Ministry of Higher Education and
Scientific Research
University of Diyala
College of Science
Department of Computer Science

Effect of Rescale Parameter on Some Classification Methods in Diagnosis of Lung Cancer

A Thesis

Submitted to the Department of Computer Science/ College of Science/ University of Diyala in a Partial fulfillment of the Requirements for the Degree of Master in Computer Science

By

Hafssa Ahmed Shukur

Supervised By

Assist.Prof. Dr. Bashar. Talib

2021 A.D. 1442 A.H.

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

﴿قَالَ الَّذِي عِنْدَهُ عِلْمٌ مِنَ الْكِتَابِ أَنَا آتِيكَ بِهِ
قَبْلَ أَنْ يَرْتَدَّ إِلَيْكَ طَرْفُكَ فَلَمَّا رَآهُ مُسْتَقِرًّا عِنْدَهُ
قَالَ هَذَا مِنْ فَضْلِ رَبِّي لِيَبْلُوَنِي أَأَشْكُرُ أَمْ أَكْفُرُ
وَمَنْ شَكَرَ فَإِنَّمَا يَشْكُرُ لِنَفْسِهِ وَمَنْ كَفَرَ فَإِنَّ رَبِّي
غَنِيٌّ كَرِيم﴾ 40

صدق الله العظيم

سورة النمل اية (40)

ACKNOWLEDGMENTS

Initially, I extend my sincere thanks to Allah who blesses me and helps me in the achievement of this work. I would like to express my appreciation to my supervisor, Dr. Bashar for his faithful guidance, valuable instructions, and constructive comments which have made the completion of this work possible.

I want to thank my parents, my husband, and my son, and my brothers for their love and guidance which helped me in achieving my goals. Also, I would like to express my gratitude and thanks to all the teaching staff who have taught me. Special thanks are extended to all my friends for their help and all people who have contributed to achieved my thesis. Special thanks to the members of the evaluation committee for discussing my thesis.

At long last, there are no words enough to thank my husband for being strong and having faith in me constantly and his support to complete this master project.

Dedication

To the flower of life and its light and the most precious person in my life, my tender mother.

To whom I proudly carry your name, teach me how to make success and instill confidence in myself my dear father, may God extend your life.

To those who supported me in adversity and the source of my happiness, my companion to my path and my love, my dear husband.

To whom their love and blood flow in my veins, and I lived with them the most beautiful moments, my brothers and sisters.

To the fruit of my love and my soul, my dear son.

To the candles that illuminate the path of knowledge in my path, my distinguished teachers.

To everyone who loved me and supported me in my scientific and practical life.

Abstract

Lung cancer is one of the deathliest diseases in the world and it is a topic of concern because the actual treatment of this disease has not been found yet. Patients with this disease can only be saved if and only if the disease is early diagnosed.

In This thesis presents intelligent lung tumor diagnosis system is developed using various processing technique for the classification of lung cancer after its detection with the help of machine learning algorithms. Where several steps are used in the form of stages which include, the data acquisition stage obtained from the (data world) archive, which contains 1000 samples and 25 features.

according to the high complexity of the patterns that appear in the dataset, which are used by the classifiers to extract the required knowledge, data preprocessing can have a significant influence on the performance of the classifiers. Accordingly, two different data preprocessing techniques are evaluated and used in this study (min max-normalize and z-score standardize), to illustrate their influence on the performances of different classifiers. and the classification stage using four Radial Support Vector Machine(SVM) and linear support vector machine(LSVM) classifier and Back Propagation Neural Network (BPNN) classifier and Naïve Bayes (NB) classifier and comparison between the accuracy and time taken of each model

The proposed system has been tested by using lung cancer dataset. The comparison results show that the proposed intelligent system has a good diagnosis performance, where the accuracy rate of first proposed model using (SVM) with normalization had an average accuracy of 98.21% while with standardization had an accuracy of 100.00%. The second proposed model Linear Support Vector Machine (LSVM) with normalization had an average accuracy of 94.63% while standardization had an accuracy of 99.55%. The third proposed model using (BPNN) with normalization had an 97.57% and standardization had an accuracy of 100.0%. The fourth proposed model using (GNB) with normalization and standardization had an accuracy of 89.1%.

