EVALUATION AND ANALYSIS OF A COMPUTER AIDED DIAGNOSTIC SYSTEM FOR LUNG NODULE ASSESSMENT IN CT SCANS

by

Berna Eser

B.S., in Biomedical Engineering, Başkent University, 2011

Submitted to the Institute of Biomedical Engineering
in partial fulfillment of the requirements
for the degree of
Master of Science
in
Biomedical Engineering

Boğaziçi University
2019

EVALUATION AND ANALYSIS OF A COMPUTER AIDED DIAGNOSTIC SYSTEM FOR LUNG NODULE ASSESSMENT IN CT SCANS

APPROVED BY:

Assoc. Prof. Dr. Albert Güveniş . . . .
(Thesis Advisor)

Prof. Dr. Cengizhan Öztürk . . . .

Prof. Dr. Olcay Çizmeli . . . .

DATE OF APPROVAL: 22 August 2019

ACKNOWLEDGMENTS

I would like to whole-heartedly express my appreciation and special thanks to my advisor Assoc. Prof. Dr. Albert Güveniş, for his continuous encouragement and support on my M.Sc research for many years.

I dedicate this study to my beloved uncle Selman Bayraktar, whom I lost to lung cancer. I hope you're sleeping in peace. I am thrilled and delighted to do something on behalf of you.

I would like to specially thank my dear friends, Başak Başar, Miray Aksın and Tuba Özdamar for their motivation.

My research would have been impossible without the aid and support of my parents and my sister, who supported me throughout the writing of this thesis and my years of study. I am proud to say that I am your daughter.

ACADEMIC ETHICS AND INTEGRITY STATEMENT

I, Berna Eser, hereby certify that I am aware of the Academic Ethics and Integrity Policy issued by the Council of Higher Education (YÖK) and I fully acknowledge all the consequences due to its violation by plagiarism or any other way.

Name : ______________________________ Signature: ______________________________

Date: ______________________________

ABSTRACT

EVALUATION AND ANALYSIS OF A COMPUTER AIDED DIAGNOSTIC SYSTEM FOR LUNG NODULE ASSESSMENT IN CT SCANS

Throughout the process of detecting lung cancer, using CT scans to predict the malignancy level of pulmonary nodules will be complicated process for radiologists. CAD gives a second opinion to radiologists to identify lesions properly and distinguish malignant nodules at the early stage of lung cancer. In order to develop the CAD scheme, a coherent and consistent database such as the Lung Image Database Consortium (LIDC) database is the most crucial point to consider. In that database, CT scans are evaluated by four different radiologists and their annotations on nodule characteristics are highly efficient for researchers. One of these characteristics is malignancy that has 5 ratings: Highly - moderately unlikely, indeterminate, moderately - highly suspicious. In this study, the classifier performances of SVM, RF and ANN are compared using 1018 cases, 907 nodules and 110 extracted features. Experimental results demonstrate that best performing classifiers are respectively ANN, SVM and RF on malignancy prediction. The most critical gap of LIDC Database is the lack of ground truth data that is mainly caused by the absence of biopsy results. Therefore, by using arithmetic mean voting, this problem might be avoided and desired information might be acquired. The results of analyses show that grouping radiologists' malignancy ratings increases classification accuracy. Classifiers are examined with the use of 5 class, 3 class (benign, indeterminate, malignant) and 2 class (benign, malignant) ratings on malignancy datasets. Experiments show that the classification performance is enhanced by grouping malignancy ratings. Three groups of datasets' classification results indicate that moderately and highly malignant separation assessments affect classification performance negatively. However, using two classes under the name of benign and malignant, increases the accuracy rate up to 98%.

Keywords: CAD, Lung Cancer Classification, ANN, SVM, RF.

ÖZET

BİLGİSAYARLI TOMOGRAFİ TARAMALARINDA AKCİĞER NODÜLÜ YORUMLAMALARI İÇİN BİLGİSAYAR DESTEKLİ TANI SİSTEMİ DEĞERLENDİRME VE ANALİZİ

Akciğer kanseri teşhisi sürecinde BT taramalarından nodüllerin kötü huyluluğunu tahmin etmek radyologlar için karmaşık bir süreçtir. Bilgisayar Destekli Tanı Sistemleri, lezyonları doğru şekilde tanımlamak için radyologlara ikinci bir fikir verir. BDT sistemi geliştirirken göz önünde bulundurulması gereken en önemli nokta anlaşılır ve tutarlı bir veri tabanıdır. Lung Image Database Konsorsiyumu (LIDC) Veri Tabanı, araştırmacılara dört radyoloğun nodüllerin karakteristik özellikleri ve konumları ile ilgili değerlendirmelerini içeren BT taramaları sunar. Kötü huyluluk 5 ayrı derecelendirme ile değerlendirilmiştir: Yüksek ve orta olasılıkla iyi huylu, belirsiz, orta ve yüksek olasılıkla kötü huylu. Bu çalışmada, SVM, RF ve ANN sınıflandırıcılarının performansı 1018 vaka, 907 nodül ve 110 özellik kullanılarak karşılaştırılmıştır. Deneysel sonuçlar, en iyi performans gösteren sınıflandırıcıların malignite tahmininde sırasıyla ANN, SVM ve RF olduğunu göstermektedir. LIDC veri tabanının en kritik eksiği, biyopsi sonuçlarının bulunmaması nedeniyle standart referans verisinin olmamasıdır. Bu nedenle, standart referansın belirlenmesinde ortalama oylama kullanılmıştır. Öte yandan, radyologların malignite sonuçlarının gruplandırılmasının sınıflandırma sonuçların doğruluğunu artırdığı görülmektedir. Sınıflandırıcılar, malignite veri kümeleri üzerinde 5 sınıf, 3 sınıf (iyi huylu, belirsiz, kötü huylu) ve 2 sınıf (iyi huylu, kötü huylu) olmak üzere test edilmiştir. Deneyler, radyologların malignite derecelendirmelerini gruplamanın, sınıflandırma performansını arttırdığını göstermektedir. Sonuçlar, orta ve yüksek derecede kötü huylu gibi ara değerlendirmelerinin sınıflandırma performansını olumsuz yönde etkilediğini göstermektedir. Buna rağmen, iyi huylu ve kötü huylu adı altında iki sınıf kullanılması, doğruluk oranını 98%'e kadar arttırmaktadır.

Anahtar Sözcükler: BDT, Akciğer Kanseri Nodül Nitelendirme, ANN, SVM, RF.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CAD	Computer Aided Diagnosis
BDT	Bilgisayar Destekli Tanı
CT	Computer Tomography
BT	Bilgisayarlı Tomografi
SVM	Support Vector Machines
ANN	Artificial Neural Networks
RF	Random Forest
LIDC	Lung Image Database Consortium
ROI	Region Of Interest
ROC	Receiver Operating Characteristics
AUC	Area Under Curve
MSE	Mean Square Error

1. INTRODUCTION

Cancer occurrence and accordingly death rates are highly increasing worldwide. Among these occurrences, lung cancer comes forward as the primary type, which causes a fatal threat for both men and women. However, it is possible to cure incidents since most of the cancer types might be detected and treated at an early stage. Detecting the disease at an early stage allows for malignancy rates to be decreased as well. Tracking the nodules that have been identified by imaging systems is also helpful in decreasing malignancy rates. The first imaging method used in analysing a suspicious case of lung cancer is the chest X-ray. The chest X-ray itself, which is a confidential method, is not sufficient enough for the ideal characterization and staging. However it helps determining prior knowledge on the situation [1].

In cases where further treatment is necessary, Computed Tomography is an of choice method applied for lung cancer imaging, which provides crucial information to radiologists. Radiologists utilize the CT scan and the information it provides to analyse malignancy predictions.

With the increase in the success and the susceptibility of the scan, the data produced by these systems increase accordingly as well. In addition, the increase in quantity of the data causes difficulties for the radiologists, which causes them to make mistakes in their interpretations. However, the difference of intensity in CT scan images and anatomical structure misinterpretation by clinicians might lead to a problem in marking the location of nodules [2], [3].

Consequently, schemes of Computer Aided Detection (CAD) are enhanced in order to detect suspicious lesions in medical images. CAD also provides an alternative point of view to radiologists, which decreases the amount of misjudgements. The CAD System consists of two processes.

The first stage is image processing to recognize nodules' coordinates via preprocessing methods and segmentation and last stage is classification using image features in order to decide nodules were identified as malignant or benign.

The most commonly used and referred dataset to enhance the CAD system which contains 1018 cases collected from five different universities (including University of Iowa, University of Chicago, Weill Cornell Medical College, University of California at Los Angeles and University of Michigan) is LIDC (Lung Image Database Consortium). Each one of the cases contains more than one lung CT scan with annotations obtained from four different radiologists. The lack of ground truth data and the insufficiency of knowledge on the degree of expertise of four radiologists compose the negative sides of LIDC dataset [4].

Up to date studies have provided consistent results in the development of detection of lung nodules and to identify malignancy rate using the Computer Aided Diagnosis System. The CAD System includes five major steps: Reading DICOM slices in the working database, pre-processing, segmentation of the ROI(Region of Interest), detecting tumour, extracting features and classification.

The LIDC database includes marked nodule areas that are determined by four different radiologists. As a consequence, the study proceeds without the need for nodule segmentation and tumour detection steps.

In view of eventual achievement of CAD scheme, the classification of defined nodule candidates is the fundamental stage. In addition, classification stage allows to reduce false positive rate. Thus, radiologists effectively interpret CT images and result of that increase the reliability of the CAD system [5]. Classification error is increasing caused by the unbalanced distribution of malignancy voting LIDC data. However, these challenges provide an opportunity to improve machine learning techniques to computer-aided diagnosis.

The main aim of this study is to support the development of a CAD scheme to assist the radiologists using with classification methods of ANN, SVM and RF in the characterization of lung nodules. Additionally, experiments show that grouping malignancy ratings have improved classification performance on publicly available LIDC database.

1.1 Lung Cancer Statistics

As reported by the World Health Organization in 2018, heart disease and stroke together comprise the first main reason of death around the world, followed by cancer subsequently as the second main reason of death which takes an estimated of 9.6 million people's lives [6]. Cancer comes into existence with uncontrolled cell growth which eventually turns into a disease. This happens when cells grow beyond their usual boundaries. The cells that exceed the limitations of enlargement might also expand their limits, spreading to other tissues or organs. The process of expanding is called metastasis. Likewise, the abnormal tissue masses that are created by overly evolved cells are called tumours, which are either classified as cancerous (malignant) or non-cancerous (benign) [7].

While numerous cancer types exist across the globe, lung cancer still holds the first place in being the best type of cancer which results in incidence and mortality rates. Lung cancer alone represents the reasoning behind approximately 1 in 5 (18.4%) cancer deaths which is comprised of 2.1 million newly diagnosed lung cancer patients new lung cancer incidences, and 1.8 million deaths estimated in 2018 [8].

There are two main subtypes of lung cancer. First of them being small-cell lung carcinoma, (SCLC) accounting for 15% of all cases of lung cancer and correspondingly the second one being non-small-cell lung carcinoma (NSCLC), accounting for the remaining 85% of the cases. Dissimilar cancer types require the necessity for different methods of treatment, which may include surgery for non-small cell lung carcinoma (NSCLC).

Conversely, in cases where small cell lung carcinoma (SCLC) is present, utilizing chemotherapy and radiation help in acquiring a better response [9], [10]. The main issue concerning the diagnosis is determining whether the nodules of the patients are benign or malignant. Nodules that are classified as highly likely to be malignant stop being examined closely and instead, a treatment period is started accordingly. However, in cases where nodules are thought to be benign, they are examined periodically with CT scans. Nevertheless, if there is a contradiction regarding the decision process of the probable situation, additional operations such as a biopsy might be required.

Accordingly, the diagnosis is affirmed with a biopsy, which usually takes the form of a bronchoscopy or a CT-guided biopsy. The histological type of malignant growth alongside the extent of spread (the stage) and the patient's performance rate combined designate the method of treatment and prognosis. If treated, patients' five-year survival percentage that follow the diagnosis increases to 14%. Correspondingly, a surgical approach, as well as chemotherapy and radiotherapy might be counted as possible treatment methods [11].

Being exposed to smoke that is exhausted from tobacco products in a long time period is the most common reason for lung cancer, effecting 90% of the cases. However genetic and biological aspects of the patient also play an important role. In addition to that, being exposed to toxic minerals such as asbestos and breathing air that contains harmful gases such as radon are amongst the reasons of lung cancer occurrence. Overall air pollution and breathing second hand smoke that comes from smokers are also within the causes of this specific type of cancer. One of these reasons or all of them combined might be why non-smoker patients, who constitute only the 10% of the incidents, are diagnosed with lung cancer [12], [13].

When it comes to lung cancer, there are no actual early symptoms, so it is often diagnosed simply when the status of the disease reaches an advanced stage. Accordingly, symptoms of lung cancer include cough, hemoptysis, chest pain and abnormal weight loss [8].

For this reason, the fact that there are no symptoms specifiable seen in the early stages of the disease indicates the necessity and emphasises the importance of the CAD systems.

1.2 Image Analysis in Chest CT

The simplest way to decrease high death rates caused by lung cancer is to diagnose the nodule as early as possible. However, the process of determining and evaluating the nodules especially with a smaller size requires a complicated process for radiologists.

