Al- Aqsa University-Gaza
Faculty of computing and IT
Dep. Information and computer science

# Lung cancer detection model

## Using Convolution Neural Networks

## LCD – CNN Model

Student Names:

| | | |
|---|---|---|
| Mohammad Harara | : | 1301183127 |
| Omar Ahmad | : | 1301205694 |
| Mahmoud Abu Aisha | : | 1301194912 |
| Isa Abu Salmiya | : | 1301200584 |

**Supervisor:** Dr. Rasha Atallah

A project submitted in partial fulfillment of the requirements for the BSc degree of computer science.

2021 – 2022

# Abstract

Lung cancer is considered one of the primary life-threatening cancers worldwide with the highest mortality rate. Early detection of lung cancer plays an important role in early diagnosis and subsequent treatment.

Through this research, a convolutional neural network (CNN) was developed to detect the absence or presence of lung cancer in the human body using the agile methodology. Grayscale .PNG radiographs were used to diagnose lung cancer.

The proposed model built using CNN, then trained and validated using a set of x-rays, whose title is the Lung Cancer Survey.

Evaluation of the model showed that the CNN model is able to detect the absence or presence of lung cancer with 98% accuracy.

# الملخص

يعتبر سرطان الرئة أحد السرطانات الأولية التي تهدد الحياة في جميع أنحاء العالم مع أعلى معدل وفيات. يلعب الاكتشاف المبكر لسرطان الرئة دورًا مهمًا في التشخيص المبكر والعلاج اللاحق.

من خلال هذا البحث ، تم تطوير شبكة عصبية تلافيفية (CNN) للكشف عن غياب أو وجود سرطان الرئة في جسم الإنسان باستخدام المنهجية الرشيقة. تم استخدام الصور الشعاعية بتدرج الرمادي PNG.لتشخيص سرطان الرئة.

تم بناء النموذج المقترح باستخدام CNN ، ثم تم تدريبه والتحقق من صحته باستخدام مجموعة من الأشعة السينية ، وعنوانها هو مسح سرطان الرئة.

أظهر تقييم النموذج أن نموذج CNN قادر على اكتشاف غياب أو وجود سرطان الرئة بدقة ٩٨ ٪.

# Acknowledgement

First of all, We would like to express our deepest gratitude to Allah who provide guidance to complete our project.

We would like to express us special appreciation and thanks to our great supervisor Dr. Rasha Atallah for hir encouragement, guidance, valuable criticism and careful reading that resulted in producing the project in its current form.

Also, a great thank to all the computer sciences program academic stuff for their teaching and guidance.

To our fathers, who encouraged us to be the best we can be, to have high expectations and to fight hard for what we believe. The men to whom we will be grateful, for the rest of our lives.

To our mothers, our sources of inspiration for the whole of our lives. we would never forget their continuous prayer for the sake of our success.

A special thanks to our families. Words cannot express how grateful we are to our sisters and brothers.

Thank you for supporting us for everything, and especially we cannot thank you enough for encouraging us throughout this experience.

Finally, our appreciation goes to all colleagues and friends for their suggestions and encouragement during the work of this project.

# Contents

# Chapter 1

# Introduction

# 1. 1) Introduction:

Lung cancer is a condition that causes cells to divide in the lungs uncontrollably. This causes the growth of tumors that reduce a person's ability to breathe. Lung disease is standout amongst the most widely recognized malignancies, representing more than 225,000 people, 150,000 deaths, $12 billion cost yearly. It is as well one of the deadliest diseases; just 17% of individuals determined to have lung tumor survive 5 years after identification, survival rate is bringing down in developing nations. The most important cause of lung cancer is smoking (cigarette, pipe, and cigar). It has been found that 90% of all lung cancer cases have been due to tobacco smoking. Tobacco smoke contains around 4000 chemicals, out of which about 60 are found to be carcinogenic Other causes of lung cancer are, Radon(Naturally occurring gas),Asbestos(Natural mineral used in construction),Metals such as cadmium, chromium, arsenic exhausted from diesel engine. Chemicals in and certain diseases that affect the lung (e.g. tuberculosis).Family history of cancer can also put a person at greater risk. The closer the relative, the greater is the risk.

Comparing with the other cancer the prediction of the lung cancer at the early stages is a quite challenging task to the doctors Also it requires a lot of experience for a doctor but also with the less accuracy. The doctor need more experience to predict the cancer at the early stages although he is experienced sometimes he may fail to predict the cancer at early stages.

Based on this, a model was built that detects lung cancer, and classifies whether a person has lung cancer or not, through deep learning which is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called convolution neural networks.

# 1. 2) Problem statement :

In general there are medical errors in detecting the lung cancer, and there is also a low of accuracy[1] where the possibility of obtaining correct results for the presence or absence of lung cancer is not the required ratio, and also it was noted that Previous models need time in processing[3] images to get the result .

## 1. 3) Project questions :

1) What are the problems in the current cancer detection models ?

2) How can  improve the accuracy of the cancer detection model?

3) How can evaluate the proposed model?

## 1. 4) Project aim and objective

1) Identify the problems in the current cancer detection models.

2) To build a model based on CNN using to detect lung cancer.

3) To evaluate the proposed model using chest CT scan images Dataset.