List of Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

Abbreviations	Meaning
(CAD)	Computer-Aided Diagnosis
(SVM)	Support Vector Machine
(LSVM)	Linear Support Vector Machine
(ANN)	Artificial Neural Network
(NB)	Naïve Bayes
(GNB)	Gaussian Naive Bayes
(CXR)	chest X-ray
(LDCT)	low-dose computed tomography
(CAD)	Computer-aided detection
(MRI)	magnetic resonance imaging
(CT)	computed tomography
(ROI)	region of interest)
(ML)	Machine Learning
(DL)	Deep Learning
(RNN),	recurrent neural networks
(CNN)	convolutional neural networks
(KNG-CNN).	Kernel-based Non- Gaussian Convolutional Neural Networks
(BPNN),	Back Propagation neural
(KNN)	k-nearest neighbors
(ODNN)	(Optimal Deep Neural Network)
(LDA)	Linear Discriminate Analysis
(m-MKL)	Multiclass Multiple kernel learning classifier
(PCA)	Principal component analysis
(LIDC-IDRI)	Lung Image Database Consortium image selection
(CLAHE)	Contrast limited adaptive histogram equalization
(ACR).	American College of Radiology
(PN)	Pulmonary nodules
(RADS)	Reporting and Data System
(CV)	Computer vision
(MLP-NN)	Multi-Layer Perceptron Neural Network
(EDA)	Exploratory Data Analysis
(SRM)	Structural Risk Minimization
(1V1)	One-Versus-One
(1VR)	One-Versus-Rest
(RBF),	Radial Basis Function
(FNN),	Feed-forward Neural Networks

(RBMs).	Restricted Boltzmann Machines
(Tanh)	Hyperbolic tangent
(ReLU):	Rectified linear unit
(BP)	Back-Propagation
(TP)	True Positive
(TN)	True Negative
(FP)	False Positive
(FN)	False Negative

Chapter One

Introduction

Chapter One

INTRODUCTION

1.1 Introduction

Lung Cancer is a prevalent disease where deaths due to it are increasing nowadays. Through all lung cancer is predominant for men and women compared to all other cancers. It is caused due to smoking and eating tobacco causing cancer in lung tissues. The cancerous nodules are called malignant tumors it occurs because of uncontrollable cell growth in lung tissues. Despite advances in detection and diagnosis and therapies, many people still develop fatal lung cancer [1].

Lung Cancer is classified into two major groups based on cell size: i. cell size is small and ii. cell size is large. Cancer is divided into four stages and this staging has been done according to the size of the tumor and node location. Lung cancer is one of the most dangerous cancers, with a low survival rate following diagnosis. In general, lung cancer affects 75 % of females and 84 % of males who smoke. About 10-15% of cases occur in people who have never smoked [2]. Recent advances in computed tomography (CT) imaging have resulted in early diagnosis of lung cancer. Two common screening tests involve the use of chest X-ray (CXR) and the use of low-dose computed tomography (LDCT) scan of the chest. Below in Figure (1.1) are some pictures that show the spread of the tumor in the lung.

