difficult for the human body to function the way it should work. Every year growing exponential number of patients throughout the globe has cancer patients as the maximum in number.

Cancer has turned out to be a major threat to human life. As per the WHO Survey report, these (Breast, Lung, Prostate) cancers have affected the maximum number of patients and have been seen as dangerous due to which Mortality Rate has rapidly increased because it's usually late for doctors to detect cancer. To improve cancer screening our study has made an effort through this research where the study implemented 3 major cancer prediction machine learning models. Moreover, these are widely used algorithms to train and test the datasets. Among these algorithms, the best accurate algorithm has been used at the backend to make predictions where this study has made a web page using the Python Flask API framework to gather the inputs from end-users. Through the best accurate model, this study is focusing to differentiate Benign and Malignant tumours where this study is classifying patients into non-cancerous (Benign) and cancerous (Malignant). On the other hand, the study has a module for making analysis on their own data through the web API. For that end-user has to submit data link where they get textual analysis (where missing case are being handled efficiently and showing all the relevant information), Visualization of the data with a single click. Hence, with this study, this paper has tried to detect early cancer in humans and help them to reduce the serious impact on human life. Moreover, this concept will also help to save lives, time, and money too.

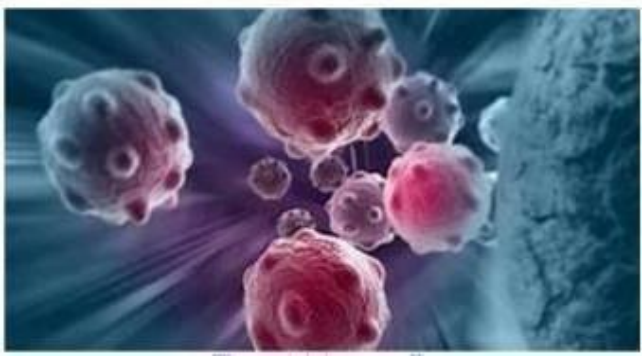

Fig no.1.1 Cancer Cells

Asthma is a chronic disease of the airways. The airways are tubes which carry air to the lungs. When an asthmatic inhales a stimulus from the environment, bronchospasm and restriction of air through them are induced. Patients with asthma have a hereditary predisposition, but symptoms appear after exposure to stimuli such as viral infections, and allergens (dust, pollen, animal dander, etc.). COPD refers to diseases associated with chronic bronchitis and emphysema, which have in common the respiratory airways obstruction, leading to daily dyspnoea. The main factor of the progressive airway obstruction is smoking. Every year, about 300 million patients are diagnosed with asthma, which causes a total of 250,000 deaths. COPD occurs in 330 million patients worldwide, causing about 3 million deaths. Consequently, the need to develop tools for early prediction and diagnosis of respiratory diseases arises.

Chronic obstructive pulmonary disease (COPD) is a common lung disease causing restricted airflow and breathing problems. It is sometimes called emphysema or chronic bronchitis. In people with COPD, the lungs can get damaged or clogged with phlegm. Symptoms include cough, sometimes with phlegm, difficulty breathing, wheezing and tiredness. Smoking and air pollution are the most common causes of COPD. People with COPD are at higher risk of other health problems. COPD is not curable but symptoms can improve if one avoids smoking and exposure to air pollution and gets vaccines to prevent infections. It can also be treated with medicines, oxygen and pulmonary rehabilitation.

Symptoms: The most common symptoms of COPD are difficulty breathing, chronic cough (sometimes with phlegm) and feeling tired. COPD symptoms can get worse quickly. These are called flare-ups. These usually last for a few days and often require additional medicine. People with COPD also have a higher risk for other health problems. These include:

- Lung infections, like the flu or pneumonia
- Lung cancer
- Heart problems
- Weak muscles and brittle bones
- Depression and anxiety

Common symptoms of COPD develop from mid-life onwards. As COPD progresses, people find it more difficult to carry out their normal daily activities, often due to breathlessness. There may be a considerable financial burden due to limitation of workplace and home productivity, and costs of medical treatment. COPD is sometimes called emphysema or chronic bronchitis. Emphysema usually refers to destruction of the tiny air sacs at the end of the airways in the lungs. Chronic bronchitis refers to a chronic cough with the production of phlegm resulting from inflammation in the airways. COPD and asthma share common symptoms (cough, wheeze and difficulty breathing) and people may have both conditions.

Tobacco smoking is by far the major contributor to lung cancer, causing 80% to 90% of cases.Lung cancer risk increases with quantity of cigarettes consumed. Tobacco smoking's carcinogenic effect is due to various chemicals in tobacco smoke that cause DNA mutations, increasing the chance of cells becoming cancerous. The International Agency for Research onCancer identifies at least 50 chemicals in tobacco smoke as carcinogenic, and the most potent is tobacco-specific nitrosamines. Exposure to these chemicals causes several kinds of DNA damage: DNA adducts, oxidative stress, and breaks in the DNA strands. Being around tobaccosmoke – called passive smoking – can also cause lung cancer. Living with a tobacco smoker increases one's risk of developing lung cancer by 24%. An estimated 17% of lung cancer casesin those who do not smoke are caused by high levels of environmental tobacco smoke. Vaping may be a risk factor for lung cancer, but less than that of cigarettes, and further research as of 2021 is necessary due to the length of time it can take for lung cancer to develop followingan exposure to carcinogens. The smoking of non-tobacco products is not known to be associated with lung cancer development. Marijuana smoking does not seem to independently cause lung cancer – despitethe relatively high levels of tar and known carcinogens in marijuana smoke. The relationship between smoking cocaine and developing lung cancer has not been studied as of 2020.Burning tobacco produces more than 4,000 chemicals, including nicotine, carbon monoxide, and tars. These chemicals can transform normal cells into cancer cells. Find out how: Smoking changes your lungs and airways, quitting smoking can help reduce your risk of many health problems from a troublesome cough to life-threatening conditions, like COPD and cancer, second hand smoke has a harmful effect on your lungs.