The computerized tomography devices used today are systems that are capable of doing a scan, with high resolution and under millimetric susceptibility, on the entire chest. As stated by Awai et al. [14], using CT instead of analogy radiography increases the chance of detecting lung cancer. The detection success is almost 2.6 to 10 times higher with CT scans. With the increase in the success and the susceptibility of the scan, the data produced by these systems increase accordingly as well. In addition, the increase in quantity of the data causes difficulties for the radiologists, which causes them to make mistakes in their interpretations. A misdiagnosis on CT scans might be caused by several reasons, however most of them are caused by observer error. Observers might conduct an unsuccessful scanning session or they might be mistaken while recognizing the state of the nodules. They might also make mistakes in the decision-making process and misinterpret some of the precise characteristics such like size, location or conspicuity. Some technical issues might also cause problems. Regardless, these errors result in inaccurate information [15].

Utilizing multiple detector scanners and reshaping the slices into thinner portions help in detecting more nodules. This also results in using a higher quantity of thinner slices while conducting a study on imaging lung cancer. This increase in the quantity of images that are obtained at each CT examination makes the process of CT analysis much more time consuming, laborious and wearisome.

As a consequence, detection sensitivity concerning the nodules might be diminished, caused by the fatigue of the reviewer [16]. Meanwhile, processes such as detecting the nodule's location, eliminating the irrelevant data from the data set, coming up with suggestions on the characteristics of the nodule detected and separating different components of the lungs make it convenient for users. As researches conducted on this subject reveal, computer-aided diagnosis systems decrease the amount of misinterpretation made by radiologists and accordingly increase the amount of early diagnosis.

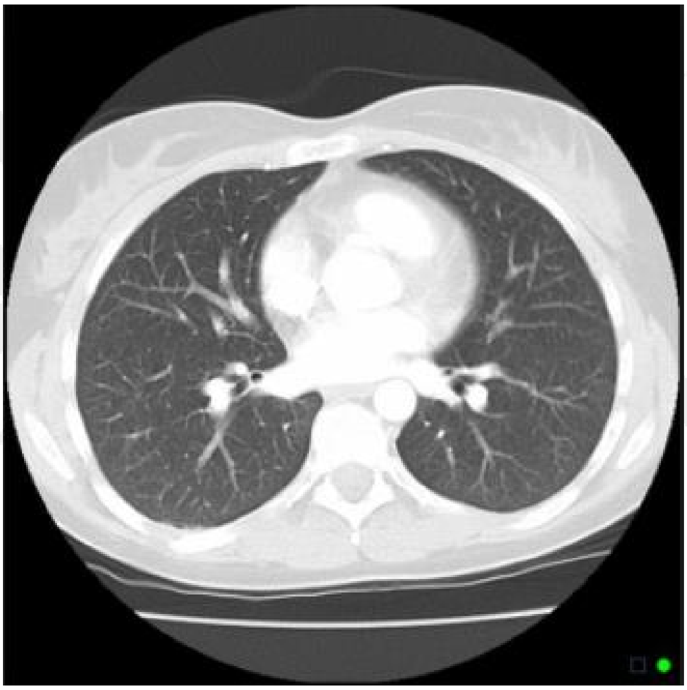


Figure 1.1 Example of CT Scan.

With technology progressing from day to day, imaging systems used in medical radiology has changed in a way that enables diagnosing the disease at an early stage. Today, screening methods such as chest radiography, CT, MRI and PET-CT make possible detecting and diagnosing even small size tumours. CAD systems are developed in order to help radiologists analyse the data procured from these devices. These systems provide radiologists with interpretation of images, detecting nodules and determination of their qualifications. They also assist radiologists by giving a second opinion in which the system help in to classifying nodules, marking conspicuous structures and sections to calibrate the lesion characterization [17].

Advantages of using CAD systems include [16], [18]:

- Assist in detecting cancer in an early stage,
- Improve accuracy rates in diagnosis,
- Decreasing the time spent by radiologists during an exam evaluation,
- Reducing misinterpretation ratio,
- Eliminating false positive,
- Prevents consuming extra time.

CAD systems are advanced for the purpose of enabling detecting the pulmonary nodules automatically in chest CT scans. In order to do so, researchers use a database, which is generally publicly available with using artificial intelligence methods.

A CAD system that helps in identifying nodules as benign or malignant is generally organized to conclude the main steps of feature extraction, classification and validation. Figure 1.2 demonstrates the block diagram representing all of these stages.

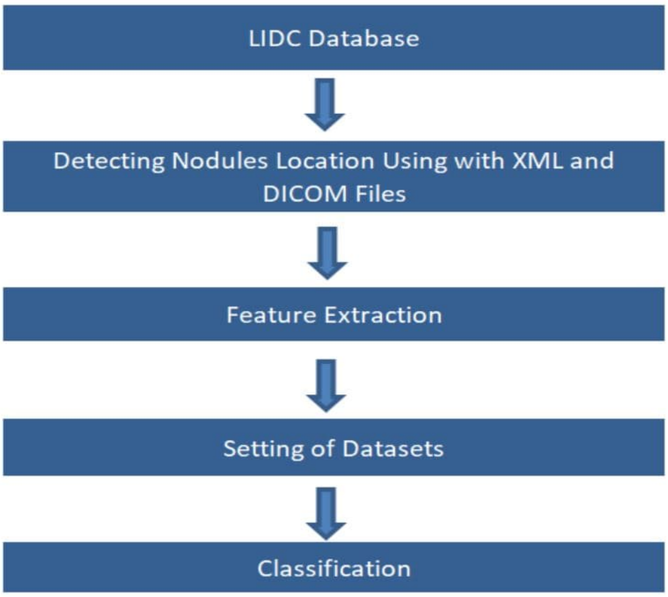


Figure 1.2 Flowchart of CAD System.

1.3 Literature Review

The aim of this study is to show that radiologists consensus on malignancy, influences the results of the classification. The LIDC Dataset lacks a base of ground truth and accordingly, mean voting is used for calculating the ground truth. In addition to that, the results of classification obtained from the SVM, RF and ANN are compared using literature.

Therefore, examining the works in literature under three groups will be useful in coming up with a better evaluation process for the outcomes of this work.

1.3.1 Impact of Reader Agreement on Reported CAD Performance

As proposed by Han et al. [19], 3D Haralick, Gabor, and local binary are three features to be compared consequently. The method of comparing the qualifications was validated with the SVM classifier, using the LIDC database by using 1012 nodules. The method recommended includes 3 groups of malignancy rate, showcasing “1” and “2” classified as benign whereas “4”and “5” as malignant and “3” as uncertain. The group “3” where classification is defined to be uncertain, two possible scenarios might be deployed, first of the scenarios being “3” to be grouped into benign class and on the contrary into malignant class in the second. By acknowledging the nodules with composite rank of malignancy “1” and “2” as benign and simultaneously “4” and “5” as malignant, the maximum accuracy rate of 0.94 is acquired, in order to utilize SVM classifier.

Kaya et al. [20], suggest a Computer Aided Diagnosis system for detecting a malignant nodules. In this study, LIDC database nodules are separated into 2 different groups, which voted 5 ratings and 3 ratings from radiologists. Ground truth is set by majority voting. In the 5 class experiment, the Rain Forests classifier obtained the highest classification accuracy, which is 0.8040. In the 3 class experiment, LDA classifier has a great accuracy, which is 0.8200.

Chen et al. [21], combined two different database, which are the LIDC and Beijing Friendship Hospital Clinical Database. The LIDC database categorized 3 different nodule sets (probably benign, uncertain, probably malignant) for training and clinical dataset for testing to classify all of the nodules with ANN. Additionally, building a system in order to compare radiologists' assessment for malignancy rate and biopsy results. The system reached AUC value of 79%, which was higher than the performance of the radiologists.

Moreover, a method suggested by Liu [22] states using 252 nodules that are larger than 3 mm in diameter from the LIDC database, and constituting a group of tumours for region agreement between radiologists in order to classify with ANN. The method of leave-one-out was implemented in the interest of verifying the productiveness of the ANN which has the highest correlation rate of 0.72.

A.Setio et al. [23] extend their work to determine pulmonary solid nodules larger than 10 mm with 3D segmentation algorithm. The total number of large solid nodules is 238 from the LIDC and majority voting used for ground truth. After feature extraction, 24 features based on shape, spatial context, blobness and intensity are computed. A nodule was considered to be solid in the event that larger part of the radiologists scored the texture characteristic higher than 3 (1 = ground-glass/non solid, 3= part-solid, 5= solid). SVM results and agreement levels evaluated together discloses a sensitivity rate of 98.3% and 94.1% large nodules at an average of 4.0 and 1.0 false positives rates, respectively.

Zinovev et al. [24] propose a scheme for evaluating the nodule characteristics, which includes combining probabilistic classifiers based on belief decision trees and ADABoost learning which helps in handling the uncertainty of the diagnosis process as a consequence of various annotations. With the intention of resolving the unbalanced data problem of LIDC, several adaptations were made to these system.

1.3.2 Studies Conducted on Designating the Ground Truth

Different ground truth assumptions, varying classification methods and dataset are caused by difficult comparison between these studies. Undoubtedly the success rate of a CAD system may vary considerably based on the selection of data that is being tested and what the ground truth holds. Although, the results are remarkable within related studies in literature.

Wiemker et al. [25] analysed eight different ground truth configurations with the usage of the LIDC Database. In the event of confirmation acquired by four of the radiologists, the CAD system designed reaches a sensitivity rate of 89%. It also acquires the same result, which is an average of false positive rates that comes forward as two per patient. An occurrence of conflict is much probable in the case of smaller nodules. Averagely, the radiologists detected only 66% of the nodules that have already been marked as a nodule by the other three radiologists. These results point out that a vast quantity of nodules, which are agreed upon by human expert observers, exist. In contrast, an equivalently broad set of noted nodules are not agreed upon, which means a consensus is not reached by the observers.

Accordingly, Angel et al. [?] implied nodules in computerized tomography scans that are 90 in number, with the method of dividing the process into two steps. Firstly, the process is executed blinded and the following unblinded with the purpose of seeing the results of the other three observers in the LIDC study. As the study suggests, for nodules that are 3 $\geq$ mm, there were exactly 174 nodules corresponding, whereas at least 1 of the observers decided that it was a nodule. For exactly 146 of the nodules corresponding, no less than 2 of the observers agreed and for a sum of 121 of them, at least 3 of the observers agreed. For 90 of the nodules corresponding, all four of the observers reached a consensus. These results demonstrate that human expert observers agree on a large amount of nodules. However, considering an equally large proportion of the nodules, there is no consensus reached among the observers.

1.3.3 Studies on ANN, SVM and RF Classifiers

By using the LIDC and ELCAP database, Hiram et al. [17] executed a test on 23 malignant and 22 benign nodules that are 2 mm to 30 mm in diameter. Features extracted from the GLCM in the wavelet domain and classified with SVM, classifier's accuracy is 82.22% whereas the sensitivity is at 90.90% and specificity is at 73.91%. Ground truth is not explained clearly.

Akram et al. [26] suggests using a method which using intensity based 2D and 3D statistical features. In the LIDC dataset, the nodules' range in diameter differentiate from 3 to 30 mm. Moreover, 47 computerized tomography scans that include nodules within are utilized in the computer-aided diagnosis system. The sensitivity rate of 96.31% is acquired by the SVM classifier with an accuracy rate of 96.54%, which emphasises an improvement concerning the existing CAD systems. Ground truth information is not sufficient.

By utilising 1191 nodules from a private dataset, Demir et al. [27] developed a CAD system which requires 2D and 3D pre-processing steps. The textual features of the surface are obtained by Gray Level Co-occurrence Matrix (GLCM) and extracted morphological characteristics from the volume of interests. SVM classification results obtain a sensitivity percentage of 98.03%, a selectivity percentage of 87.71%, an accuracy percentage of 90.12%, as well as a 2.45 ratio regarding false positive incidents.

In the study proposed by Matsuki et al. [28], a CAD system, which using the ANN, is developed in order to classify malignant nodules, which establishes 16 radiological findings and 7 clinical parameters. The numerical mean of the area that is under the curve of ROC has enhanced from 0.831 to 0.959 ($P < .001$). Through implementing the ANN results on high-resolution computerized tomography scans, it is understood that the accuracy ratio concerning the diagnosis, might be increased by utilizing a computerized design. Also differentiating 56 nodules that are benign from 99 nodules that are malignant serves as a helping medium during this process.

1.4 Organization of This Thesis

In this thesis, three different classification methods have been compared using the LIDC database in accordance with lung CT images. By grouping the malignancy scores of four different radiologists using images on the LIDC database, an observation is made on the effects of the radiologists consensus in classification outcomes.

The first chapter of the thesis includes an introduction and importance of Computer Aided Diagnosis System. The second chapter focuses on the materials and the methods deployed where the parts of the developed system are explained specifically. In this section, the phase that starts with the LIDC Database properties and feature extraction of the images are explained. At the concluding part of the chapter, the evaluation of the classification methods is given in terms of accuracy. Furthermore, the third chapter explains and discusses the results of the classification process by comparing the results of other studies. Finally, the last chapter summarizes the study and informs on the results of the study, and also commentates on future plans regarding the subject.

2. MATERIALS AND METHODS

2.1 LIDC Database

The development of computed tomography systems has a significant impact on the frequent application of this imaging technique. The systems are less affected by the patient's movements, and they are able to screen more frequently in order to obtain more detail, resulting in an increase in image quality, as well as a significant increase in data quantity. With these developments, the burden on radiologists is also increased accordingly.