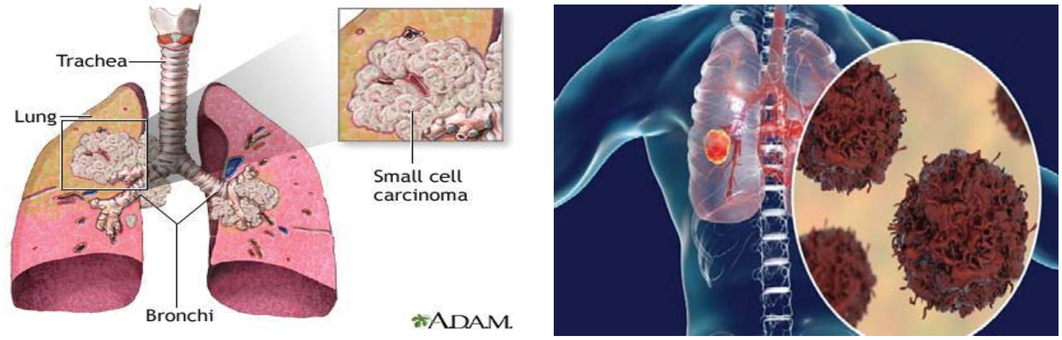

Figure 1.1 the spread of the tumor in the lung

In Computer-Aided Detection (CAD) systems have the ability to improve/assist radiologists and their workflow decreases error outcomes. CAD is a technique for locating abnormal regions (masses, nodules, and polyps) in the body and providing a location to radiologists. CAD has been used in a variety of medical imaging techniques, including magnetic resonance imaging (MRI), computed tomography (CT) and ultrasound. In particular, region of interest (ROI) based segmentation, feature extraction (kind of feature) and classification are three important stages in CAD systems used for cancer diagnosis and detection[3]. A region generating mechanism is deployed to produce a large number of candidate regions having a high likelihood of containing the pulmonary nodules [4].

The purpose of different ML and DL algorithms may be to improve data interpretation quality, consistency, and/or capacity in diagnostics. This paved the way for the formation of a sub-area called Deep Learning (DL) under the field of ML [5]. ML is a set of methods in mathematics and statistics for training a computer for specific tasks by finding hidden and useful patterns from inputs[6] .

There have been a lot of important efforts for automatic lung cancer classification by recurrent neural networks (RNN), convolutional neural networks CNN and other approaches used CT [7]. and other technique such as the Back Propagation neural (BPNN) and the Nearest K (KNN) [8]. Support vector machines (SVM) can be used when our data has exactly two classes [9]. Developed automated diagnosis technique for lung image CT scans (Deep Neural Network (DNN) and Linear Discriminate Analysis (LDA). Multiclass Multiple kernel learning classifier (m-MKL) [10] [11]. In Figure (1.2), an example of CT scans, some of which are for patients with lung cancer and polyps with some deep and machine learning algorithms applied.

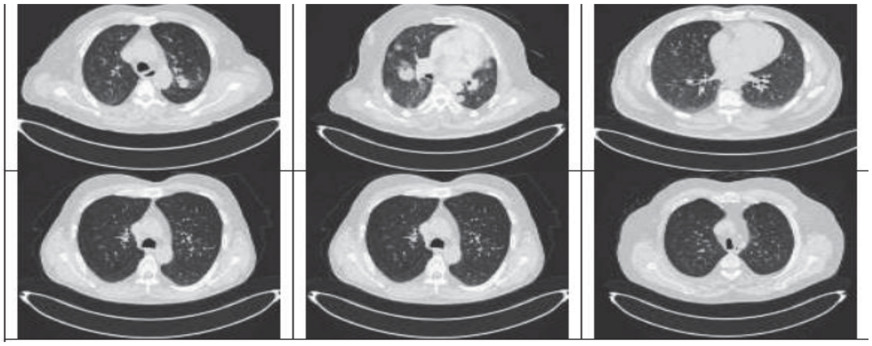

Figure1.2: an example of CT scans, some of which are for patients with lung cancer and polyps

In this work, SVM are supervised learning for diagnosis of lung cancer. When our data has exactly two groups or more, Support Vector Machines (SVM) can be used. To classify the data, the best hyper-plane is found, creating two groups of different data points.

The Naive Bayes algorithm is a basic probabilistic classification that calculates a probability set by calculating the rate and value combinations in a given data set. The algorithm uses the theorem of Bayes and concludes that the value of the class variable is independent of all variable.

Artificial neural networks (ANNs) the mathematical model that resolves classification and prediction problems. Inputs, outputs and (usually) hidden layers are neural network layers that transform the input into anything that can be used in the output layer, in many ways, hide layers translate the input into something that can be used by the output layer in two phases of testing and evaluation the ANN Model passes when a neural cancer prediction network is used when a dataset is used to train the network.

In this study, the artificial neural networks and Naïve Bayes and SVM algorithms have been used and good results are obtained from algorithms classification.