## 1. 5) Project scope and limitation :

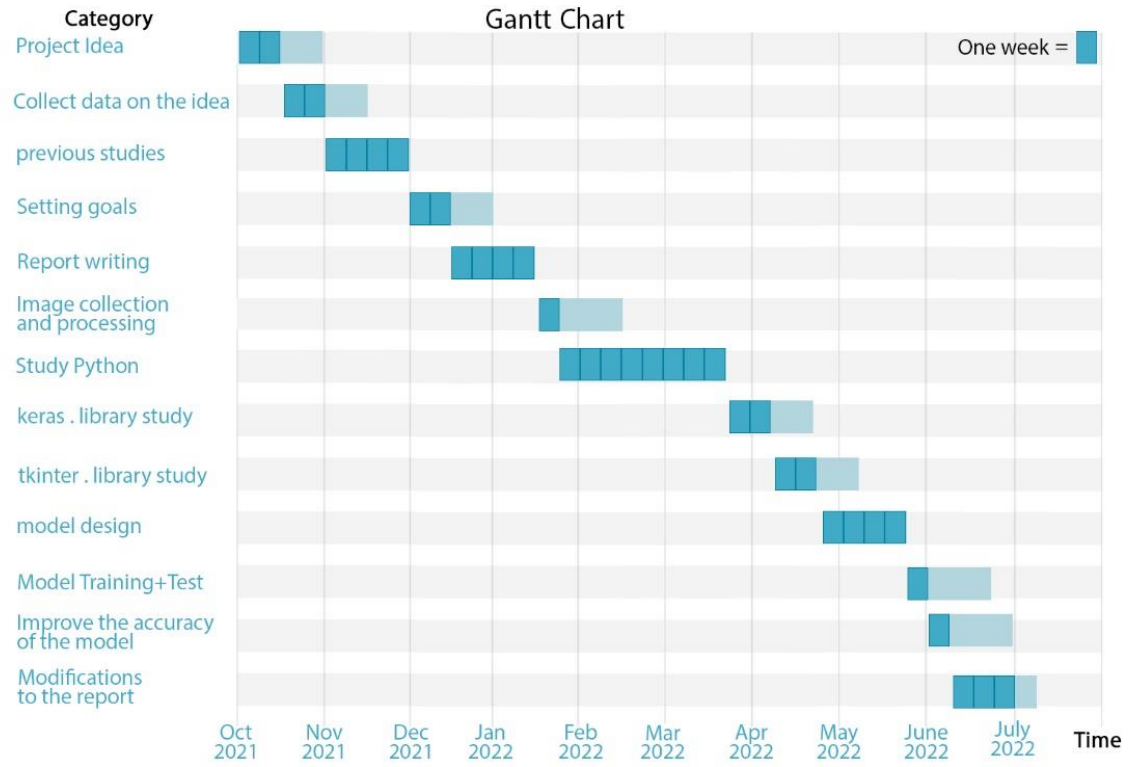
1) Scope :

- Only to detect lungs cancer.
- Extension for the image should be .png .

2) Limitation:

- The type of images used in the project are x-ray .
- Image dimensions must be [64 x 64].

# 1. 6) Gantt chart :




Figure(1.1) : Gantt chart for the project

As shown in Figure(1.1) the time period during which the work of this model was completed. This period was from the beginning of October 2021 until the end of July 2022, and this period was divided into weeks to complete the works shown in the plan at weekly intervals during this period.

# 1. 7) Conclusion :

In this chapter, an introduction to lung cancer has been clarified, the problems of detection lung cancer, also the objectives of the proposed model LCD – CNN, and then explain scope and limitations .

# Chapter 2

# Lecture Research

# (LR)

# 2. 1) Introduction:

This chapter discuss the previous studies of research papers and projects similar to LCD_CNN model, and the gaps in the previous models, also introduce an overview of the mechanism of work of this proposed model.

# 2. 2) Existing system :

Table (2.1) Existing systems

| Ref | Model | Objective | Dataset | No-Images | Method | Result | Criticize |
|---|---|---|---|---|---|---|---|
| [1] | Automatic Lung Cancer Prediction from Chest X-ray Images | chest x-ray-based lung cancer diagnosis using the deep learning approach. and solve the problem of a small dataset. | By Author | 247 frontal chest x-ray images; 154 of which have lung nodules )100 malignant cases, 54 benign cases) and 93 are images without lung nodules | Deep Learning Approach | ±74.43 6%.01 | - chest x-rays produce lower quality images compared to LDCT or CT<br>- All images have a pixel size of 2048 |
| [2] | Lung Cancer Detection using CT Scan Images | cancer nodule detection and Classifies the detected lung cancer as malignant or benign | By Author | Different 16 DICOM images from LIDC are used for training the classifier and result is validated using 5 images with total 15 nodules | Subtraction method between two serial mass chest radiographs is proposed. | 92% | - It classifies the cancer as just malignant or benign but does not classify into different stages like stage I, II, III, IV.<br>- LIDC database was very large in size (124 GB) which was very tedious to download.<br>- Use images in JPEG format only. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [3] | Lung Cancer Detection: A Deep Learning Approach | approach to detect lung cancer from CT scans using deep residual learning. | LIDC-IDRI dataset. | (LIDC-IDRI) images | (Deep Learning) employ deep residual networks to extract features from preprocessed images which are fed to classifiers, the predictions of which are ensembled for the final output. | 84% | DICOM image format |
| [4] | Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes | A liquid biopsy that captures and detects tumor-related biomarkers in body fluids has great potential for early-stage diagnosis. | by author | - | (Deep Learning) demonstrate an accurate diagnosis of early-stage lung cancer, using deep learning-based surface-enhanced Raman spectroscopy (SERS) of the exosomes. | 95% | Plasma exosome analysis is time consuming |
| [5] | Detection of lung cancer using artificial neural network and fuzzy assembly methods . | Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. | by author | 1000 sputum color images to test both methods, | artificial neural network (ANN) | 60% | - Use color image just.<br>- The accuracy is low |
| [6] | Early Detection of Lung Cancer Using Wavelet Feature Descriptor and Feed Forward Back Propagation Neural Networks Classifier | early detection of cancer method using wavelet and ANN for identifying cancerous lung nodules and without cancerous lung nodules. | by author. | 50 testing images | artificial neural network (ANN) | 92.61% | - images with diameter ranges from 260 to 400 mm with the layer thickness of 0.75–1.25 mm<br>- When maximum number of epochs is reached |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [7] | Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer | develop a radiomics prediction model to improve pulmonary nodule (PN) classification in low-dose CT. To compare the model with the American College of Radiology (ACR) Lung CT Screening Reporting and Data System (Lung-RADS) for early detection of lung cancer. | Database Consortium image collection (LIDC-IDRI) in The Cancer Imaging Archive (TCIA) contains 1018 cases | 72 PNs (31 benign and 41 malignant) | constructed a prediction model by using a support vector machine (SVM) classifier coupled with a least absolute shrinkage and selection operator (LASSO). A tenfold cross-validation (CV) was repeated ten times (10 × 10-fold CV) to evaluate the accuracy of the SVM-LASSO model. | 84.6% | Development on the model is still little |
| [8] | Serum Protein Markers for the Early Detection of Lung Cancer: A Focus on Autoantibodies | provide differentially expressed protein, antigen, and autoantibody biomarkers that combined with CT imaging for early detection of lung cancer. | by author | - | Lung cancer detection using biomarkers | 85% | - 2D image<br>- Low sensitivity |
| [9] | CT-Derived Features for Accurate Detection of Lung Cancer | for pulmonary nodule diagnosis using various features extracted from a single computed tomography (CT) scan , system fuse texture and shape features to get an accurate diagnosis for the extracted lung nodules. | Lung Image Database Consortium (LIDC) | 1018 CT scans | ANN | 95% | - The process of converting images to 3D takes a lot of time |