CAD systems are widely used to facilitate the work of researchers in radiology as in many other medical fields. The biggest challenge for the systems developed is obtaining data that are correctly labelled and usable for generalization.

Nevertheless, a database that is directly derived from National Cancer Institute, which collects a various number of computed tomography (CT) scans exists, with the assistance brought in by the works of five different academic institutions which are University of Chicago, Cornell University, University of Michigan, University of California and University of Iowa. This database, which might be accessed publicly as the "Lung Image Database Consortium", has the objective to solely create a basis for developing, training and evaluating CAD techniques to researchers [29].

For the analysis of the CT scans, each one of the four institutions designate a radiologist. Each scan is conducted with annotations by a chest radiologist, using a process that consists of two phases.

Thus, the LIDC proposes a two phased data collection process that would [**?**]:

- Enable each scan to be reviewed by multiple expert readers,

- Clearly give the location of the nodule and spatial scope information acquired from each review's results in the form of a radiologist annotation,
- In the identification process of nodules, denotes between readers, the differences and the variability in the characterization of the boundaries of the nodule, as well as allowing for them,
- Permit the data accumulation process to be executed asynchronously with the purpose of not needing all radiologists to take part simultaneously in the imaging of a single scan.

Among the qualifications of the LIDC dataset, concerning the large majority of nodules, the lack of standard references acquired through a biopsy or through a follow up comes forward. Because of the fact that radiologists were not obliged to agree while analyzing the nodules present in the dataset, a variation of semantic ratings and outlines acquired from particular radiologists exists. There are several reasons behind this variability which includes the lack of ground truth data, the insufficiency of knowledge on the expertise degree of different radiologists and them being anonymous over varied nodules [24].

Consequently, the nodule is connected with several semantic ratings instead of one. In addition, these challenges give an opportunity that computer aided diagnosis to be fulfilled by implementing untraditional machine learning techniques.

Radiologists obliged to outline the boundaries of each nodule that is labelled as "nodule $\geq$ 3 mm" and to allocate subjective nodule qualifications. For every single nodule labelled as "nodule $<$ 3 mm", the radiologists are obliged to solely point out the centre of the nodule without defining the boundaries of the nodule and to illustrate the features of the nodule. Identically for the "non- nodule $\geq$ 3 mm", only marking the center is required. In the second phase of the process, every radiologist might review other radiologists' evaluations and they are also able to review and modify their own evaluations.

As a result of the two phased review process, a concluding annotation folder is generated for each CT scan. From the Web site of National Biomedical Image Archive (NBIA), a total group of CT images that conclude the relevant XML annotation files are accessible [30].

The characteristics of a nodule are malignancy, sphericity, texture, calcification, subtlety, lobulation, internal structure, spiculation, and margin. The table 2.1 demonstrates and expresses these qualifications, which are also called radiographic descriptors. Each of them is rated from 1 to 5 or 6 by four different radiologists.

Table 2.1
The Distribution of LIDC Nodule Characteristic.

Characteristics	Ratings		Characteristics	Ratings
Calcification	1. Popcorn		Sphericity	1. Linear
	2. Laminated			2. .
	3. Solid			3. Ovoid
	4. Non-central			4. .
	5. Central			5. Round
	6. Absent		Spiculation	1. Marked
Internal Structure	1. Soft Tissue			2. .
	2. Fluid			3. .
	3. Fat			4. .
	4. Air			5. None
Lobulation	1. Marked		Sublety	1. Extremely Subtle
	2. .			2. Moderately Subtle
	3. .			3. Fairly Subtle
	4. .			4. Moderately Obvious
	5. None			5. Obvious
	1. Highly Unlikely		Texture	1. Non-Solid
Malignancy	2. Moderately Unlikely			2. .
	3. Indeterminate			3. Part Solid
	4. Moderately Suspicious			4. .
	5. Highly Suspicious			5. Solid

In order to advance the level of development concerning Computer Aided Detection or facilitating diagnosing methods for lung nodules, accumulating a vast quantity of computed tomography scans and building a standard reference is at substantial importance.

In the LIDC Dataset, there is no ground truth data present, since annotators are not obliged to agree on characteristic ratings. On account of limited number of studies found in literature, in order to summarize the ratings, this study suggests to use mean voting method with numerous annotators. Also providing that the average produced has a comprehensive ratio that is not an integer number. Consequently, it is rounded to the nearest whole number.

Table 2.2
Rounding Values.

Rounding Values		
1.5	→	1
2.5	→	3
3.5	→	3
4.5	→	5

The final release of the LIDC Database includes 1018 cases right along with 2635 distinct nodules. Each scan in the LIDC dataset is paired with a correlating XML file that contains the outputs of a two-phased progress in which various amount of radiologists annotated on the image.

In this proposed study, selecting 907 distinct nodules which have ≥ 3 mm diameters and have an annotation by all of the 4 radiologists for malignancy degree.

2.2 Feature Extraction

The features informs about the characteristics of the item which might also be identified as the determinable properties of an image. The lung nodules detected after the scanning might be found on one or more CT scans.

Table 2.3
Distribution of Malignancy Rates.

Malignancy Degree	Number of Nodules
1	97
2	92
3	491
4	156
5	71
Total	907

In this work, by using MATLAB, a several features are calculated from the nodule areas that are created as a result of segmentation. Accordingly, in literature, features are seen to be classified as 2D, 2.5D and 3D. In this study, the 110 features extracted from the largest area of the nodule (2D) are subtracted from all the sections and the average of the features (2.5D) is taken into consideration. The features include shape, size and texture based forms. Whereas on MATLAB, some features are obtained by utilising the functions of “RegionProps” and “Properties”: Standard deviation eccentricity, solidity, circularity, aspect ratio, area of bounding box.

The features that concern Gray level co-occurrence matrix and Haralick texture [31] are extracted which takes place on the biggest region of the nodule and a mean of all slices of the nodule.

Haralick features comprise of contrast, correlation, energy, homogeneity, entropy, autocorrelation, dissimilarity, cluster shade, cluster prominence, inverse difference normalized and inverse difference moment normalised, maximum probability, difference entropy, sum entropy, difference variance, sum variance, sum average, information measures of correlation 1, information measures of correlation 2.

Features that are obtained from the Gray Level Co-Occurence Matrix- GLCM [31], [32]:

Contrast: Obtained by measuring the contrast between the pixels and their adjacency throughout the entirety of the image.

$$Contrast = \sum_{i,j} |i-j|^2 p(i,j) \tag{2.1}$$

Correlation: Measures the joint probability of specified pixel pairs.

$$Correlation = \sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\vartheta_i \vartheta_j} \tag{2.2}$$

Energy: Supplies the sum of squared components in the GLCM in order to demonstrate i and j matrix indexes.

$$Energy = \sum_{i,j} p(i,j)^2 \tag{2.3}$$

Homogenity: Calculates the closeness concerning distributing components in the GLCM and the GLCM diagonal in order to demonstrate i and j matrix indexes.

$$Homogenity = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \tag{2.4}$$

Entropy: It is the measurement unit that expresses how complex an image is. On a flat image, the entropy value equals zero. As complications increase in the image, the entropy value also increases accordingly. It is used to determine the compressibility of the image (i inclines entropy values).

$$Entropy = \sum_{i=1}^{n} p_i \log_2 p_i \tag{2.5}$$

Zernike Moments: That moments are complex polynomial sets that form a vertical set on a unit disc. They are insusceptible to rotating and insensitive to translation. That is the reason why nodules are transported to a 128x128 square with calculating their centers of gravity [33]. In this study, 60 features are extracted using Zernike moments.

2.3 Determinig Malignancy Datasets

In the LIDC Database, malignancy predictions of radiologists have 5 different ratings: 1 - highly unlikely, 2 - moderately unlikely, 3 - indeterminate, 4 - moderately suspicious and 5 - highly suspicious.

As consensus rates increase between radiologists, the number of samples falling into each class is expected to decrease. However, the decrease in the middle evaluation (2 - moderately unlikely and 4 - moderately suspicious) classes is much higher than in other classes. Accordingly, radiologists are more likely to reach a consensus on indefinite classifications (1 - highly unlikely , 3 - indeterminate , 5 - highly suspicious). The experiments are applied in class of 2 and 3, grouped according to their malignancy rates, and in class of 5 where such a grouping is not applied.

Table 2.4
Grouping Rates to 5 Class.

Number of 5 Class				
1 - Highly Unlikely	2 - Moderate Unlikely	3 -Indeterminate	4 - Moderate Suspicious	5 - Highly Suspicious
97	92	491	156	71

Table 2.5
Grouping Rates to 3 Class.

Number of 3 Class		
1-2 Benign	3- Indeterminate	4-5 Malignant
189	491	227

Table 2.6
Grouping Rates to 2 Class.

Number of 2 Class	
1 - 2 Benign	4 - 5 Malignant
189	227

2.4 Classification

The main objective behind classification in CAD systems are deciding on whether the nodules are malignant or benign with using features as guidance. A number of studies that do classification with using LIDC Database exist. Be that as it may, in this one in particular, 3 different datasets that are privatized and 3 different classification methods are used.

2.4.1 Support Vector Machine (SVM)

Support vector machines (SVM) are used in the construction of learning machines and they minimize the generalization error. In addition, they are an innovative approach created by positioning a group of planes that divide two or more classes of input [34]. By construction of these planes, the SVM brings forward the limits between the input data. The components of the input data which determine these limits are called support vectors.

The main objective of the SVM is to install an effective way of learning that is computerized, with the purpose of separating hyperplanes in high dimensional feature space.

As stated by Vapnick [35], support vector machines differentiate a group of binary training data. These data reserve a hyperplane that is distant from the two separate classes, which is respectively called the maximal margin hyperplane. The main goal is to develop a function f by using training data that comprises of N-dimensional patterns x_i and class labels y_i.

$$f\colon R^N \to \{\pm 1\} \tag{2.6}$$

$$(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2), \ldots\ldots (\mathrm{x}_1, \mathrm{y}_1) \in \left(\mathrm{R}^{\mathrm{N}}\mathrm{x}\{\pm 1\}\right) \tag{2.7}$$

Consequently, f will do a proper classification on new examples, which are x and y. In the case of separating the training data linearly being impossible, support vector machines might work efficiently using with the kernel techniques (comprises using the kernel trick). In the input space, the hyperplane which defines the SVM matches with a non-linear decision boundary. Thereby, support vectors help in expressing the function f, whereas it is computerized and calculated accordingly.

$$f(x) = \sum_{i=1}^{N_S} \left(a_i y_i K\left(s_i, x\right) + w_0\right) \tag{2.8}$$

In the case of K being the kernel function, a_i weight that is connected to the output of support vector machines is obtained as S_i being the support vector.

An additional advantage of the SVM is an automatic model selection, which helps in automatically obtaining the proper number and locations of the basis functions as long as training. Accordingly, the achievement of the SVM is mostly dependent on the kernel [36], [37].

The desired result concerning sets of training vectors, which are connected to different classes and are “m” in total by numbers, is: $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_m, y_m)$, where $x_i \in R_n$ indicates the i_{th} input vector and $y_i \in 1, 1$.

In addition, the objective obtained from the maximal margin classifier includes finding a hyperplane (where w: wx + b = 0) which is used to differentiate training samples [38]. Within the potential hyperplanes, just one of them maximizes the nearest data point of every class and the margin, which is the distance among the hyperplane.

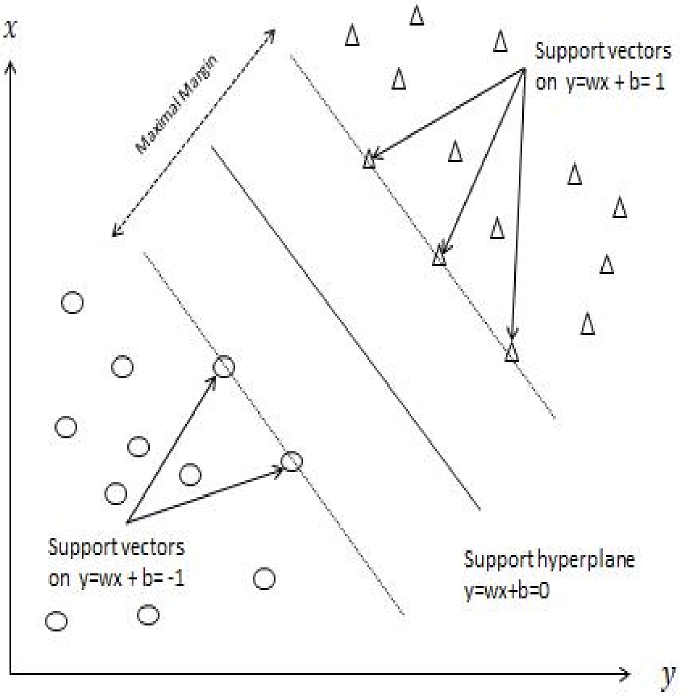


Figure 2.1 An example demonstrating the optimal separation of hyperplane that comes from the SVM, in accordance with the maximal margin between two hyperplanes that are parallel.

Figure 2.1 shows an example concerning the optimal dividing hyperplane alongside the largest margin. Points positioned on the verge of the margin are denoted by the support vectors. The function for convolution of the kernel, which is partly aligned with decision function and the support vectors, is the result to the classification. Consequently, the polynomial kernels, gaussian kernels, radial kernels and anova kernels are among the options in SVM-based CAD applications.