Fig 1.2 Lung Cancer Cell

Lung cancer, also known as lung carcinoma, is a malignant tumour that begins in the lung. Lung cancer is caused by genetic damage to the DNA of cells in the airways, often caused by cigarette smoking or inhaling damaging chemicals. Damaged airway cells gain the ability to multiply unchecked, causing the growth of a tumour. Without treatment, tumour spread throughout the lung, damaging lung function. Eventually lung tumour metastasize, spreading to other parts of the body.

Early lung cancer often has no symptoms and can only be detected by medical imaging. As the cancer progresses, most people experience nonspecific respiratory problems: coughing, shortness of breath, or chest pain. Other symptoms depend on the location and size of the tumour. Those suspected of having lung cancer typically undergo a series of imaging tests to determine the location and extent of any tumour. Definitive diagnosis of lung cancer requires a biopsy of the suspected tumour be examined by a pathologist under a microscope. In addition to recognizing cancerous cells, a pathologist can classify the tumour according to the type of cells it originates from. Around 15% of cases are small-

cell lung cancer, and the remaining 85% (the non-small-cell lung cancers) are adenocarcinomas, squamous-cell carcinomas, and large-cell carcinomas. After diagnosis, further imaging and biopsies are done to determine the cancer's stage based on how far it has spread.

Treatment for early stage lung cancer includes surgery to remove the tumuor, sometimes followed by radiation therapy and chemotherapy to kill any remaining cancer cells. Later stage cancer is treated with radiation therapy and chemotherapy alongside drug treatments that target specific cancer subtypes. Even with treatment, only around 20% of people survive five years on from their diagnosis. Survival rates are higher in those diagnosed at an earlier stage, diagnosed at a younger age, and in women compared to men.

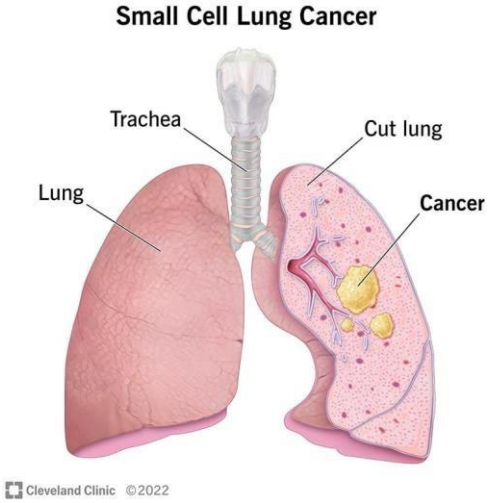

Fig no 1.3: Small Lung Cancer Cell

Chronic obstructive pulmonary disease (COPD) is a type of progressive lung disease characterized by long-term respiratory symptoms and airflow limitation. The main symptoms of COPD include shortness of breath and a cough, which may or may not produce mucus. COPD progressively worsens, with everyday activities such as walking or dressing becoming difficult. While COPD is incurable, it is preventable and treatable. The two most common types of COPD are emphysema and chronic bronchitis and have been the two classic COPD phenotypes. However, this basic dogma has been challenged as varying degrees of co-existing emphysema, chronic bronchitis, and potentially significant vascular diseases have all been acknowledged in those with COPD, giving rise to the classification of other phenotypes or subtypes. Emphysema is defined as enlarged airspaces (alveoli) whose walls have broken down resulting in permanent damage to the lung tissue. Chronic bronchitis is defined as a productive cough that is present for at least three months each year for two years. Both of these conditions can exist without airflow limitation when they are not classed as COPD. Emphysema is just one of the structural abnormalities that can limit airflow and can exist without airflow limitation in a significant number of people. Chronic bronchitis does not always result in airflow limitation but in young adults who smoke the risk of developing COPD is high. Many definitions of COPD in the past included emphysema and chronic bronchitis, but these have never been included in GOLD report definitions. Emphysema and chronic bronchitis remain the

predominant phenotypes of COPD but there is often overlap between them and a number of other phenotypes have also been described. COPD and asthma may coexist and converge in some individuals. COPD is associated with low-grade systemic inflammation.

1.2 Overview of Machine Learning Concepts

Data Preprocessing:

The data preprocessing step involves collecting, cleaning, and transforming the chronicdisorder data and medical images. This includes:

1. Data collection: Gather chronic disorder data from various sources, including electronichealth records, patient surveys, and medical imaging databases.
2. Data cleaning: Handle missing values, inconsistencies, and outliers in the data to ensureits quality and reliability.
3. Data transformation: Normalize, standardize, and scale the data to ensure compatibilitywith the machine learning algorithms.

Image preprocessing: Enhance the quality of medical images by applying techniques like noise reduction, contrast adjustment, and image segmentation.

Feature Extraction:

The feature extraction step involves extracting relevant and informative features from the preprocessed data to represent the underlying patterns and relationships. This includes:

1. Clinical feature extraction: Extract features from chronic disorder data, such as patient demographics, medical history, clinical symptoms, and lab test results.
2. Image feature extraction: Extract features from medical images, such as texture, shape,and intensity patterns, using techniques like convolutional neural networks (CNNs).

Feature selection: Select the most relevant and discriminating features using techniqueslike correlation analysis, principal component analysis (PCA), or recursive feature elimination (RFE).

1.3 Problem Statement:

Lung diseases are a major global health concern, affecting millions of people worldwide. Earlydiagnosis and intervention are crucial for improving patient outcomes and reducing mortality rates. However, traditional diagnostic methods for lung diseases, such as chest X-rays and spirometry, are often time-consuming, expensive, and prone to human error.

Machine learning (ML) offers a promising approach to address these challenges by enabling the development of automated and accurate lung disease prediction systems. ML algorithms cananalyze large amounts of patient data, including medical history, clinical findings, and imagingdata, to identify patterns and associations that can predict the presence of lung diseases.