As Show In Table (2.1) the existing *Systems* that low dose computed tomography (LDCT) and computed tomography (CT) scans provide greater medical information than normal chest x-rays, access to these technologies in rural areas is very limited. There is a recent trend toward using computer-aided diagnosis (CADx) to assist in the screening and diagnosis of cancer from biomedical images.

The first model show 121-layer convolutional neural network, also known as DenseNet121 along with the transfer learning scheme is explored as a means of classifying lung cancer using chest x-ray images. The model was trained on a lung nodule dataset before training on the lung cancer dataset to alleviate the problem of using a small dataset. The proposed model yields 74.43±6.01% of mean accuracy, 74.96±9.85% of mean specificity, and 74.68±15.33% of mean sensitivity. The proposed model also provides a heat map for identifying the location of the lung nodule. These findings are promising for further development of chest x-ray-based lung cancer diagnosis using the deep learning approach. Moreover, they solve the problem of a small dataset. [1]

The second model have real patient CT scan images are obtained from Lung Image Database Consortium(LIDC) archive. It is the database of lung cancer screening CT images for development, training, and evaluation of computer assisted diagnostic methods for lung cancer detection and diagnosis. It was initiated by National Cancer Institute. It consists of 1018 cases of dataset contributed by seven academic center and eight medical imaging companies. Images are in DICOM format with size 512*512 pixel. DICOM format is difficult to process; therefore those images are converted to JPEG Gray scale image using software MicroDicom software. MicroDicom opens the DICOM CT scan images and can also convert to appropriate JPEG format. The proposed model is then developed in MATLAB R2016a. MATLAB is one of the tools for research development and analysis . Both detection and features extraction are implemented in MATLAB and classification is implemented using machine learning toolbox. Classification learner toolbox aids in developing the trained prediction model from the features extracted easily and very fast. 5 folds cross validation was used to prevent from overfitting during the training process. Different 16 DICOM images from LIDC are used for training the classifier and result is validated using 5 images with total 15 nodules. seen that there is progressive increase in accuracy 92%. Sensitivity remained same. Specificity increased to 50% From the detected cancer nodes, features like Area, Perimeter, Centroid, Diameter, Eccentricity and Mean Intensity of the Pixels were extracted. Extracted features were used to Train Support vector machine and trained model was developed. Training time for classification learner app was 5.93 seconds. Classification learner app evaluates the prediction time for the developed trained model to be 310 observations per second. [2]

The third model employ deep residual networks to extract features from preprocessed images which are fed to classifiers, the predictions of which are ensembled for the final output. We explain in this paper the proposed methodology, evaluation, and results using the LIDC-IDRI dataset. The Lung Image Database Consortium image collection (LIDC-IDRI) contains diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. It consists of more than thousand scans from high-risk patients in the DICOM image format. Each scan contains a series of images with multiple axial slices of the chest cavity. Each scan has a variable number of 2D slices, which can vary based on the machine taking the scan and patient. The DICOM files have a header that contains the details about the patient id, as well as other scan parameters such as the slice thickness. The images are of size (z, 512, 512), where z is the number of slices in the CT scan and varies depending on the resolution of the scanner . [3]