The most frequently used kernel functions concerning SVM analyses are as follows [39]:

- The Polynomial Kernel

$$K\left(x_{i,}x\right)=\left(x_{i}^{T}+1\right)^{p} \tag{2.9}$$

where $p>0$ is a constant and it explains the order of the Kernel.

- The Gaussian RBF Kernel

$$K(x,y) = \exp\left(-\frac{\|x - x_i\|^2}{2\delta^2}\right) \tag{2.10}$$

where $\delta > 0$ is a constant and it explains width of the Kernel.

During the process of training, the parameters p in Eq. 2.9 and δ in Eq. 2.10 are designated.

The resulting value of the SVM is -1 or 1. When the final value obtained from the output of a test sample is bigger than 0, the lesion present in the image will be classified as malignant. On the contrary, when the final value obtained from the output is smaller than 0, the lesion will be classified as benign [38].

Apart from other methods used when statistically classifying the nodules, SVMs reduce the structural risk to a minimum, which includes the possibility of misclassifying data that have never been seen. Thus in theory, the SVM is highly generalizable to test data. Also the SVMs transfigure the corresponding information present in the training set to a little quantity of support vectors which are solely used to classify new data. By doing so, the differentiating qualifications of the two classes are clearly distinguished.

As a result, it is obtained that the SVM is a usefull machine learning method used for classifying the system. It analyzes and identifies the categories using trained data. SVM is basically used for the classification purpose for diagnosing of the disease.

2.4.2 Random Forest (RF)

Random Forest is a ensemble algorithm which contains a large number of random decision trees in which it specifies the most appropriate one. No other generalization approach is necessary since this randomness is applied on selecting feature as well as defining trial/training data sets [40].

Some qualifications of the RF Algorithm:

- Produces reliable results on data with few examples or data with large size such as microarray data unlike many classification algorithms,
- Works efficiently under big data sets,
- Informs on which variable is more important for the classification process,
- Has an efficient way of predicting the missing information,
- Contains solutions for a possible scenario where data is unbalanced,
- Forests obtained might be saved for later use on other data,
- By expanding some qualifications, it can be used in setting and detecting the outlier on data that is not labeled.

Every decision tree made in the RF is constituted on another bootstrap sample obtained from the initial data set. A part of the data set is separated for testing and is not included in the training at the event creating each tree. Usually the training data is made up of 66% of the bootstrap sample, whereas trial data covers 34% of it [40].

During the process of making the decision trees and deciding on every node of the trees, k ($k < K$) feature selected from features that equals K in total is used. Features to be used in the node are designated with a measuring medium called gini importance. Correspondingly, evaluation of variable importance is made with calculations obtained from this phase. In the scope of the thesis, the parameter of the number of trees that is used for comparing is set to be 100 in classifications.

2.4.3 Artificial Neural Network (ANN)

The Artificial Neural Network or briefly named as ANN, comprises of a group of neurons that are interconnected. Neurons are cellular structures that have the ability to acquire and keep information with using experiential knowledge. In order to copy the structural construction of a neuron in a computer system, a structure named perceptron was developed in 1943 [41].

In case where $x^i \in R^{N_D}$ a 1 into x^i is appended, making it a R^{N_D+1} vector. The perceptron transfigures the input into a binary output $f^w(x) \in \{0, 1\}$ by considering [42], [43]:

$$f^w(x) = \begin{cases} 1 & if \quad w.x > 0 \\ 0 & otherwise \end{cases} \tag{2.11}$$

where $w \in R^{N_D+1}$ which equals to linear function $x \to w \cdot x$ followed by a non-linear activation function $\varphi(x)$ that can be equally written, is a vector of real-valued

$$\mathrm{f}^{\mathrm{w}}\left(\mathrm{x}^{\mathrm{i}}\right) = \varphi\left(\mathrm{w} \cdot \mathrm{x}^{\mathrm{i}}\right) \tag{2.12}$$

where $\varphi(x)$ is the Heaviside step function, defined by $\varphi(\mathrm{x}) = 1$ if $x > 0$ and 0 otherwise.

A substantial point is the selection of the parameter w, hence the perceptron defined by $f^w(x)$ might perform a specific task. The supervised setting in which there is a dataset

$$D = \left\{\left(x^2, t^2\right)_{i \in \{i_i \ldots Na\}}\right\} \tag{2.13}$$

with, for all i,

$$x^i \in R^{N_D+1}, \quad t^i \in \{0,1\} \tag{2.14}$$

By re-assessing the predictions at each one parameter bring up to date in the original perception algorithm, the parameters are repeatedly updated and the parameters that correspond with the incorrect predictions are modified.

As formality, $\mathrm{y^i} = \varphi\left(\mathrm{w} \cdot \mathrm{x^i}\right) \in \{0,1\}$ is the final product of the perceptron model. For the purpose of obtaining a set of parameters where the dataset D is thoroughly explained, stochastic parameter updates are utilized for every pair of training $(\mathrm{X_i}, \mathrm{y_i})$ in D, as follows:

$$\mathrm{w} \leftarrow \mathrm{w} + \left(\mathrm{t^i} - \mathrm{y^i}\right) \mathrm{x^i} \tag{2.15}$$

Moreover, the updates that are obtained from Eq.2.15 that implemented either for a number of iterations, fixated prior to the performance. Also, when errors concerning iterations are less than the predefined threshold, they are fixated again prior to the performance.

$$Error = \frac{1}{\mathrm{N_a}} \sum_{\mathrm{i=1}}^{\mathrm{N_a}} \left|\mathrm{t^i} - \mathrm{y^i}\right| \tag{2.16}$$

This design, which is basically constructed, can show its force when various perceptrons come together in order to work coherently. These perceptrons are usually put in order in layers. Each one of these layers take input from the previous one, apply weights and after points out to the next layer if possible.

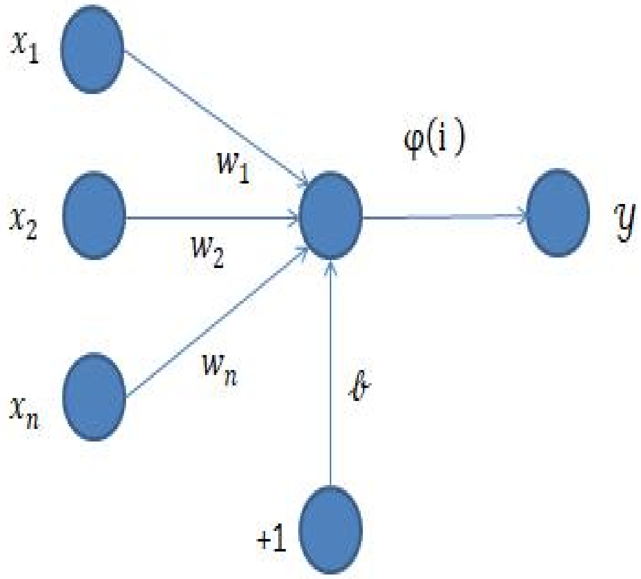


Figure 2.2 Simple Design of Perceptron.

In ANN, optimal performance is acquired when the weights are linked with connections in-between the layers. There are two manners of achieving this outcome, and most of the processes involve putting initials on the weights and letting the network use it as an example. Afterwards, a process called "back-propagation" takes place, which involves calculating the errors made by the network and feeding the outcome backwards. This process is used to update the weights. By repeatedly increasing the use of this process, the network acquires the knowledge that is required to categorize numerous different classes [44].

Accordingly, another one of these is feed forward Artificial Neural Network, which includes activation function that means neuron transferring data to another neuron. This function has the possibility to be linear or sigmoid [45]. When the sigmoid activation function is utilized, the calculation of the output of the j_{th} neuron is made accordingly:

$$z_j = \frac{1}{1+e^{-\delta\left(z_{-in_j}\right)}} \tag{2.17}$$

where z_{-in_j} defines the input of the neuron, which is obtained from the neuron on the previous layer, it is calculated as such:

$$z\_in_j = b_j + \sum_{i=1}^{n} x_i w_{ij} \tag{2.18}$$

where x_i describes the final output of the i_{th} neuron that was positioned on the former layer, n represents the overall amount of neurons that were positioned on the former layer accordingly. Moreover, w_{ij} defines the connection weight derived from the i_{th} neuron, with i_{th} neuron which is positioned on the previous layer. Whereas b_j represents the j_{th} neuron bias and r represents the steepness of the function connected to the sigmoid activation. During the learning process, both of w and j values are updated. The update requires the following equations [46]:

$$w_{ij}(new) = w_{ij}(\text{old}) + \Delta w_{ij} \tag{2.19}$$

$$b_j(new) = b_j old + \Delta b_j \tag{2.20}$$

$$\Delta w_{ij} = \alpha \delta_j x_i \tag{2.21}$$

$$\Delta b_j = \alpha \delta_j \tag{2.22}$$

where α defines the learning rate and δ defines the correction factor, the productivity of Artificial Neural Network is appraised by utilizing Mean Square Error (MSE) that is described as

$$MSE = \frac{1}{L} \sum_{k=1}^{L} \sum_{j=1}^{m} \left(t_j^k - y_j^{\mathrm{k}} \right)^2 \tag{2.23}$$

where the quantity of training pairs are shown with L, m defines the total number of neurons in the output layer; actual and target outputs at j_{th} neuron for k_{th} training pair are represented by y_j^k and t_j^k [47].

The set of weights, which defines a network architecture with a training array of input templates, determine the output of the network designated for each entries patterns. Together with the error occurred between the acquired network performance and the desired target performance, they constitute a potential multimodal reaction surface over a hyperspace which has sizes coinciding in according to the quantity of weights [39].

One of the problems occurring with the ANN approach is the over-fitting of the data. This problem occurs when the classifier recognizes excellent training examples, at the cost of being able to recognize a general input. This problem might be prevented by using cross-validation where the network is trained on one, and evaluated on another set of data. The network can be over-fitted if some error occurs in the validation set. As long as the previous networks are saved, the network can be taken back to the ground where the smallest error was present [45].

3. RESULTS AND DISCUSSIONS

The LIDC Database enables seeing annotations from four different radiologists on CT scans. For this reason the effect of the consensus on classification performance, which derives from radiologists â malignancy predictions are observed in this research. Correspondingly, the experiments are conducted on data sets with 2 and 3 classes, where malignancy rates are grouped, and data sets with 5 classes, where such grouping is not present. The performances of the SVM, RF and ANN are compared by using 907 nodules with 110 features.

Performances of the methods are compared by calculations on their accuracy, sensitivity and specificity. As explained in part 2.3, according to malignancy scores obtained from radiologists, the dataset is divided into three:

- Class 2, comprising scores of benign (1-2) and malignant (4-5),
- Class 3, comprising scores of benign (1-2), indeterminate (3) and malignant (4-5),
- Class 5, comprising scores of highly unlikely (1), moderately unlikely (2), indeterminate (3), moderately suspicious (4) and highly suspicious (5).

The accuracy ratio regarding the classification process is measured by detecting cases where the test groups are classified properly.

TP: True Positives, where the number of true value is malignant and the prediction number is also malignant.

TN: True Negatives, where the number of true value is benign and the prediction number is also benign.

FN: False Negatives, where the number of true value is benign while the prediction number is malignant.

FP: False Positives, where the number of true value is malignant while the prediction number is benign.

Analyzing the ROC graph and confusion matrix concerning the trained method is highly enough in terms of evaluating the accuracy rate of the designed classifier [48]. In the confusion matrix, the sensitivity and specificity for the tests are presented with including sensitivity as properly classified malignant nodules and specificity as properly classified benign nodules.

Confusion Matrix: The cross matrix holes contain the count of classes, which are classified properly, whereas the off diagonal cells contain instances that were misclassified. The whole percentage of properly classified instances are shown in green and rate of misclassified instances are shown in red by the blue matrix holes in the right down.

ROC Graph: Receiver Operating Characteristics curves for each one of the outputs are demonstrated by the colored lines in the ROC graph. In accordance with threshold being diversed, the ROC curve is a plot that shows the true positive rate means sensitivity or the false positive rate (1-specificity), depending on the variation. An ideal test would come up with a 100% sensitivity and a 100% specificity rate, which are shown by points in the upper side of the left edge.

3.1 Result of SVM

In this research, first method is SVM which selected for calculating the effects of consensus achieved by radiologists on the classification method.

The method is used in classifying nodules as benign or malignant, whereas test data sets and training sets are shaped separately. The shaping of them includes utilizing a five-fold cross validation technique, which requires the present data to be dissected into 5 random pieces, surely in accordance with their malignancy scores. The purpose of this is being able to generalize the results by using cross validation.

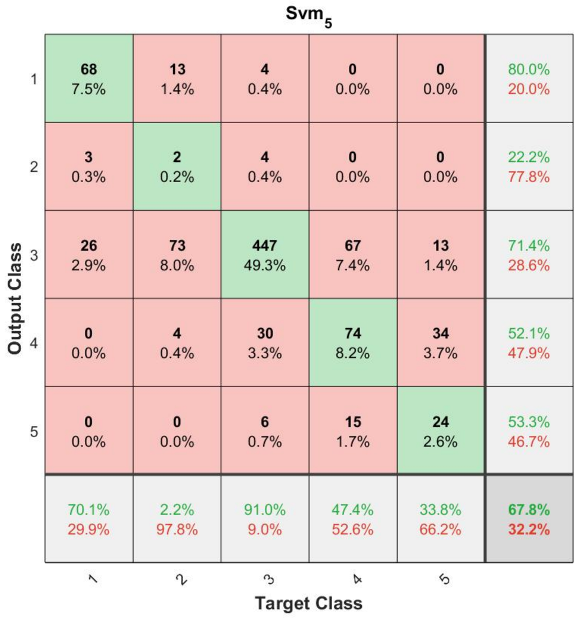
For the purpose of detecting less error connected to cross validation, the parameters of BoxConstraint, PolynomialOrder, KernelFunction and KernelScale were optimized. The higher accuracy was achieved when Kernel function Linear, the cost value is 0.10, the margin of error is 0.001.