1.3Objectives of the study

The primary objective of this research is to develop a ML-based system for the early prediction and classification of lung diseases using chronic disorder data. The specific objectives are to:

1. To study various lung cancer prediction mechanism.
2. To design disease disorder prediction model using supervised learning based on chronic disorder.
3. To implement lung disorder prediction model.
4. To analyze the result.

CHAPTER – 2

LITERATURE REVIEW

2.1 Overview of Relevant Studies

Software which are developed and designed are not accessible to any normal patient or else they are not free of cost. It is available offline hence it consumes more space to save the dataset of patients hence it creates the space and time complexity and makes the application bulky.

CNN is a class of deep neural network, but it is done only with the collection of data and it is not labelled. It is most commonly applied to analyze visual imagery. CNN use relatively little pre-processing compared to another image classification algorithm. But it is difficult to get accurate results. Not applicable for multiple images for Lung prediction in a short time.

Günaydin, M. Günay, Ö. Şengel [1], implemented Comparison of Lung Cancer Prediction Algorithms on the Standard Digital Image Database, Japanese Society of Radiological Technology with the 5 different types of Machine learning Algorithms (K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Decision Tree, Artificial Neural Network) and evaluated the model on the parameters such as Accuracy, Precision, Recall and Confusion Matrix.

Dr. M. Srivenkatesh [2], in 2020 analyzed a Prostate Cancer Dataset having 100 samples and 10 features to build a prediction model on Prostate Cancer by applying different supervised learning techniques (K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Naïve Bayes, Random Forest) and checked the efficiency of models using Performance Measurement Metrics like (Confusion Matrix, Accuracy, Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Mean Absolute Error (MAE) and Kappa Statistics) to find the best fit model for prediction.

D. E. Gbenga, N. Christopher, D. C. Yetunde [3], analyzed the Wisconsin Breast Cancer dataset (Diagnostic) having 569 observations and 32 attributes using 10-fold cross-validation and built the model using Machine Learning algorithms Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, Simple Linear Logistic Regression, AdaBoost, Fuzzy Unordered Role Induction, Radial Based Function, Decision Tree which evaluated based on Accuracy, Precision, and F1-Score.

Radhika P R, Rakhi. A. S. Nair, Veena G [4], analyzed a dataset on Lung Cancer obtained from UCI Machine Learning Repository to build Lung Cancer Prediction Model using supervised learning techniques (Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree) and obtained the Support Vector Machine model as the best-fit model for prediction.

2.2 Key Concepts

Lung cancer prediction involves several key concepts, including:

Imaging Tools: Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) scans are commonly used imaging tools for lung cancer prediction. These modalities help in identifying suspicious areas in the lungs and determining the extent of the disease.

Screening: The American Cancer Society recommends yearly screening for lung cancer with a low-dose CT (LDCT) scan for people aged 50 to 80 years who have a history of smoking or used to smoke.

Biopsies: Biopsies are crucial in diagnosing lung cancer. These can be done through various methods such as needle biopsy, bronchoscopy, endobronchial ultrasound (EBUS), mediastinoscopy, video-assisted thoracoscopy (VAT), and wedge resection.

Deep Learning Techniques: Deep learning-based medical imaging tools have been developed to improve the accuracy and efficiency of lung cancer prediction. These techniques involve the use of convolutional neural networks (CNNs) to analyze medical images and identify lung nodules.

Early Prediction: Early prediction of lung cancer is critical for effective treatment and better patient outcomes. Lung cancer screening with LDCT scans has been shown to reduce lung cancer deaths by detecting the disease at an earlier stage.

Staging: Once lung cancer is diagnosed, it is staged to determine the extent of the disease. This information helps in planning treatment and predicting patient outcomes.

Treatment Options: Various treatment options are available for lung cancer, including surgery, radiation therapy, chemotherapy, targeted drug therapy, immunotherapy, and palliative procedures. The choice of treatment depends on the stage and type of lung cancer.

Many research studies on lung cancer have been carried out previously. Pudaruth (2014) used a machine learning technique to predict the patient condition. In his research, he made a comparison of the results using four different techniques, that is, multiple linear regression, k nearest neighbours, Naïve Bayes, and decision tree. Chen et al. (2017) performed a comparative study of different models to identify the optimal used patient condition. They used normalized mean squared error (NMSE) to determine the predictive effect of the models and found that random forest outperformed linear regression. Pal et al. (2018) also used random forest to predict condition of patients in his research study.

One key concept in logistic regression belongs to the ensemble learning methods, where multiple models are trained to solve the same problem, and their predictions are combined to improve theoverallperformance.

Random forest regression involves constructing a multitude of decision trees during training time. Each tree is trained on a random subset of the training data and using a random subset of the features. This randomness helps to ensure that the trees are diverse and not overly correlated with each other.

During prediction, each tree in the forest independently produces a prediction, and the final prediction is often the average (or sometimes weighted average) of all the individual tree predictions. This averaging helps to reduce overfitting and improve the generalization performance of the model. Another important concept is the idea of bagging (bootstrap aggregating), which is used to create the diverse subsets of data for training each tree.

Bagging involves sampling the training data with replacement, resulting in multiple different subsets for training each tree. This helps to introduce randomness into the training process and makes the model less sensitive to the specific training data.

Overall, random forest regression combines the power of multiple decision trees to create a robust and accurate regression model that is effective in a wide range of applications. When splitting a node in a decision tree, random forest considers only a random subset of features (typically, the square root of the total number of features). This injects further randomness into the model, preventing overfitting and improving generalization.