The Fourth model deep learning-based surface-enhanced Raman spectroscopy (SERS) of the exosomes. Our approach was to explore the features of cell exosomes through deep learning and figure out the similarity in human plasma exosomes, without learning insufficient human data. The deep learning model was trained with SERS signals of exosomes derived from normal and lung cancer cell lines and could classify them with an accuracy of 95%. In 43 patients, including stage I and II cancer patients, the deep learning model predicted that plasma exosomes of 90.7% patients had higher similarity to lung cancer cell exosomes than the average of the healthy controls. Such similarity was proportional to the progression of cancer. Notably, the model predicted lung cancer with an area under the curve (AUC) of 0.912 for the whole cohort and stage I patients with an AUC of 0.910. demonstrate that the deep learning analysis of the nano plasmonic sensing technique can be used to identify early-stage lung cancer patients, with high accuracy. Without specific biomarkers, method was able to detect the signal feature of lung cancer cell-derived exosomes among plasma exosomes. The deep learning model supervised by cellular exosomes successfully identified the lung cancer patients and even detected stage I patients. Our method basically relies on a noninvasive, safe, and sensitive analytic method for detecting lung cancer cell-derived exosomes in blood. These suggest that our method can be used as a routine prescreening tool for lung cancer. [4]

The Fifth model the early detection of the lung cancer is a challenging problem, due to the structure of the cancer cells. This paper presents two segmentation methods, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improves the chances of survival for the patient. The two methods are designed to classify the image of N pixels among M classes. In this study, we used 1000 sputum color images to test both methods, and HNN has shown a better classification result than FCM, the HNN succeeded in extracting the nuclei and cytoplasm regions. [5]

The Sixth model a CADs system to classify lung nodules done using Wavelet feature descriptor which computed from the gray level co-occurrence matrix of a Daubechies wavelet transform and feed forward back propagation neural networks classifier. The region of interest is obtained from the segmented single slices containing 2 lungs. The neural network that constructed using four training functions (Traingd, Traingda, Traingdm, and Traingdx), the traingdx training function gives the maximum classification accuracy. The proposed NN feed forward back propagation classifier produced Accuracy of 92.61%, specificity of 100% and sensitivity of 91.2% and a mean square error of 0.978. The sensitivity is calculated by evaluating the percentage of segmented lung nodules containing cancerous nodule that is correctly classified as cancerous. From this new approach, radiologists can use our CADs for lung cancer detection accurately and easily in the early stage of cancer. [6]

The Seventh model to develop a radiomics prediction model to improve pulmonary nodule (PN) classification in low-dose CT. To compare the model with the American College of Radiology (ACR) Lung CT Screening Reporting and Data System (Lung-RADS) for early detection of lung cancer.

We examined a set of 72 PNs (31 benign and 41 malignant) from the Lung Image Database Consortium image collection (LIDC-IDRI). One hundred three CT radiomic features were extracted from each PN. Before the model building process, distinctive features were identified using a hierarchical clustering method. We then constructed a prediction model by using a support vector machine (SVM) classifier coupled with a least absolute shrinkage and selection operator (LASSO). A tenfold cross-validation (CV) was repeated ten times (10 × 10-fold CV) to evaluate the accuracy of the SVM-LASSO model. Finally, the best model from the 10 × 10-fold CV was further evaluated using 20 × 5- and 50 × 2-fold CVs. [7]

The Eighth model this autoantibody response to tumor-associated antigens starts during early stage lung cancer and may endure over years. Identification of tumor-associated antigens or the corresponding autoantibodies in body fluids as potential noninvasive biomarkers could thus be an effective approach for early detection and monitoring of lung cancer. We provide an overview of differentially expressed protein, antigen, and autoantibody biomarkers that combined with CT imaging might be of clinical use for early detection of lung cancer. Lung cancer tissue generates lung cancer-associated proteins to which the immune system might produce high-affinity autoantibodies. This autoantibody response to tumor-associated antigens starts during early stage lung cancer and may endure over years. Identification of tumor-associated antigens or the corresponding autoantibodies in body fluids as potential noninvasive biomarkers could thus be an effective approach for early detection and monitoring of lung cancer. We provide an overview of differentially expressed protein, antigen, and autoantibody biomarkers that combined with CT imaging might be of clinical use for early detection of lung cancer. [8]

The ninth model pulmonary nodule diagnosis using various features extracted from a single computed tomography (CT) scan. The proposed system fuse texture and shape features to get an accurate diagnosis for the extracted lung nodules. 3D Local Binary Pattern (LBP) and higher-order Markov Gibbs random field (MGRF) models are utilized to model the texture appearance due to their capability to give a precise description for the spatial non-uniformity in the texture of the nodules. Spherical Harmonic expansion and some basic geometric features are utilized to model the shape features due to their capability to give a full description of the shape complexity of the nodules. Finally, all the modeled features are fused and fed to a stacked autoencoder to differentiate between the malignant and benign nodules. Our framework is evaluated using 727 nodules which are selected from the Lung Image Database Consortium (LIDC) dataset, and achieved classification accuracy, sensitivity, and specificity of 92.66%, 95.70%, and 90.40% respectively. [9]

# 2. 3) The gaps in the existing models :

There is a low of Accuracy [1] where the possibility of obtaining the correct results for the presence or absence of lung cancer is not the required percentage, as it was noted that previous models need time in processing[3] images to obtain the result and also many deep learning techniques are applied to detect lung cancer . This is due to the Lack of a Large Dataset[3] for medical images especially lung cancer and also some models only support one type of image [3].

# 2. 4) Architecture for cancer detection models:

As shown in figure 2.1 the Architecture for Cancer Detection , have four stages In this model, the x-ray images collected through the x-ray machines, which were previously discussed, are entered, and then these images enter the reprocessing stage to be processed, and then the re-processed image is entered To the training stage and the model was trained on these images, then we moved to the classification stage, where the model classifies the images into a cancer patient and a non-patient, and at the end of these stages a new x-ray is tested to discover whether he is a cancer patient or not a cancer patient .