When the data that are grouped under class of 2,3 and 5 are tested with the SVM classifier, the highest accuracy is obtained with the data that are grouped under a class of 2. Respectively, the accuracy levels of classification are 0.95, 0.76 and 0.68; whereas the CV errors are 0.048, 0.24 and 0.31.

ROC graphics are beneficial in eliminating possibilities, concerning the organization of the classifiers, as well as visualizing their quality performance.

After the calculation of the Receiver Operating Characteristics (ROC) curve, the area under the curve (AUC) was procured. The AUC assists in finding the discriminatory capability of the Support Vector Machine. Admittedly, a rating of 1.0 equals perfect discriminatory ability, whereas a rating of 0.5 inclines that there is no discriminatory ability [36].

As a results of the training data, The AUC acquired ratings of 0.9, 0.8 and 0.7.



Svm$_5$

Output Class \ Target Class	1	2	3	4	5	
1	**68** 7.5%	**13** 1.4%	**4** 0.4%	**0** 0.0%	**0** 0.0%	80.0% 20.0%
2	**3** 0.3%	**2** 0.2%	**4** 0.4%	**0** 0.0%	**0** 0.0%	22.2% 77.8%
3	**26** 2.9%	**73** 8.0%	**447** 49.3%	**67** 7.4%	**13** 1.4%	71.4% 28.6%
4	**0** 0.0%	**4** 0.4%	**30** 3.3%	**74** 8.2%	**34** 3.7%	52.1% 47.9%
5	**0** 0.0%	**0** 0.0%	**6** 0.7%	**15** 1.7%	**24** 2.6%	53.3% 46.7%
	70.1% 29.9%	2.2% 97.8%	91.0% 9.0%	47.4% 52.6%	33.8% 66.2%	**67.8%** **32.2%**

Figure 3.1 Confusion Matrix of SVM Classifier with 5 Class.

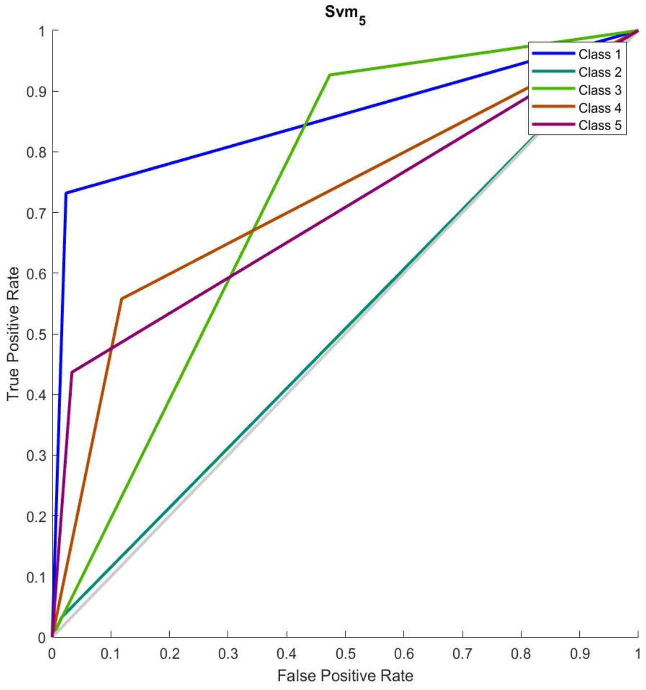
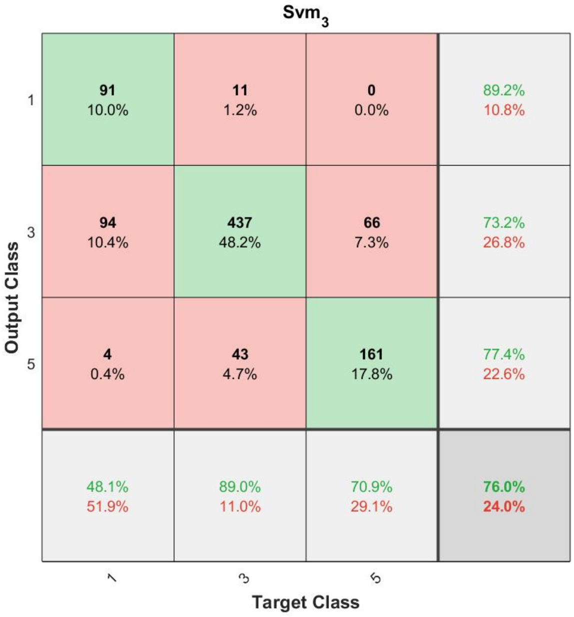


Figure 3.2 ROC Graph of SVM Classifier with 5 Class.



Svm$_3$

Output Class \ Target Class	1	3	5	
1	91 10.0%	11 1.2%	0 0.0%	89.2% 10.8%
3	94 10.4%	437 48.2%	66 7.3%	73.2% 26.8%
5	4 0.4%	43 4.7%	161 17.8%	77.4% 22.6%
	48.1% 51.9%	89.0% 11.0%	70.9% 29.1%	**76.0%** **24.0%**

Figure 3.3 Confusion Matrix of SVM Classifier with 3 Class.

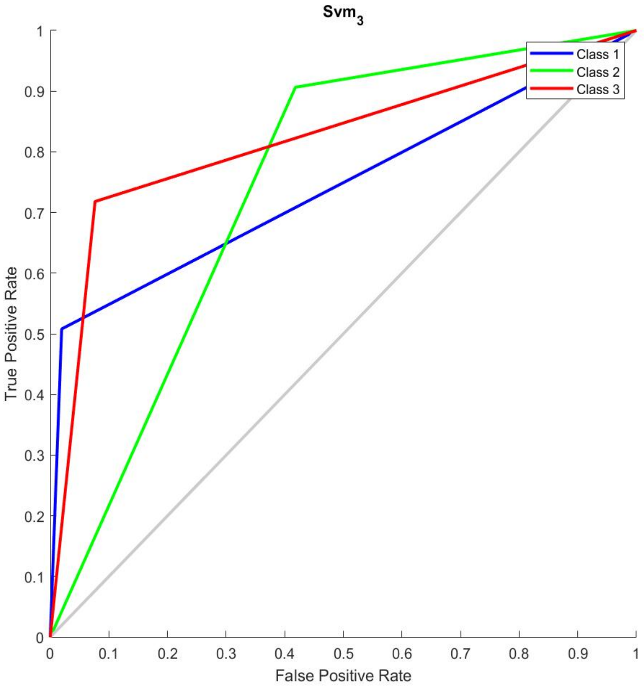


Figure 3.4 ROC Graph of SVM Classifier with 3 Class.

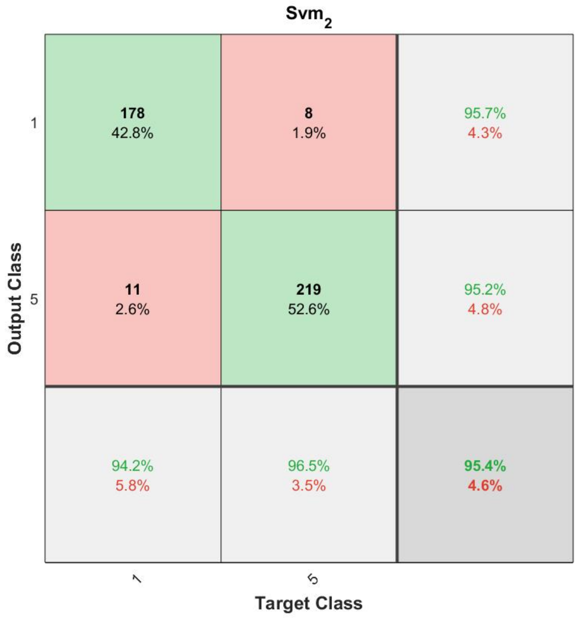


Figure 3.5 Confusion Matrix of SVM Classifier with 2 Class.

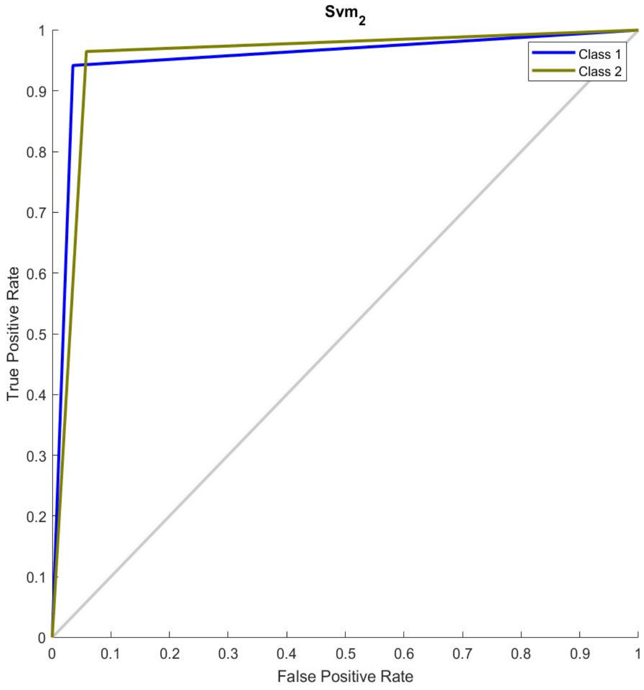
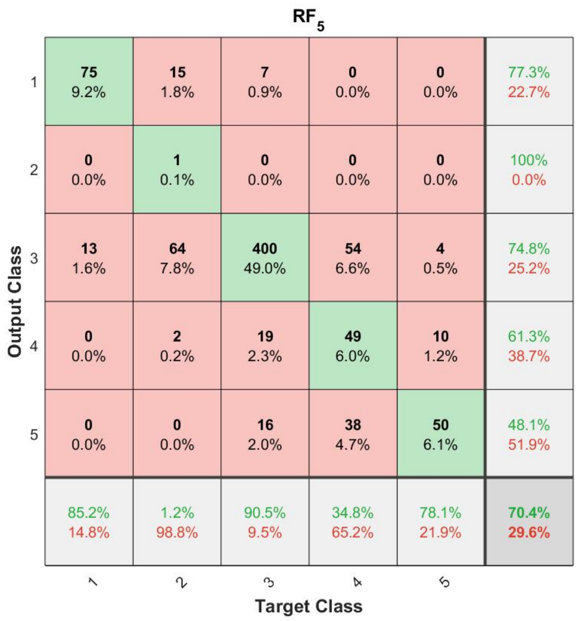


Figure 3.6 ROC Graph of SVM Classifier with 2 Class.

3.2 Result of RF

In this study, the RF method, which comprises of various decision trees, was utilized in order to conduct the classification. By the medium of sequential forward selection, nodes of the decision trees were formed with various treasured features. Accordingly, each one of the decision trees has the ability to make judgment. They are independent in their judgments, which are based on the features of the tree, subsequent to the phase of data training [49]. What this study tried to obtain was a comparison made by using different numbers of trees, utilizing a sum of 110 features, whose results are shown as follows:



RF_5

Output Class \ Target Class	1	2	3	4	5	
1	75 9.2%	15 1.8%	7 0.9%	0 0.0%	0 0.0%	77.3% 22.7%
2	0 0.0%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	13 1.6%	64 7.8%	400 49.0%	54 6.6%	4 0.5%	74.8% 25.2%
4	0 0.0%	2 0.2%	19 2.3%	49 6.0%	10 1.2%	61.3% 38.7%
5	0 0.0%	0 0.0%	16 2.0%	38 4.7%	50 6.1%	48.1% 51.9%
	85.2% 14.8%	1.2% 98.8%	90.5% 9.5%	34.8% 65.2%	78.1% 21.9%	70.4% 29.6%

Figure 3.7 Confusion Matrix of RF Classifier with 5 Class.