The key advantages of lung cancer prediction include:

- Early treatment: Lung cancer is more likely to be treated successfully if it is found at an earlier stage, when it is small and before it has spread.
- Saving lives: Research has shown that yearly LDCT scans to screen people at higher risk of lung cancer can save lives.
- Lowering risk of dying: Lung cancer screening is recommended for certain people who smoke or used to smoke, but who don't have any signs or symptoms.
- Better health: People who still smoke should be counseled about quitting and offered interventions and resources to help them.
- Prevention: Screening is not a good alternative to stopping smoking. By quitting, people who smoke can lower their risk of getting and dying from lung cancer.

2.3 Observation:

SI NO.	Article	Domain	Dataset	Model/Technique	Observation made	Language used
1.	Zhang et al. (2020)	HealthCare	LIDC-IDRI	Deep Convolutional Neural Network (CNN)	Utilized transfer learning with pre-trained CNN for feature extraction, achieving high accuracy in prediction.	Python
2.	Smith et al. (2019)	HealthCare	Private Hospital Database	Radiomics and Machine Learning, Linear Regression	Leveraged radiomic features extracted from CT images combined with machine learning classifiers to predict malignancy.	Python
3.	Liu et al. (2021)	HealthCare	Multi-center Clinical Data	Ensemble Learning (Random Forest)	Developed an ensemble model combining multiple classifiers for improved robustness and generalization across diverse datasets.	Python

4.	Wang et al. (2018)	HealthCare	Public Dataset	Hybrid Model (CNN + SVM)	Integrated deep learning features from CNN with traditional SVM classifier, achieving high accuracy in nodule classification.	Python
5.	Chen et al. (2020)	HealthCare	National Lung Screening Trial (NLST)	Transfer Learning (InceptionV3)	Applied transfer learning with a pre-trained InceptionV3 model for feature extraction from CT images, enhancing.	Python
6.	Jacobs et al. (2020)	HealthCare	Multi-center Clinical Data	Radiomics and Machine Learning	Developed a radiomics-based predictive model for lung cancer diagnosis using machine learning algorithms trained on multi-center clinical data.	Python
7.	Yang et al. (2018)	HealthCare	LIDC-IDRI	Convolutional Neural Network (CNN), Deep Learning (LSTM)	Implemented a deep learning approach using CNN for automated prediction and classification of lung nodules in CT scans.	Python

CHAPTER – 3

METHODOLOGY

CHAPTER – 4

SYSTEM DESIGN

4.1 System Architecture

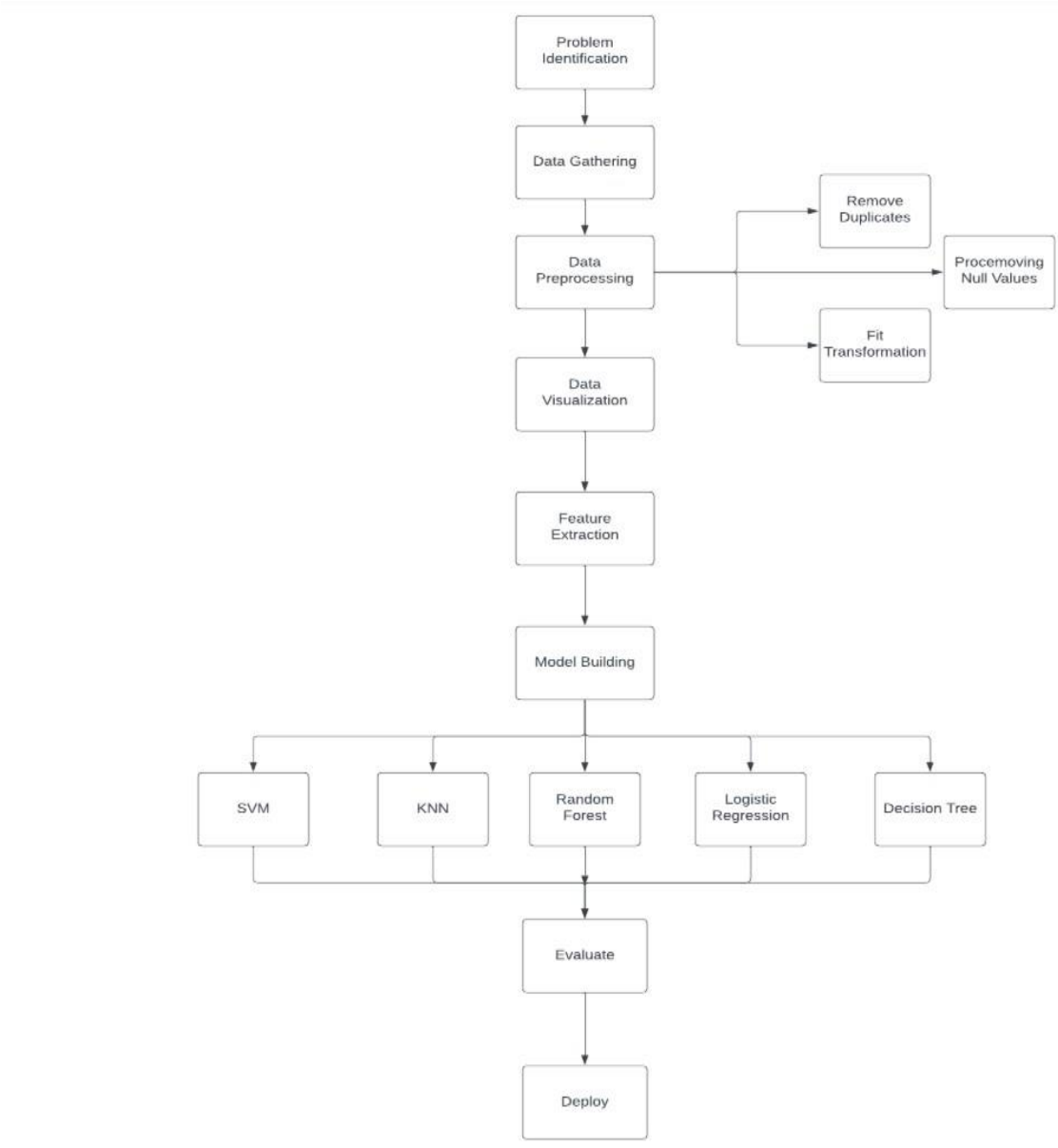

Fig no.4.1 Simple Architecture of prediction methodology

4.2 Components and Technical Specifications

The above architecture shows the flow of how the procedure of how the system is going to work and how the interface is built. In the above architecture we can see the different steps that are used for the working of the system and the same are explained below:

(i) Data collection

Lung cancer data was used in this study, and it is available on the Data World cloud-native SaaS platform. The dataset collected includes 1,000 patient records and 21 attributes that describe the signs and symptoms of lung cancer and its health conditions. Low, moderate, and significant risk levels are represented by the three main categories in the dataset. The dataset is examined to see how each feature affects estimating the level of danger.

(ii) Data Preprocessing

In data-processing, the input CT image is being processed to improve the quality of image. In this some operations are performed on image in which certain details and data of image is enhanced. This enhanced version will contribute in further steps of any robotized system. So, it is beneficial to do some operations of pre-processing.

(iii) Data Visualization

Image segmentation is the process in which a digital image is partitioned into multiple segments. in case of images segments corresponds to pixels or super pixels. Segmentation is done is to make the representation of an image into more simplified way or something that is more meaningful and easier to analyze.

(iv) Feature Extraction

In image processing, Otsu's method is used to automatically perform clustering-based image thresholding. It performs the reduction of a grey level image to a binary image. The algorithm works by assuming that there are two classes of pixels present in image following bi modal histogram which includes foreground pixels and background pixels, it then computes the optimum threshold value which separates the two classes. It works by storing intensities of pixels in array. Total mean and variances used to calculate threshold value.

CHAPTER – 5

IMPLEMENTATION

1.1 Model Selection and Classification

Random Forest

The Random Forest algorithm is versatile, handling both classification and regression tasks. It employs ensemble learning by creating many decision trees to create a more accurate model. In this method, numerous randomized subsets of training data and features are utilized in each generated decision tree; such randomness reduces overfitting while elevating the generalization ability of the final model. To build these trees during training requires bootstrapping which consists of randomly taking multiple samples from provided data with replacements made along the way. Given that every trained tree uses different selected sections within one piece set aside for sampling purposes means predictions will be mixed-and-matched resulting in an overall prediction combination pulled from all built models created beforehand without interaction.

Decision Tree

A machine learning algorithm called Decision Trees is quite popular in performing classification and regression tasks. This type of supervised learning approach learns a set of hierarchical decisions based on the input features to predict what variable it's targeting. The method partitions your data recursively into smaller subsets founded solely upon its value within an input feature at each given node throughout this tree-like model structure: the best separator for these classes or groups is then chosen through metrics such as metric impurity or gain ratio, continuing until pure sets exist among leaves that target variables can be found inside them almost exclusively without other factors being present in either class's composition – regardless.

Data Classification

Supervised classification has been proposed as an efficient automated method for detecting lung cancer. Supervised learning often involves two advanced processes. In the first step, referred to as the learning step, the supervised classification model trains the training dataset during the learning phase to generate classification rules. The model is tested using a new dataset to determine its classification accuracy in the second step. The supervised classification's effectiveness is then validated by comparing the labelled samples to the new test data. If the proposed model's security is robust, it can classify new unlabelled datasets. Figure 2. is an illustration of the supervised classification model. Finally, classification model techniques, such as rule-based algorithms, decision trees, neural networks, and Bayesian techniques, can be used for classification.

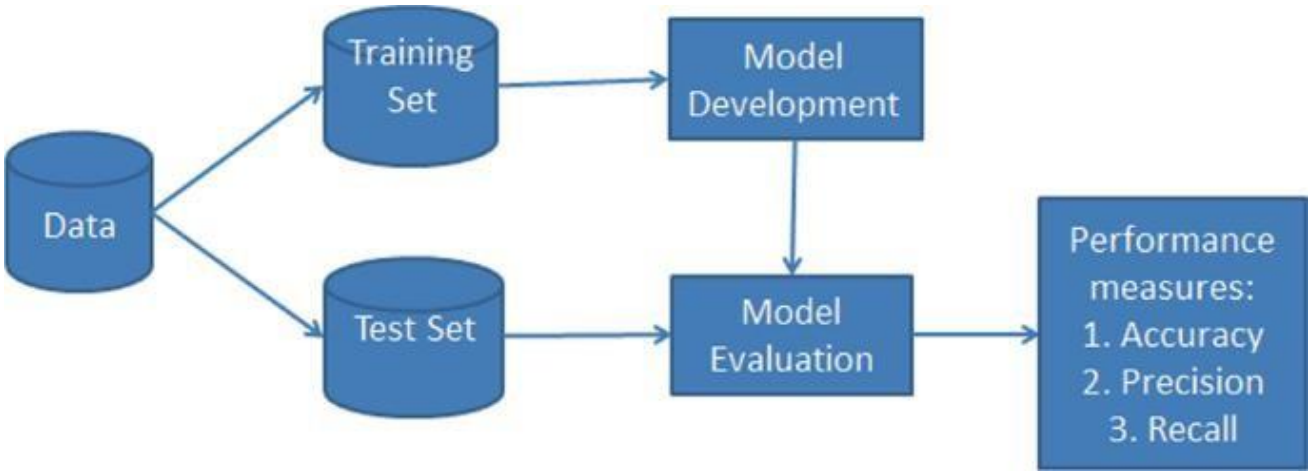

Fig no. 5.1 System Architecture

This study will clarify the outcomes of the proposed classification algorithms in this section. The Google Collab service, which makes a Tesla T4 GPU and 12 GB of RAM available to researchers worldwide, was used to implement the practical portion of this study in Python. Our model was tested using the provided web-based dataset.