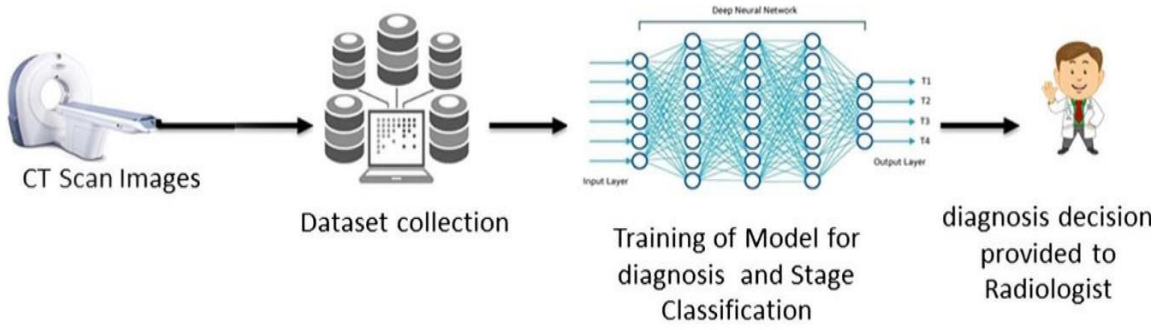



Figure (2.1): Architecture for Cancer Detection

# 2. 5) Conclusion :

This chapter clarify the previous studies and the gaps in the previous systems, such as: low of accuracy, long processing time, small size of the data set, and an overview of the work structure of the proposed model was given.

# Chapter 3

# Methodology

# 3. 1) Introduction:

In this chapter clarify the methodology that used to build LCD-CNN model and why it used, also it is explain the stages for building the proposed model.

# 3. 2) Agile methodology :

Agile is a project management methodology that breaks down large projects into smaller manageable parts known as iterations. At the end of each iteration (which usually takes places over a fixed period of time), something of value is produced. The product produced during each iteration must be able to put it out in the world to receive feedback from stakeholders or users.

This methodology consists of several recurring stages, namely analysis, design, programming, testing, and review as shown in Figure (3.1) . These stages will be discussed during the next chapters.


Figure (3.1): Agile Phases

# 3. 3) Choosing agile methodology:

This methodology was used because the model is a medical system that must be accurate and not tolerate errors Accordingly, the agile methodology was used because it consists of several stages, and these stages are repeated so that the required outputs come out with the highest accuracy, as the accuracy came in several stages. 98.6%, so this methodology is suitable for Moodle.

# 3. 4) Conclusion :

In this chapter, the methodology that was used and why it was used are explained.

# Chapter 4

# System Analysis

# And Design

# 4. 1) Introduction:

In this chapter, the proposed model is explained in more detail than last time, the dataset used to build this model, and a simple UML to show the image path in the model are explained.

# 4. 2) Architecture for LCD-CNN model:

In LCD-CNN the x-ray images collected through the x-ray machines, which were discussed earlier, are entered, and these images are then entered into a set of stages described in Figure (4.1):

The first stage is preprocessing: where the entered image is processed to make its size (64 x 64), and then the image is converted to black and white if the image is color.

The second stage is training the model: at this stage, the features are selected through training, such as: tumor area, tumor circumference, tumor diameter, average tumor density in pixels., where the model is trained on these extracted features to perform classification when tested.

The third stage of classification: where the trained model classifies the images into (normal and cancer), according to the extracted features.

In the end, the final result of the picture is given. Is the person sick with cancer or normal (not sick)?

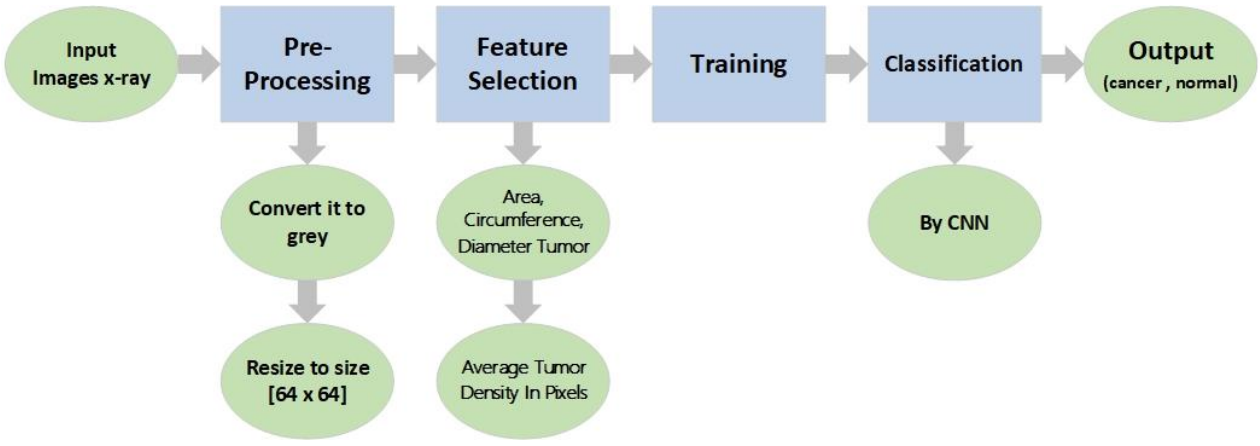



Figure (4.1): Architecture for LCD-CNN model

Clarify what is meant by the CNN algorithm mentioned in Figure(4.1). The CNN algorithm in is an abbreviation of the Convolutional Neural Network algorithm:

1) Convolutional layer: through which the resulting matrix of the image is multiplied by the specified filter using one of the arithmetic operations (relu, sigmoid).