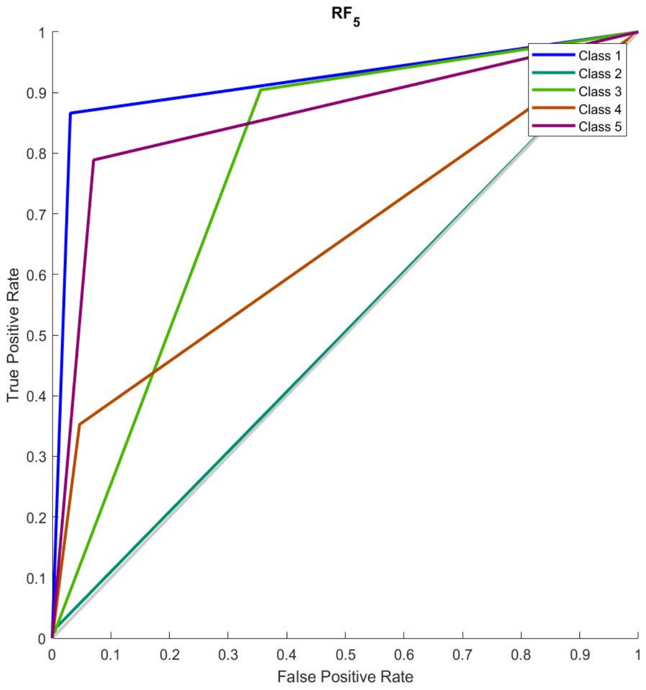
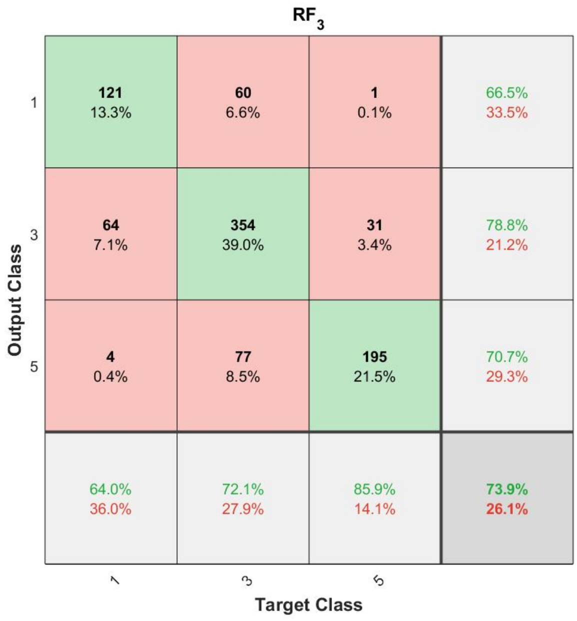


Figure 3.8 ROC Graph of RF Classifier with 5 Class.



	1	3	5	
1	121 13.3%	60 6.6%	1 0.1%	66.5% 33.5%
3	64 7.1%	354 39.0%	31 3.4%	78.8% 21.2%
5	4 0.4%	77 8.5%	195 21.5%	70.7% 29.3%
	64.0% 36.0%	72.1% 27.9%	85.9% 14.1%	73.9% 26.1%

Figure 3.9 Confusion Matrix of RF Classifier with 3 Class.

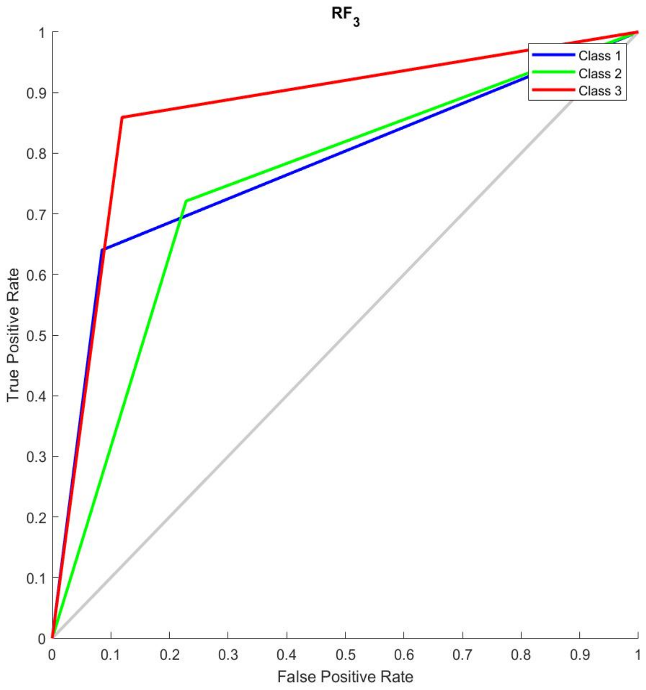
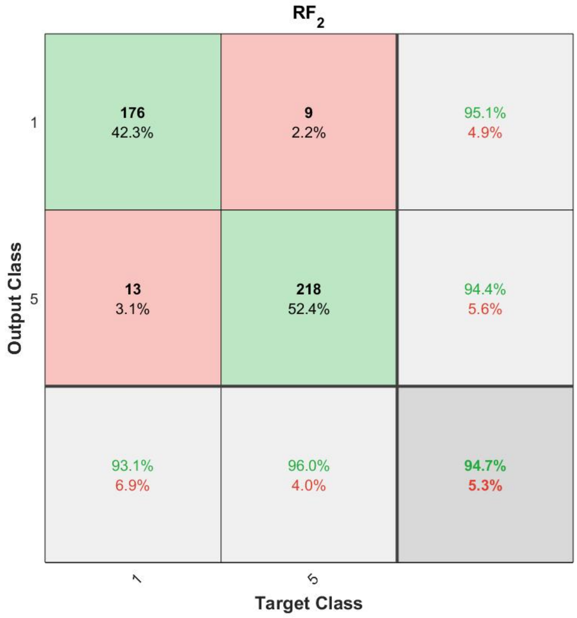


Figure 3.10 ROC Graph of RF Classifier with 3 Class.



	1	5	
1	**176** 42.3%	**9** 2.2%	95.1% 4.9%
5	**13** 3.1%	**218** 52.4%	94.4% 5.6%
	93.1% 6.9%	96.0% 4.0%	**94.7%** **5.3%**

Figure 3.11 Confusion Matrix of RF Classifier with 2 Class.

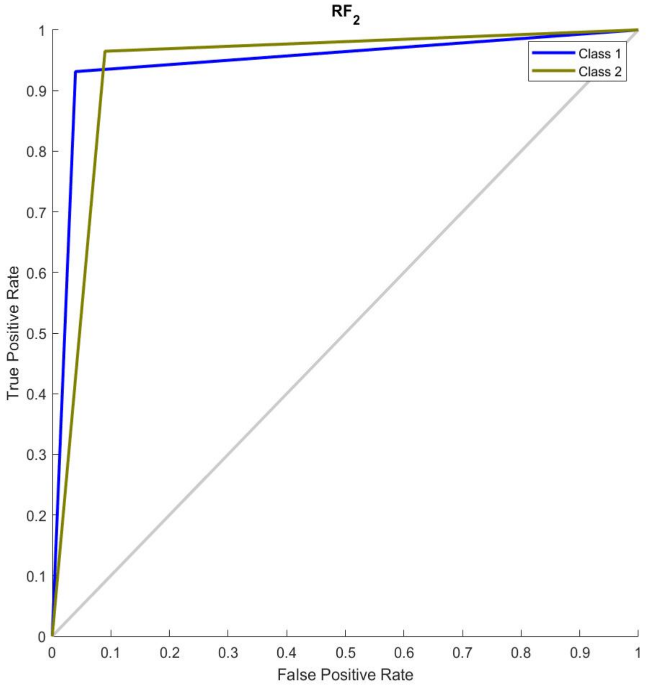


Figure 3.12 ROC Graph of RF Classifier with 2 Class.

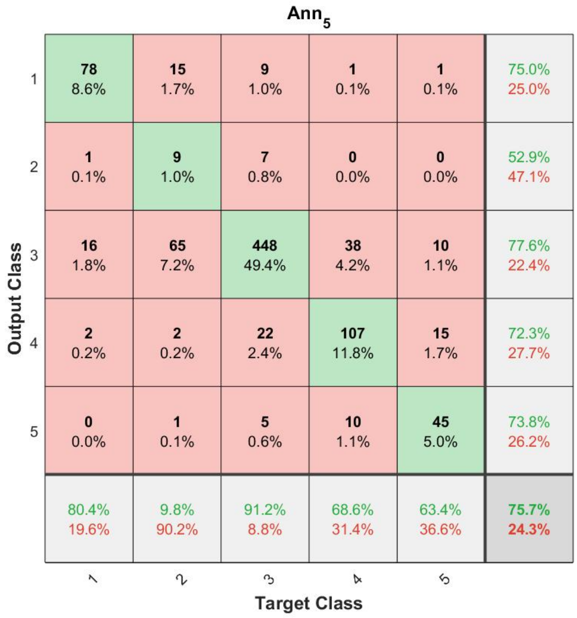
The RF classifier is optimized with the parameters of “Num Learning Cycles”, “Num Variables to Sample” and “Split Criterion”.

As demonstrated by the confusion matrixes above, in the cases where radiologists reach a consensus, classification made by utilizing 2 class scored higher in terms of accuracy, in comparison with class of 3 and 5. The highest accuracy is obtained with the data that are grouped under a class of 2. Respectively, the accuracy levels of classification are 0.947, 0.75 and 0.68; whereas the CV errors are 0,04, 0,23 and 0.27. However, the RF classification system is not more successful than the SVM method.

3.3 Results of ANN

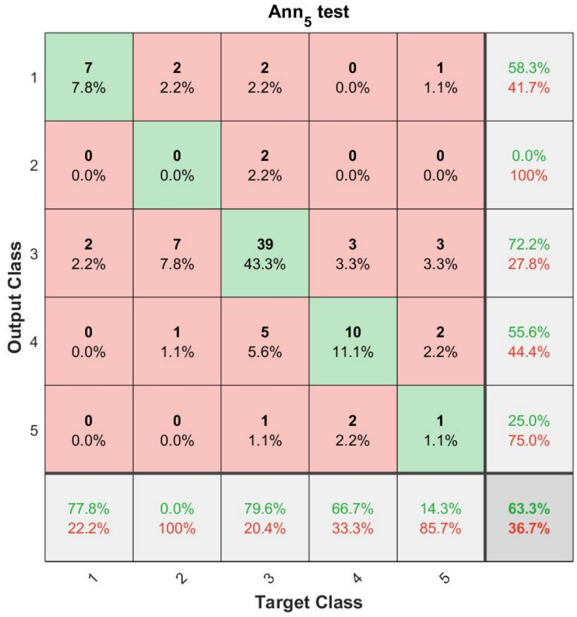
For the purpose of advancing the level of success of the network and aiming it to be trained afterwards, the target and input data must be placed in the network in order to be fed in it. Thereafter, a division is made by the network, which results in three differentiated samples instead of one. The samples are associated with the input data and they are classified under three categories in terms of what they will be used for. The different kinds of samples are respectively comprised of; ones that are used for training the system, ones that are used for validating and the ones that help to test the system. The samples that are used for training the network helps the network to change former knowledge in the case of an error. Moreover, the network generalization is calculated by the validation samples. They also stop the generalization when the generalization is no longer improving. In the event where the network errors still occur largely, a new training of the network might be required in order to obtain more accurate and effective result [50].

Out of the training dataset, 72% of the data is used for training, whereas 18% of it is used for validation and the remaining 10% for the intention of testing. In addition, the set of the test never changes, which is obtained by dissecting it to 10 parts. The amount of hidden neurons are 34 for class of 5, 25 for class of 3 and finally 18 for class of 2.


Ann$_5$
Output Class
Target Class

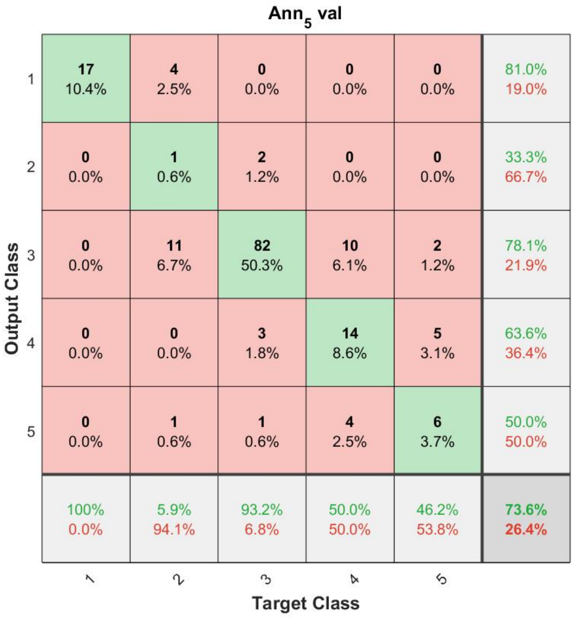
	1	2	3	4	5	
1	**78** 8.6%	**15** 1.7%	**9** 1.0%	**1** 0.1%	**1** 0.1%	75.0% 25.0%
2	**1** 0.1%	**9** 1.0%	**7** 0.8%	**0** 0.0%	**0** 0.0%	52.9% 47.1%
3	**16** 1.8%	**65** 7.2%	**448** 49.4%	**38** 4.2%	**10** 1.1%	77.6% 22.4%
4	**2** 0.2%	**2** 0.2%	**22** 2.4%	**107** 11.8%	**15** 1.7%	72.3% 27.7%
5	**0** 0.0%	**1** 0.1%	**5** 0.6%	**10** 1.1%	**45** 5.0%	73.8% 26.2%
	80.4% 19.6%	9.8% 90.2%	91.2% 8.8%	68.6% 31.4%	63.4% 36.6%	**75.7%** **24.3%**

Figure 3.13 Confusion Matrix of ANN Classifier with 5 Class.


Ann$_5$ test
Output Class
Target Class

	1	2	3	4	5	
1	**7** 7.8%	**2** 2.2%	**2** 2.2%	**0** 0.0%	**1** 1.1%	58.3% 41.7%
2	**0** 0.0%	**0** 0.0%	**2** 2.2%	**0** 0.0%	**0** 0.0%	0.0% 100%
3	**2** 2.2%	**7** 7.8%	**39** 43.3%	**3** 3.3%	**3** 3.3%	72.2% 27.8%
4	**0** 0.0%	**1** 1.1%	**5** 5.6%	**10** 11.1%	**2** 2.2%	55.6% 44.4%
5	**0** 0.0%	**0** 0.0%	**1** 1.1%	**2** 2.2%	**1** 1.1%	25.0% 75.0%
	77.8% 22.2%	0.0% 100%	79.6% 20.4%	66.7% 33.3%	14.3% 85.7%	**63.3%** **36.7%**

Figure 3.14 Confusion Matrix Testing Results of ANN Classifier with 5 Class.