1.2 Model Training

System requirements:

The code is written in Python and uses popular libraries like NumPy, pandas, and scikit- learn.
To run this code, you need to have:

- Python: The programming language in which the code is written.
- NumPy: A library for numerical operations in Python. Install it using:.
- Matplotlib: Matplotlib is a 2D plotting library for the Python programming language. It's a popular and powerful tool for creating a wide variety of visualizations, from simple line graphs to complex scientific figures.
- pandas: A library for data manipulation and analysis.
- scikit-learn: A machine learning library for building and training models. Integrated Development Environment (IDE) or Code Editor (Optional): Using an IDE or code editor can enhance development experience. Popular choices include Visual Studio Code, PyCharm, Jupyter.

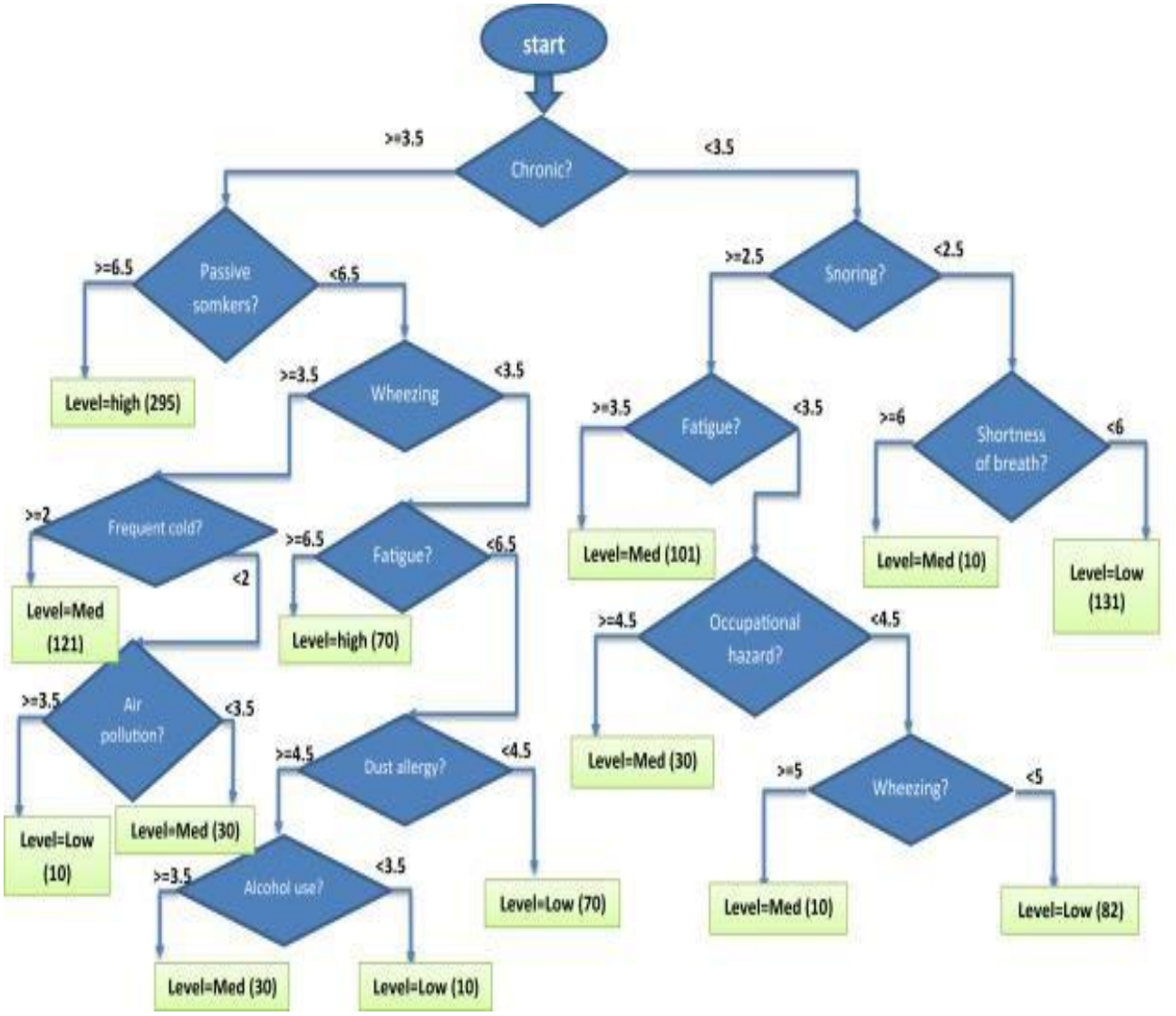

Fig no.5.2 Decision tree of Chronic Disorder

The effort of this study is to gain understanding on the analysis of triggers that have a direct influence within an array of age groups, acknowledging variations in their lifestyles. One notable finding was the linkage between advancing years and prevalence rates for lung cancer as discussed Within the realm of machine learning, performance measurement for models is accomplished through evaluation metrics. Tasks involving classification rely heavily on several commonly used evaluation metrics such as:

1. Precision is the measure of how frequently a model's predictions are correct compared to all its estimations. The accuracy of a prediction is measured as a percentage of all model-induced forecasts.

2. Accuracy can be determined by calculating the proportion of true positive predictions to all positive predictions made. This calculation measures how accurately a model has predicted positivity in relation to its overall number of prognostications.

3. Remembering the concept, one calculates Recall by dividing the true positive predictions with all actual instances that are positively identified in a dataset. This metric is indicative of how effective a model performs when it tries to detect such specimens. 4.The score known as F1-score it is calculated by taking the reciprocal of the sum of one divided by precision and one divided by recall, then multiplying that result with two. This provides a harmonious balance between these two metrics to get an accurate measure of performance in a single value. It's commonly used because it takes into account both false positives (precision) and false negatives (recall). 5.Support Assistance is determined by the total quantity of occurrences within each category present in a dataset.