Rule : $$R(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$ , Sigmoid : $$f(x) = \frac{1}{1 + e^{-x}}$$

2) Max Pooling layer: here the size of the resulting matrix from the previous stage is reduced to a size that is specified in the function.

Note: the previous two stages are sometimes repeated to increase the accuracy of the model .

3) Flattening layer: It is a stage through which the matrix is converted from two-dimensional to one-dimensional.

4) Full Connection layer: here the previous stages are executed and assembled, and the accuracy and error rate are calculated.

# 4. 3) The dataset :

A data set called Chest CT-Scan Images Dataset [10] was used, which was collected through imaging with x-ray devices from several hospitals around the world and is classified as cancer patient or non-cancer patient. In the operations of the model that was created for the detection of lung cancer. The total of these images used in the operations was 3,439 images with the extension .png, and this image was divided as follows:

The total number of images classified as cancer patients: 2618 images, The total number of images classified as normal people: 821 images, The number of images used in the training process: 2264 images, divided into 1672 images classified as patients and 592 images were classified as normal Number of images used in Test process: 1112 images, divided into 896 images classified as diseased and 216 normal images. Number of images used for verification: 63 images, divided into 50 images classified as sick and 13 images as normal.

Table (5.1): showing the distribution of images for the dataset

| Classification | Training | Testing | Validation | Total |
|---|---|---|---|---|
| Cancer | 1,672 | 896 | 50 | 2,618 |
| Normal | 592 | 216 | 13 | 821 |
| Total | 2,264 | 1,112 | 63 | 3,439 |

# 4. 4) UML use case :

As Figure (4.2) At first, the radiographer takes an x-ray of the patient's lung, then the image is entered into the program to be examined, by uploading the image to the program, where the program processes it before entering it into the CNN algorithm to predict the outcome of the injury, where the algorithm code gives either zero or one, Where 0 is a patient (cancer) and 1 is a (normal) , and in the end the result is given to the patient.

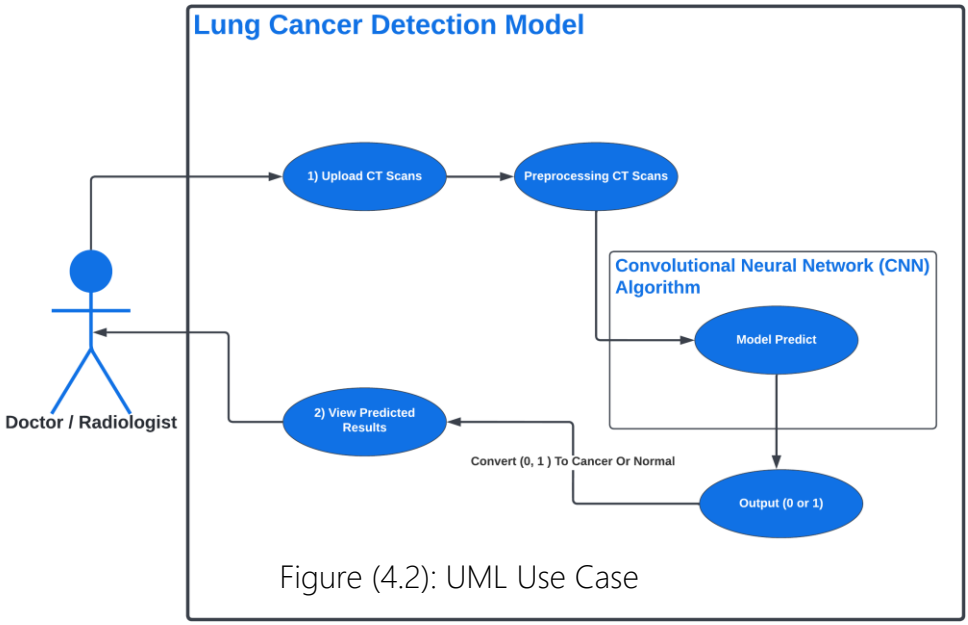



Figure (4.2): UML Use Case

# 4. 5) Conclusion :

In this chapter, the proposed model has been explained in detail, the dataset used in building this model, and the UML to illustrate the path of the image in the model have been clarified.

# Chapter 5

# Implementation

# 5. 1) Introduction:

During this implementation chapter, a lung cancer detection model is built using the Python language and a CNN algorithm call. In this chapter, the CNN algorithm is explained and the code for this model is explained.

# 5. 2) Implementation :

The code was generated in Python using the Convolutional Neural Networks (CNN)

algorithm, and this is the code:

```
# Importing the keras libraries and packages
import numpy as np
from keras.models import Sequential
from keras.layers import Conv2D, MaxPooling2D, Flatten, Dense
from keras.preprocessing.image import ImageDataGenerator
from keras.utils import load_img, img_to_array
```

All libraries required to create the CNN algorithm and to load images for the algorithm are called here.

```
# Initializing the CNN ---------- [ Part One ]
model = Sequential()
```

Here a Sequential object is created.

# Step 01 -> Convolution

*model.add(Conv2D(32, (3, 3), input_shape=(64, 64, 3), activation="relu"))*

# Step 02 -> Max Pooling

*model.add(MaxPooling2D(pool_size=(2, 2)))*

- Here, 32 layers of the image are created, and a 3 x 3 filter is applied to it, and the image is scaled to a size (64 x 64) and a depth of 3, i.e. a color image (RGB), and then a modified (Relu) is applied to it. The output is either zero or the value of the input x)).

Here, the matrix resulting from the previous operation is reduced to a size (2 x 2).

- Now the previous two stages are applied twice, to get more accurate results.