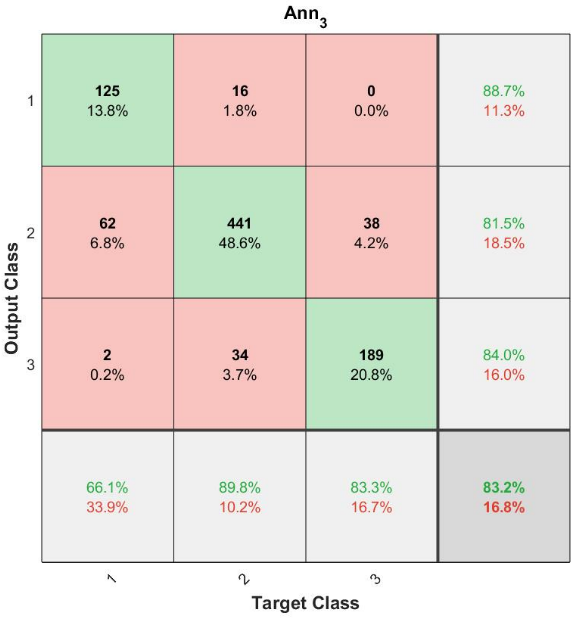


Ann$_5$ val

Output Class \ Target Class	1	2	3	4	5	
1	17 10.4%	4 2.5%	0 0.0%	0 0.0%	0 0.0%	81.0% 19.0%
2	0 0.0%	1 0.6%	2 1.2%	0 0.0%	0 0.0%	33.3% 66.7%
3	0 0.0%	11 6.7%	82 50.3%	10 6.1%	2 1.2%	78.1% 21.9%
4	0 0.0%	0 0.0%	3 1.8%	14 8.6%	5 3.1%	63.6% 36.4%
5	0 0.0%	1 0.6%	1 0.6%	4 2.5%	6 3.7%	50.0% 50.0%
	100% 0.0%	5.9% 94.1%	93.2% 6.8%	50.0% 50.0%	46.2% 53.8%	73.6% 26.4%

Figure 3.15 Confusion Matrix Validation Results of ANN Classifier with 5 Class.

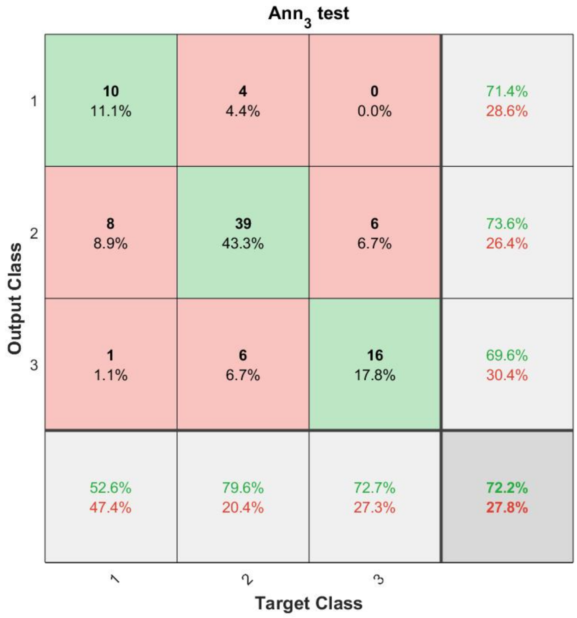
After evaluating the test results of class of 5, 3 and 2;it is seen that ANN is the most successful classification method with a performance of 0,983, 0.832, 0.757 respectively.



Ann$_3$

Output Class \ Target Class	1	2	3	
1	125 13.8%	16 1.8%	0 0.0%	88.7% 11.3%
2	62 6.8%	441 48.6%	38 4.2%	81.5% 18.5%
3	2 0.2%	34 3.7%	189 20.8%	84.0% 16.0%
	66.1% 33.9%	89.8% 10.2%	83.3% 16.7%	83.2% 16.8%

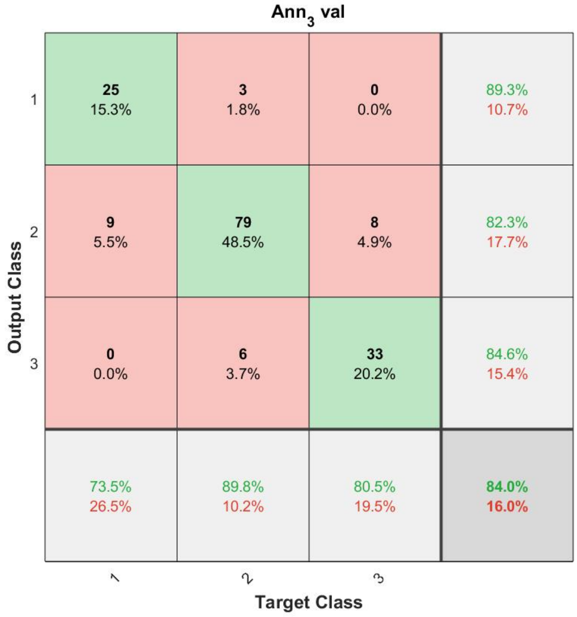
Figure 3.16 Confusion Matrix of ANN Classifier with 3 Class.



Ann$_3$ test

Output Class \ Target Class	1	2	3	
1	**10** 11.1%	**4** 4.4%	**0** 0.0%	71.4% 28.6%
2	**8** 8.9%	**39** 43.3%	**6** 6.7%	73.6% 26.4%
3	**1** 1.1%	**6** 6.7%	**16** 17.8%	69.6% 30.4%
	52.6% 47.4%	79.6% 20.4%	72.7% 27.3%	**72.2%** **27.8%**

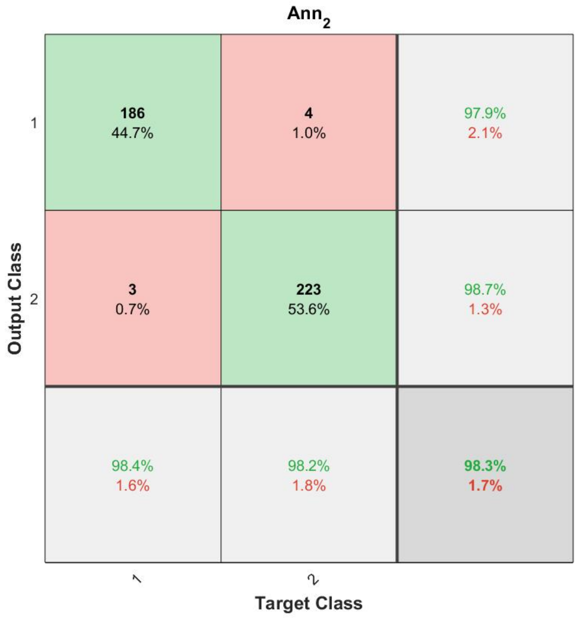
Figure 3.17 Confusion Matrix Testing Results of ANN Classifier with 3 Class.



Ann$_3$ val

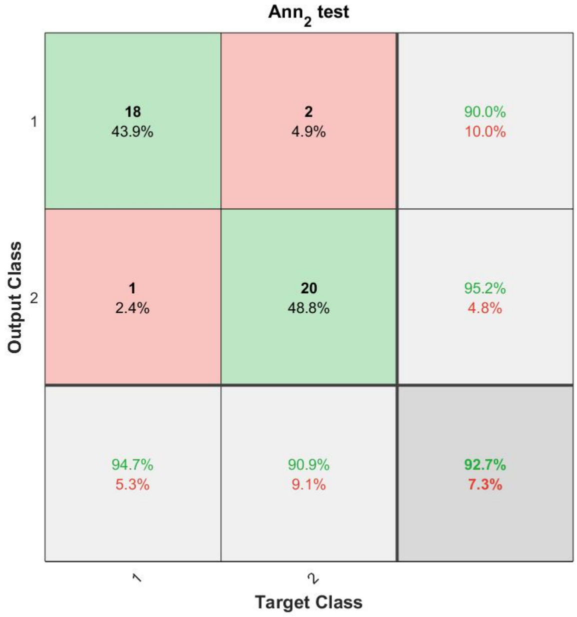
Output Class \ Target Class	1	2	3	
1	**25** 15.3%	**3** 1.8%	**0** 0.0%	89.3% 10.7%
2	**9** 5.5%	**79** 48.5%	**8** 4.9%	82.3% 17.7%
3	**0** 0.0%	**6** 3.7%	**33** 20.2%	84.6% 15.4%
	73.5% 26.5%	89.8% 10.2%	80.5% 19.5%	**84.0%** **16.0%**

Figure 3.18 Confusion Matrix Validation Results of ANN Classifier with 3 Class.



	1	2	
1	186 44.7%	4 1.0%	97.9% 2.1%
2	3 0.7%	223 53.6%	98.7% 1.3%
	98.4% 1.6%	98.2% 1.8%	98.3% 1.7%

Figure 3.19 Confusion Matrix of ANN Classifier with 2 Class.



	1	2	
1	18 43.9%	2 4.9%	90.0% 10.0%
2	1 2.4%	20 48.8%	95.2% 4.8%
	94.7% 5.3%	90.9% 9.1%	92.7% 7.3%

Figure 3.20 Confusion Matrix Testing Results of ANN Classifier with 2 Class.

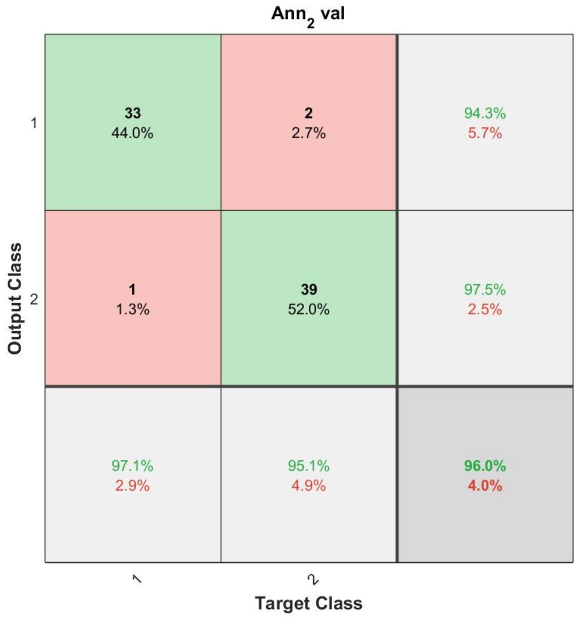


Figure 3.21 Confusion Matrix Validation Results of ANN Classifier with 2 Class.

An important issue concerning the structure of ANN is determining the learning rate constant. While training the networks, if not enough epochs are used, the model canât acquire enough knowledge, resulting in a decreased level of test and training set accuracies, which is also called an under-fitting. Conversely, if more epochs are used, the system might overflow with training samples by memorizing all of them, which is called an over-fitting. This suggests that the test has a low rating of accuracy even though it has a good accuracy rate of training sets. Accuracy mentioned corresponds with the ratio obtained by dividing the number of correct classification samples to all samples.

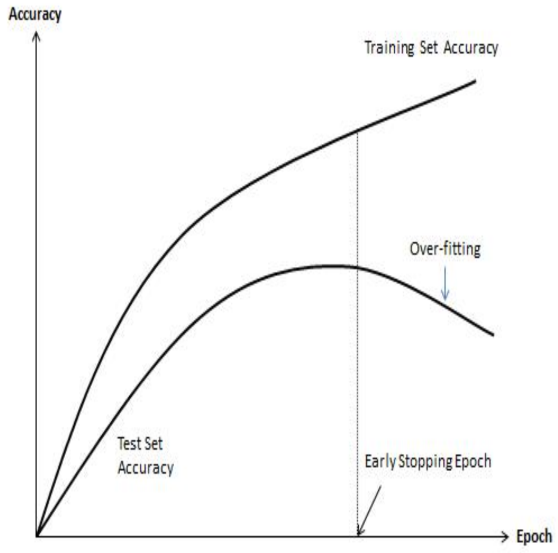


Figure 3.22 Analyzing Accuracy Rate of Different Epochs.

Analyzing accuracy rate of different epochs and their test sets, which shows the user when to stop doing a training on the network 3.3, might help with coming through this situation. However, a difference will exist at all times between the accuracy rates of training and test sets. When this difference starts to decline, it is taken as a sign to stop the process of training.

The best validation performances of groups of 2,3 and 5 are demonstrated in the graphics below. As the results show that there were no over-fitting or under-fitting obtained during the tests.