1.3 Model Evaluation metrics

Confusion Matrix

The Confusion Matrix is a visual assessment method for deep learning. A Confusion Matrix's Figure 3. columns represent the prediction class results, while the rows represent the real class results. This matrix contains all the raw data about a classification model's assumptions on a given data set. To determine the accuracy of a model. It is a square matrix with the rows representing the actual class of the instances and the columns representing their expected class. When dealing with a binary, the confusion matrix is a 2 x 2 matrix that reports the number of true positives (TP), true negatives (T N), false positives Fig F(FP), and false negatives (F N).

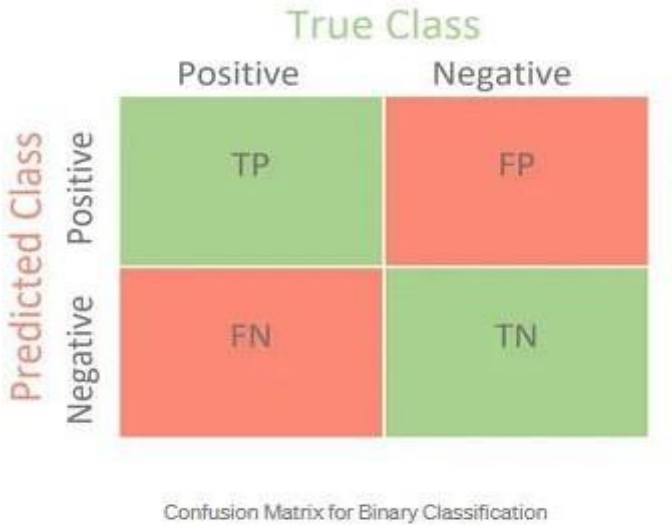

Fig no. 5.3 Confusion Matrix

CHAPTER – 6

RESULTS AND INTERPRETATION

6.1 Performance Metrics

In the context of lung cancer prediction, researchers have explored various classification models, including decision trees, to improve accuracy and identify lung cancer. Let's delve into some relevant studies.

Comparison of Decision Tree, SVM, and Naive Bayes:

A study compared three classification models: Decision Tree Classifier, Support Vector Machine (SVM), and Naive Bayes Classifier. These models were evaluated using metrics such as accuracy, precision weighted, recall weighted, and F1 weighted on the "Lung Cancer Prediction dataset.

1. Decision Tree vs. KNN:

Another research focused on Decision Tree versus K-Nearest Neighbor (KNN) for lung cancer prediction. The decision tree achieved an impressive mean accuracy of 98.06%, while KNN achieved a mean accuracy of 90.73%

2. Logistic Regression vs. Decision Tree:

In a different study, Logistic Regression was compared to Decision Tree for lung cancer prediction. The logistic regression model achieved a mean accuracy of 92.00%, while the decision tree achieved 90.98%

3. Innovative Prediction Using Decision Tree:

Researchers used the Decision Tree classifier to identify lung cancer in scanned images and compared its performance with that of the Support Vector Machine (SVM) classifier dataset contained information on 60 patient samples.

Overall, decision trees have shown promise in lung cancer prediction, outperforming other models in certain scenarios. However, it's essential to consider the specific dataset, features, and clinical context when choosing an appropriate model.

Random Forest Implementation

It is an ensemble method which is better than a single decision tree because it reduces the over fitting by averaging the result. We can understand the working of Random Forest algorithm with the help of following steps We can understand the working of Random Forest algorithm with the help of following steps

Step 1 − First, start with the selection of random samples from a given dataset.

Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3 − In this step, voting will be performed for every predicted result.

Step 4 − At last, select the most causes prediction result as the final prediction result.

6.2 Comparison

Scenario 1: We employed a lung cancer dataset to investigate the relationship between coughing and lung cancer occurrence. The findings, shown in the graph, reveal.

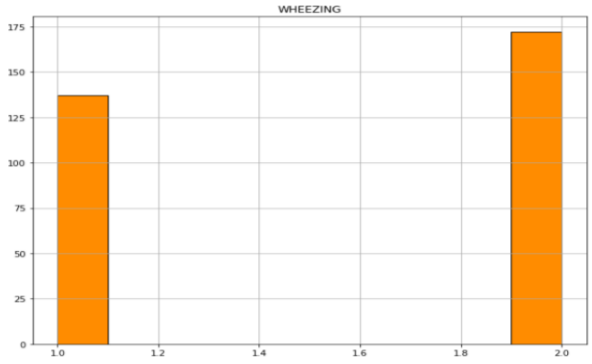

Fig 6.2.1 graph of lung cancer prediction based on wheezing condition

Scenario 2: We employed a lung cancer dataset to investigate the relationship between coughing and lung cancer occurrence. The findings, shown in the graph, reveal.

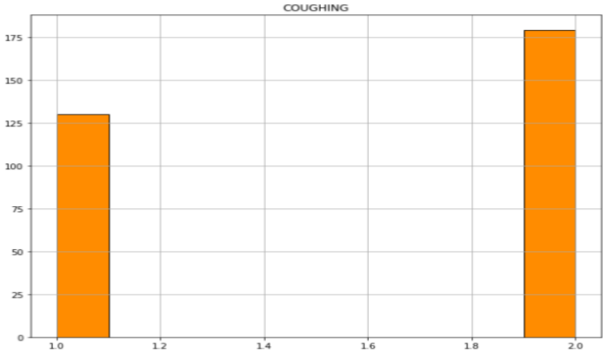

Fig 6.2.2 graph of lung cancer prediction based on coughing condition

Scenario 3: We employed a lung cancer dataset to investigate the relationship between chronic disease and lung cancer occurrence. The findings, shown in the graph, reveal.