1. 2 Related Works

Many types of research and studies are dealing with the detection and diagnosis of lung cancer, among which includes the following:

B. Manju et al.(2021)[13]: They focuses on the need for an immersive learning system that enables effective condition for lung cancer detection in patients. Principal component analysis (PCA) stands out as a particularly effective algorithm for classifying the target groups. PCA successfully combines related qualities and creates a dissipated showcase of its constituents.

The number of principal components to be preserved is determined by examining the Scree plot. With a small amount of input, Support Vector Machines (SVM) outperforms other classification algorithms for classifying lung tumors. The collected components will be fed into the support vector machine for classification.

The pre-dangerous stage will remind health specialists to pay certain patients extra attention. The confusion matrix is used to calculate the expectation. The developed model has an accuracy of 0.87 and an error rate of 0.3 in the early detection of various stages of malignancy.

S.Jena & S.George (2020)[3]: They introduced kernel-based Non-Gaussian Convolutional Neural Networks were used to create a classification scheme for lung cancer in CT photos (KNG-CNN). Three convolutional, two fully connected and three pooling layers make up KNG-CNN. The diagnosis of False-Positives or errors found in the work is done using kernel-based Non-Gaussian computation.

The Lung Image Database Consortium image selection (LIDC-IDRI) dataset is initially used as input files and preprocessing steps include ROI-based segmentation using the powerful CLAHE method, which enhances images for improved feature extraction. Following the segmentation process, morphological features are extracted. In the end, for effective classification of tumors larger than 30mm, the KNG-CNN procedure is employed. This technique yielded an accuracy of 87.3 %.

O. Mohammed, et al. (2020)[14]: They suggested an artificial neural network to determine whether lung cancer exists in the human body or not. The data was gathered from the website of a data scientist. Chest pain, trouble swallowing, shortness of breath, coughing, wheezing, allergies, respiratory disease, exhaustion, anxiety, and yellowing fingers have also been used to detect lung cancer.

Additional information about the patient was used for the ANN model as input variables. The model has been learned and tested with a dataset on lung cancer. Evaluating and testing the new model. The rate of accuracy obtained was 99.01%.

P. Katiyar & K. Singh (2020)[15]**:** They used image recognition and machine learning to create a variety of computer-aided systems. Different segmentation, extraction of features, classification techniques like discreet wavelet transform, Gray level co-occurrence matrix, SVM, Artificial Neural Network and more are considered. Varied segmentation techniques are considered. The Deep Neural Network (DNN) had 97 % accuracy, CNN had 94 % accuracy, ANN had 99 % accuracy, and the SVM classifier had 96 % accuracy, according to the researchers.

Provides a review of current lung cancer detection technology as well as technological advancements over the last five years.

I. Nasser (2019)[16]**:** He proposed method for the detection of lack or presence of lung cancer in the human body using the Artificial Neural Network (ANN). Symptoms such as chest pain, swallowing difficulty, shorter breathing, coughing, allergy, tiredness, chronic disease, anxiety, and yellow fingers were used for the diagnosis of lung cancer.

They were used as the ANN input variables and other information about the person. The ANN is based on the data set whose title is Lung cancer surveys, which was developed, trained, and validated. The model assessment has shown that 96.67 % accuracy of the ANN model detects the absence of involvement of pulmonary cancer.

W. Choi et al. (2018)[17]**:** They created a prediction model of radiomics to improve the classification of the lung nodule (PN) in low dose CT. The Lung CT Screening Reporting and Data System (Lung-RADS) model for the early diagnosis of lung cancer is comparable with the American College of Radiology (ACR). Reviewed 72 PNs of the Lung Image Database Image

Selection (41 malignant and 31 benign) (LIDC-IDRI). Every PN has extracted 103 CT radiomic features.

A prediction model was created using a vector support machine (SVM), coupled with a minimum absolute shrinkage and selection operator (LASSO). (10×10-fold CV) cross-validation was used to test the SVM-LASSO model accuracy. The precision of the two features was 84.6% and 0.89 AUC.