# Step 03 -> Flatten

*model.add(Flatten())*

Here the resulting matrix from the previous stage is made into a one-dimensional matrix.

# Step 04 -> Full Connection

➔ Input Nodes (128 => Hidden Node)

*model.add(Dense(units=128, activation="relu"))*

➔ Output Nodes (1 => Output Node [Cancer Or Normal])

*model.add(Dense(units=1, activation="sigmoid"))*

- Here the resulting matrix from the previous stage is entered and multiplied by 128 nodes (hidden nodes) using (*relu*) law.

- And then out of 128 nodes, one node is either (0) or (1) using the arithmetic operation (*sigmoid*).

```
# Compiling the CNN
model.compile(optimizer="adam", loss="binary_crossentropy", metrics=["accuracy"])
```

Here the previous stages are collected, the accuracy of the model is found, and the percentage of errors in the model (Loss) is calculated.

# 5. 3) Project software and hardware requirements:

When creating a program project through Python and using an algorithm such as the "CNN" algorithm, a set of public libraries must be called to build the project through. These libraries are:

1) Keras :

It is a library based on the Python environment based on its work on the TensorFlow library, through which the functions of using the CNN algorithm are called, which is the function [Dense, Flatten, MaxPooling2D, Conv2D, Sequential] and also the function [ImgeDataGenerator] is called Through which the properties of the images are modified to suit the model.

2) TKinter :

It is an internal library in the Python environment that is used to build simple application interfaces, to facilitate the process of dealing with them for the user, and through which images are also uploaded and given to Keras to predict the classification of the image, for example.

Note : Python 3.10 and the integrated development environment (IDE) PyCharm were used to build this mod.

# 5. 4) Conclusion :

During this chapter, the architecture of the CNN algorithm is explained, and the hardware and software components required to build this model are explained.

# Chapter 6

# Testing And

# Evaluation

# 6. 1) Introduction:

In this chapter, the test of all kinds is explained, the white box test and the black box test, and the evaluation of the model is explained in two metrics, the first using the code and the second manually.

# 6. 2) Testing:

In this section, the process of testing the LCD-CNN model is explained, as it is divided into :

## 6. 2. 1) White testing :

Unit testing focuses on validating the constituent units of the built model, and as mentioned in the Execution chapter (5.2), the model is divided into seven modules: the library calling module, the model initialization module, the convolutional module, the Max Pooling module, the Flatten module, and then the communication module Completely nodes and then assembly, it is this unit that constitutes the white box test.

Unit testing is a white box oriented test. First of all, the interface of the module is tested to ensure that information flows correctly in and out of the program even under test. Then the local data structure is tested to ensure that the cached data maintains its integrity during all execution steps. Boundary conditions are tested to ensure that the unit operates correctly at the limits set to restrict or limit processing. All independent paths are exercised through the control structure to ensure that all statements in the module have been executed at least once. Finally, if any errors are found, they are immediately corrected and the unit is tested again.

That is, there are no syntax errors in LCD-CNN

## 6. 2. 2) Black testing :

The black box test is where a group of pictures of patients with cancer and normal is entered, and the output of the model is according to the picture that was entered into the form. Semantic errors, but if the entered image is a cancer patient and the output is a normal image or vice versa, there will be semantic errors, in the LCD-CNN model there are no semantic errors.

# 6. 3) Evaluation :

In this section, the process of measuring the accuracy of the LCD-CNN model, which was obtained after a training process of about 10 minutes, is explained. The accuracy is measured in two ways:

## 6. 3. 1) Coding :

In the process of calculating the accuracy, the technique called metrics = ["accuracy"], which depends on the confusion matrix, calculated the accuracy of the LCD-CNN model, and the resulting accuracy was 98%

```
model.compile(optimizer="adam", loss="binary_crossentropy", metrics=["accuracy"])
```

Table (6.3) : Represents the accuracy of the models .

| Ref | Algorithm | Accuracy |
|---|---|---|
| [1] | Deep Learning Approach | 74.43 % |
| [2] | Subtraction method between two serial mass chest radiographs is proposed. | 60.1 % |
| [3] | Deep Learning | 84 % |
| [4] | (Deep Learning) using deep learning-based surface-enhanced Raman spectroscopy (SERS) of the exosomes. | 95 % |
| [5] | Artificial Neural Network (ANN) | 60 % |
| [6] | Artificial Neural Network | 92.61 % |
| [7] | support vector machine (SVM) | 84.6 % |
| [8] | Lung cancer detection using biomarkers | 85 % |
| [9] | ANN | 95 % |
| | LCD – CNN Model | **98 %** |