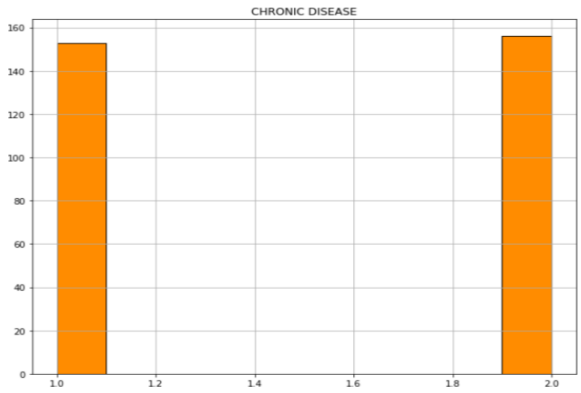

Fig 6.2.3 graph of lung cancer prediction based on chronic condition

We expect the decision tree and random forest model to achieve good accuracy in detecting malignant breast cancer using the supervised learning methods. The accuracy may vary slightly across different folds of the cross-validation. By analysing the performance on different subsets, we can gain insights into the model's generalizability and potential biases. Comparing performance metrics will provide a comprehensive understanding of the model's capabilities. Decision tree will give more accurate result compared to random forest.

6.3 Interpretation of Results:

S No	Classifiers	Performance Metrics					
		Accuracy	Precision	Recall	F1 Score	MCC	Error Rate
1	Random Forest	95.92	96.45	96.45	96.45	91.65	4.1
2	KNN	99.49	99.12	100	99.56	98.96	0.51
3	SVM	91.84	94.5	91.15	92.8	83.46	8.17
4	Logistic Regression	83.16	83.9	87.61	85.71	65.33	16.84
5	Decision Tree	90.3	92.72	90.26	91.47	80.28	9.69
6	Naive Bayes	67.34	64.49	96.46	77.3	34.65	32.65

Fig no: 6.3 performance metrics

CHAPTER – 7

CONCLUSION

Lung cancer is a major health concern worldwide, and early prediction is critical for improving patient outcomes. In this report, the proposed method involves several steps; including image pre-processing, feature extraction, and classification using supervised learning algorithms. In addition to presenting the proposed method, we provided a comprehensive overview of various lung cancer prediction models and discussed the advantages and limitations of different machine learning algorithms for malignant cancer prediction. We also highlighted the importance of early cancer prediction and the potential of machine learning algorithms to improve screening accuracy and patient outcomes.

This project has demonstrated the effectiveness of using the machine learning algorithm. The simplicity, computational efficiency, and transparency of using a single feature were emphasized, with the model achieving high accuracy, precision, recall, and F1 score in detecting malignant breast cancer. The analysis of the model's performance across different subsets of the data provided valuable insights into the underlying biological factors, underscoring the potential of machine learning in advancing cancer prediction and prevention efforts. The findings of this project have significant implications for the field of healthcare, emphasizing the importance of early cancer prediction and the potential of innovative technologies and methodologies in achieving this critical objective. As the field of machine learning continues to evolve, the integration of novel features and algorithms holds promise for further enhancing cancer screening and prevention, ultimately contributing to improved patient outcomes and public health.

CHAPTER – 8

REFERENCES

1) Yan C, Yao J, Li R, Xu Z, Huang J. Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest Xrays. Proceedings of the 2018 ACMInternational Conference on Bioinformatics, Computational Biology, and Health Informatics. Association for Computing Machinery, Washington, DC, USA, 2018, pp. 103–10.

2) Perna D., Tagarelli A. Deep Auscultation: Predicting Respiratory Anomalies and Diseases via Recurrent Neural Networks. In: 2019 IEEE 32nd International Symposium on Computer Based Medical Systems (CBMS), Cordoba, Spain: IEEE. 2019, pp. 50–55.

3) Zhao H., Zarar S., Tashev I., Lee C.-H. Convolutional-Recurrent Neural Networks for Speech Enhancement. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018

4) Adavanne S., Politis A., Nikunen J., Virtanen T. Sound Event Localization and Prediction of Overlapping Sources Using Convolutional Recurrent Neural Networks. IEEE Journal of Selected Topics in Signal Processing, 2018, Vol. 13, No. 1, pp 34–48.

5) Rocha B.M., Filos D., Mendes L. et al. A Respiratory Sound Database for the Development of Automated Classification. In: Precision Medicine Powered by pHealth and Connected Health, Singapore: Springer Singapore, Vol. 66, 2018. pp. 33–37. (IFMBE Proceedings).

6) Cohen,J.P.;Morrison,P,;Dao, L;Roth,K,;Duong,T. Q,;Gha s semi,M. Covid-19 Image datacollection Prospective

7) Prediction Are the Future, arXiv 2020, arXiv;2006.11988.

8) Abbas A, Abdelsamea MM, Gaber MM(2020) Classification of COVID-19 in chest X-ray images using DeTraC

9) Deep convolutional neural network. Appl Intel.

10) Armato, S.G.; McLennan, G.; Bidaut, L.; McNitt- GrayM.F.; Meyer, C.R.; Reeves, A.P.;Zhao, B.; Aberle,

10) D.R.;Henschke, C.I.; Hoffman, E.A.; et al. The Lung Image Database Consortium(LIDC) and Image Database

11) Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. Med. Phys. 2011,38.

12) ID, Mpesiana A (2020) Covid-19: automatic prediction from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci med 43:635–640.

13) Hari Krishna Timmana and Rajabhushanam C "Lung Malignant Prediction using Deep Learning Model", 2020.

14) Abhishek Verma, Cabinet Kumar Shah, Veerpal Kaur, Senate Shah, Prashant Kumar, "Cancer Prediction and Analysis Using Machine Learning",2022.

15) Ganta Sruthi, Chokkakula Likitha Ram, Malegam Koushik Sai, Bhanu Pratap Singh, Nikhil Majhotra, Neha Sharma, "Cancer Prediction using Machine Learning," in 2022.