□. Sathiya□riya & S. □enila (201□)[18]**:** They submitted a report to compare the various classifications to identify and estimate diseases of lung cancer. During their study they used the classification algorithm of SVM and Naïve Bayes. They got an accuracy of 80.75% for Naive Bayes, and 86% for SVM, and 86% for KNN, ND 86% FOR J48. It provides various outcomes for the lung cancer datasets for the use of classification algorithms. The consistency of the result is calculated by properly and incorrectly categorized instances by the classification techniques.

□. Christo□her□ & □. □amera □an□□ (201□)[19]**:** The predicted lung cancer has been analyzed by classification algorithms such as Naive Bayes and J48. Data from 100 cancer patients, as well as no cancer patients, were initially compiled, preprocessed, and analyzed using a lung cancer classification algorithm. There are 100 instances and 25 attributes in this dataset.

The main purpose of the paper is to alert users earlier and analyze classification algorithm results. Time Taken to build the model (seconds) is 0.03 for Bayesian Network and 0.01 for Naïve Bayes and 0.06 for j48, and after calculating the accuracy manually from the confusion matrix for each model were found to be accurate which is 100% for all model.

□. Chandran (201□)[20]**:** They proposed system for automatically classified for Lung diseases as Emphysema, Pleural effusion, and normal lung. The lung CT images are taken as input, preprocessing is applied, feature extraction is done by various methods such as Gabor filter extracts the texture features, Walsh Hadamard transform extracts the pixel co-efficient values, and a fusion method is proposed in this work which extracts the median absolute deviation values.

Feature selection including statistical correlation-based methods and Genetic Algorithm for searching in feature vector space are investigated. Four types of classifiers are used where the Multi-Layer Perceptron Neural Network (MLP-NN) classifier with the proposed fusion feature extraction method, genetic algorithm feature selection method gives the promising result of 91% accuracy than J48, K- Nearest Neighbor and Naïve Bayes classifiers.

1.□□ro□lem statement

Clinical decisions are usually decided depending on the doctor's intuition and expertise instead of the knowledge-rich data hidden in the database. This practice leads to undesired basics, excessive medical costs, and errors that affect the quality of service given to patients. lung cancer is a disease that requires early detection to determine it, whether it is benign or malignant.

Using neural networks and other Technique has shown outstanding results, with high flexibility in different environmental conditions, but its limitations in classification processes have led us to use other machine learning methods to solve this problem as it has achieved impressive results in the field of medical data classification because early detection leads to rapid treatment.

Prediction of lung cancer is a risky task, since it is directly dependent on people's life. Accuracy is a major factor, because it can be disastrous if not predict accurately. Therefore, lung tumor prediction is the problem of this thesis.

1.4 □im o□the thesis

The main aim of this thesis is to design intelligent system to detect and classify lung cancer with different models based on different preprocessing techniques (min-max normalize and z-score standardize) and their impact on machine learning algorithms to achieve a high degree of accuracy, in addition to making a comparison between These methods are to determine the best among them, the doctor can train the system on some known data with accurate detection of disease and giving reliability in decision making and rapid detection of lung cancer. These models are:

1- Design and implement such a powerful structure by using Artificial Neural Network (ANN) structure as the first classification approach.
2- The second model is Naïve Bayes (NB) that used for classification lung cancer
3- Support Vector Machine (SVM) is the third approach for classification approach lung cancer

1.□O□tline o□□hesis

The thesis contains four additional chapters in addition to the first chapter as described below:

Cha□ter □□o: □heoreti□al □a□□gro□nd

In this chapter, the linguistic aspect of the tools, techniques, and algorithms that will be applied in designing and implementing the proposed system is presented.

Cha□ter □hree: □he Pro□osed System

In this chapter the proposed system for diagnosing and detecting lung cancer is presented, where a detailed explanation of the tools and techniques used in classification is presented.

Cha□ter □o□r: □□□erimental □es□lts and □□al□ation

In this chapter, the results obtained from the implementation of the proposed system are presented, as well as the analysis and discussion of the results, their testing, and comparisons with previous studies.

Cha□ter □i□e: Con□l□sions and S□ggestions □or □□t□re □or□

A collection of observations from the design and application of the proposed system are presented in this chapter.