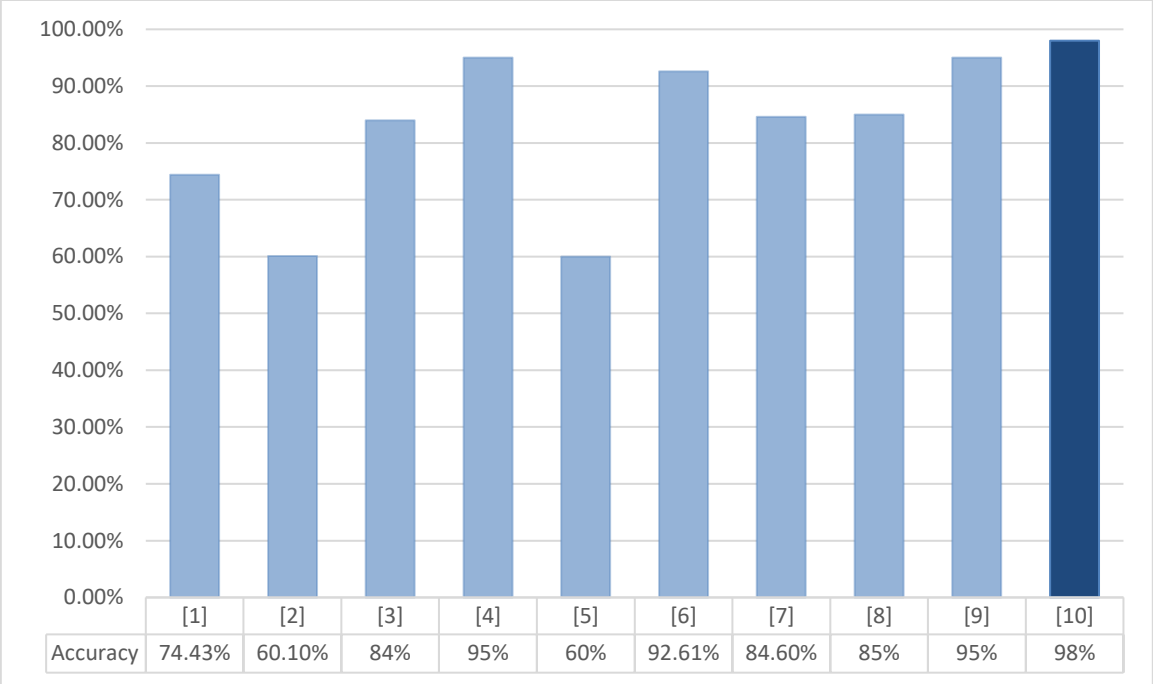



Figure (6.1): Represents The Accuracy Of The Models

## 6. 3. 2) Manually:

In the table the Pattern of the Confusion Matrix and its parts on which the calculation of accuracy depends.

Table(6.1) : Pattern of the Confusion Matrix

| | | Predicted | |
|---|---|---|---|
| | | Negative | Positive |
| Actual | Negative | True Negative | False Positive |
| | Positive | False Negative | True Positive |

Table(6.2) : The Confusion Matrix implementation of the LCD-CNN model

| | | Predicted | |
|---|---|---|---|
| | | Cancer | Normal |
| Actual | Cancer | 893 | 3 |
| | Normal | 3 | 213 |

**Precision :** talks about how precise/accurate model is out of those predicted positive, how many of them are actual positive.

**Precision** = True Positive **/ (** True Positive **+** False Positive **)**

now us apply the same logic for Precision

**Precision =** 213 **/ (** 213 **+** 3 **)** = **0,98611**

**Recall actually :** calculates how many of the Actual Positives model capture through labeling it as Positive (True Positive).

**Recall =** True Positive **/ (** True Positive **+** False Negative **)**

now us apply the same logic for Recall

**Recall =** 213 **/ (** 213 **+** 3 **)** = **0,98611**

**F1 Score :** might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives).

**F1 =** 2 *** [** Precision ***** Recall **/** Precision **+** Recall **]**

now us apply the same logic for F1 Score

**F1 =** 2 *** [** 0,98611 ***** 0,98611 **/** 0,98611**+** 0,98611 **] = 0,9860**

# 6. 4) Conclusion :

In this chapter, all types of testing are explained, the white box test and the black box test, and the evaluation of the model in two metrics, the first using the code and the second manually, was explained.

# Chapter 7

# Conclusion And Future Work

# 7. 1) Introduction:

In this chapter, the future development of that project or model is explained, and a quick summary of what was explained in all previous chapters is provided.

# 7. 3) Future work :

1) Discover whether cancer is benign or malignant .

2) Support images in any extension .

3) 3D photo support .

4) Color photo support .

5) Locating the cancer in the lung .

6) Embedded system with a website .

# 7. 4) Conclusion :

In all the previous chapters an introduction was given about this project about what it is talking about, then the reasons that led to the need for such a project were clarified, and it was clarified what are the advantages of this model over the previous models, the most important of which was accuracy, and then it was explained The structure and structure of the project, what is the algorithm that was used, the method of dealing with this model was developed, and the method of testing the model was clarified, how it was and how this model was evaluated.

# RESOURCES AND REFERENCES :


1. Ausawalaithong, W., et al. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. in 2018 11th Biomedical Engineering International Conference (BMEiCON). 2018. IEEE.

2. Makaju, S., et al., Lung cancer detection using CT scan images. Procedia Computer Science, 2018. 125: p. 107-114.

3. Bhatia, S., Y. Sinha, and L. Goel, Lung cancer detection: a deep learning approach, in Soft Computing for Problem Solving. 2019, Springer. p. 699-705.

4. Shin, H., et al., Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes. ACS nano, 2020. 14(5): p. 5435-5444.

5. Taher, F. and R. Sammouda. Lung cancer detection by using artificial neural network and fuzzy clustering methods. in 2011 IEEE GCC conference and exhibition (GCC). 2011. IEEE.

6. Arulmurugan, R. and H. Anandakumar, Early detection of lung cancer using wavelet feature descriptor and feed forward back propagation neural networks classifier, in Computational vision and bio inspired computing. 2018, Springer. p. 103-110.

7. Choi, W., et al., Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. Medical physics, 2018. 45(4): p. 1537-1549.

8. Broodman, I., et al., Serum protein markers for the early detection of lung cancer: a focus on autoantibodies. Journal of Proteome Research, 2017. 16(1): p. 3-13.

9. Shaffie, A., et al. On the integration of CT-derived features for accurate detection of lung cancer. in 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). 2018. IEEE.

10. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images