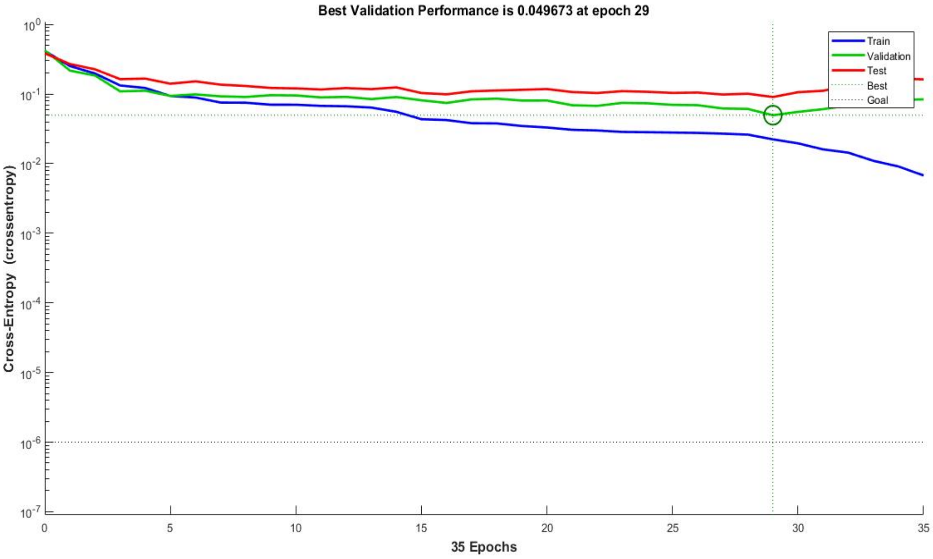


Figure 3.23 Results of Accuracy Rate of 29 Epochs.

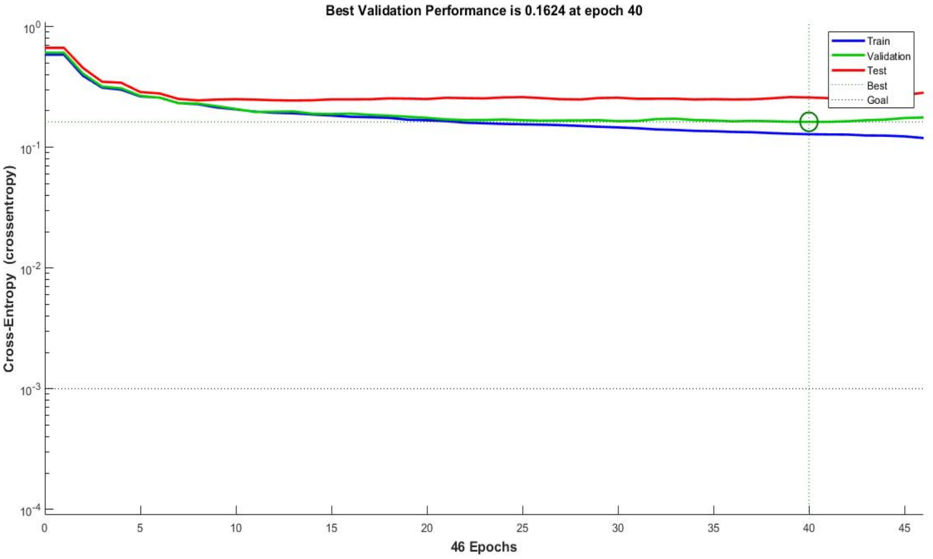


Figure 3.24 Results of Accuracy Rate of 40 Epochs.

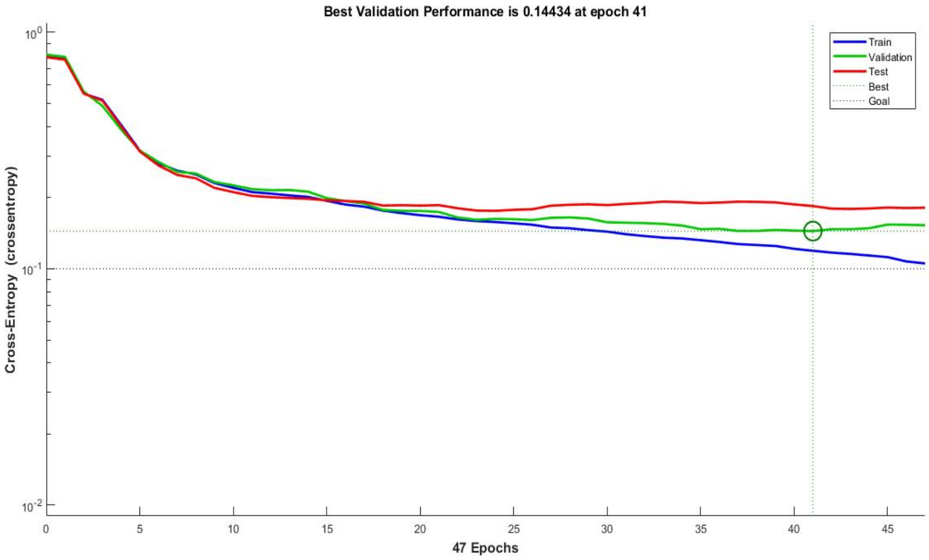


Figure 3.25 Results of Accuracy Rate of 41 Epochs.

3.4 Comparison of Classification Methods

Among the classifications that were made by three different training sets, the method of ANN acquired the highest score of accuracy, followed by the SVM and the RF methods.

In the dataset of 5 classes/groups, a limited number of moderate ratings (2 - moderately unlikely and 4 - moderately suspicious) that are not reached a consensus upon by radiologists, effected the results negatively.

The dataset of 2 and 3 classes has a higher rate of accuracy than the dataset of 5 classes. Similarly that results show that the moderate ratings effect the classification success adversely.

Table 3.1
Class of 5 Experiment Results.

Class 5 Experiment					
	Accuracy	Sensitivity	Specificity	CV Error	Hidden Neuron
SVM	0,6803	0,4225	0,6804	0,3197	-
RF	0,6836	0,6761	0,8454	0,2780	-
ANN	0,7574	0,6338	0,8041	-	34
ANNval	0,7362	0,4615	1,0000		

Table 3.2
Class of 3 Experiment Results.

Class 3 Experiment					
	Accuracy	Sensitivity	Specificity	CV Error	Hidden Neuron
SVM	0,75965	0,71366	0,46561	0,24035	-
RF	0,74972	0,87665	0,67725	0,23289	-
ANN	0,83241	0,83260	0,66138	-	25
ANNval	0,84049	0,80488	0,73529		

Table 3.3
Class of 2 Experiment Results.

Class 2 Experiment					
	Accuracy	Sensitivity	Specificity	CV Error	Hidden Neuron
SVM	0,9519	0,9471	0,9577	0,0481	
RF	0,9471	0,9604	0,9312	0,0482	-
ANN	0,9832	0,9824	0,9841	-	18
ANNval	0,9600	0,9512	0,9706		

On the basis of all methods, generally the measurements of sensitivity are lower than the specificity measurements.

This shows that the methods are more successful in determining the negative samples of methods, compared to positive ones.

In terms of calculating specificity, differences that are distinct as the success of classification exist between the methods. This signifies that the methods differ from each other in terms of determining samples as positive or negative.

3.5 Discussion

In the literature, Han et al. [19], Kaya et al. [20], Chen et al. [21], A.Setio et al. [23] and Zinovev et al. [24] have prominent studies on malignancy prediction using LIDC dataset. These studies based radiologists consensus on malignancy. In our study, malignancy ratings of radiologists were grouped as 2 classes, 3 classes and 5 classes and results are comparable with Han et al. [19] and Kaya et al. [20]'s studies.

Kaya et al. [20]'s reaches the highest classification accuracy with various classifiers, which is 0.8040 with 3 class of groups and Han et al. [19] is 94% with 2 class of groups. In our study, the highest accuracy is 98,32% with 2 class of groups. Although the classification methods are different, these two studies support of grouping the score of radiologists as well as our results.

Makes a comparison by taking into account the classification methods, literature's highest accuracies are Akram et al. [26] using SVM classifier and accuracy is 96.54%, Demir et al. [27] of 90.12% with SVM and Matsuki et al. [28]'s study has 95,9% accuracy level with ANN. Our studies highest accuracy is 98,32% with ANN classifier.

Considering the different ground truth assumptions, there are various gaps in the literature. We used mean voting as the standard reference in our study based on the Sauter et al. [51]'s study that refers using mean voting causes the most successful results for these kind of datasets. When the studies in the literature are examined, it is seen that the majority voting is used in general and there are not enough studies using mean voting to date.

4. CONCLUSION AND FUTURE WORK

4.1 Summary of The Study

Within the scope of this thesis, the malignancy of the nodule has been tried to be determined by SVM, RF and ANN methods using the LIDC database published by the American Cancer Research Institute. The effect of consensus, obtained from malignancy ratings of four different radiologists, on classification success is calculated.

One of the biggest problems of the LIDC database is the lack of standard references/ground truth and unbalanced data distribution. In order to determine the ground truth, mean voting method was used. In the literature, most of the studies conducted were made using the majority voting, and a few studies in the literature were conducted with mean voting.

Generally, the grouping of malignancy degrees has been observed to increase the classification performance, for the reason that it decreases the representation degree of intermediate degrees (2 - moderately unlikely and 4 - moderately suspicious), which were controversial for the radiologists. In the experiments on the three class data set; the aim is to reduce the scattering between evaluations by grouping the evaluations that are close to each other (5 - highly suspicious / 4 - suspicious and 1 - highly unlikely / 2 - unlikely). Additionally, in the 2-class dataset, in order to see how non-identifiable nodules effect the classification success, this class was removed from the data set.

It was observed that the consistency between the radiologistsâ evaluations and taking into consideration the linear relationship between radiographic descriptors have positive effects on the classification performance. In the statistical significance tests, it was observed that ANN classifier, which has the highest classification performance, was statistically superior to other methods.

In general, the sensitivity measurement on all methods is lower than the specificity measurement. This situation inclines that methods are more successful in detecting negative examples than detecting the positive ones.

When 5-class, 3-class and 2-class experiments were compared, considering all the methods compared and developed;

- The classification performance has increased by 25% on average,
- The sensitivity has increased by 41% on average,
- The specificity has decreased by 30% on average.

4.2 Future Work

In this study, 2 and 2.5 dimensional features were used. In future studies, 3D features can be included in the study. All features obtained from this study have been used. By using the method of feature election, the features which affect the success negatively might be eluded and the features which affect the success positively might be used. Accordingly, this would also contribute to the advancement of the work.

Additionally, median voting was used for ground truth. The study might be continued by using median or majority voting methods.

By using Deep Machine Learning methods, the classification performance on LIDC can be increased.

REFERENCES

1. Purandare, N., and V. Rangarajan, “Imaging of lung cancer: Implications on staging and management,” *Indian Journal of Radiology and Imaging*, Vol. 25, p. 109, 2015.

2. Makaju, S., P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, “Lung cancer detection using CT scan images,” *Procedia Computer Science*, Vol. 125, pp. 107–114, 2018.

3. Suzuki, K., M. Kusumoto, S. I. Watanabe, R. Tsuchiya, and H. Asamura, “Radiologic classification of small adenocarcinoma of the lung: Radiologic-pathologic correlation and its prognostic impact,” *Annals of Thoracic Surgery*, Vol. 81, pp. 413–419, 2006.

4. Vinay, K., A. Rao, and G. H. Kumar, “Predication of lung nodule characteristic rating using best classifier model,” *International Journal of Computer Applications*, Vol. 56, 2012.

5. Suzuki, K., “Machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey,” *IEICE Transactions on Information and Systems*, Vol. 96, pp. 772–783, 2013.

6. Lippie, G., and G. Cervellin, “Is digital epidemiology reliable? insight from updated cancer statistics,” *Annals of Translational Medicine*, Vol. 7, 2019.

7. Seyfried, T. N., and L. C. Huysentruyt, “On the origin of cancer metastasis,” *Critical Reviews in Oncogenesis*, Vol. 18, p. 43, 2013.

8. Bray, F., J. Ferlay, I. Soerjomataram, R. L. Siegel, A. L. Torre, and A. Jemal, “Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, Vol. 68, pp. 394–424, 2018.

9. Schiller, J. H., D. Harrington, a. C. L. C P Belani, A. Sandler, J. Krook, J. Zhu, and D. H. Johnson, “Comparison of four chemotherapy regimens for advanced non small cell lung cancer,” *New England Journal of Medicine*, Vol. 346, pp. 92–98, 2002.

10. Goldstein, S. D., and S. C. Yang, “Role of surgery in small cell lung cancer,” *Surgical Oncology Clinics of North America*, Vol. 20, pp. 769–777, 2011.

11. Jemal, A., R. Siegel, J. Xu, and E. Ward, “Cancer statistics, 2010,” *CA: A Cancer Journal for Clinicians*, Vol. 60, pp. 277–300, 2010.

12. Molina, J. R., P. Yang, S. D. Cassivi, S. E. Schild, and A. A. Adjei, “Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship,” *Mayo Clinic Proceedings*, Vol. 83, pp. 584–594, 2008.

13. Gorlova, O. Y., S. F. Weng, Y. Zhang, C. I. Amos, and M. R. Spitz, “Aggregation of cancer among relatives of never-smoking lung cancer patients,” *International Journal of Cancer*, Vol. 121, pp. 111–118, 2007.

14. Awai, K., K. Murao, A. Ozawa, M. Komi, H. Hayakawa, S. Hori, and Y. Nishimura, “Pulmonary nodules at chest CT: Effect of computer-aided diagnosis on radiologists’ detection performance,” *Radiology*, Vol. 230, pp. 347–352, 2007.

15. Ciello, A. D., P. Franchi, A. Contegiacomo, G. Cicchetti, L. Bonomo, and A. R. Larici, “Missed lung cancer: When, where, and why?,” *Diagnostic and Interventional Radiology*, Vol. 23, pp. 118–126, 2017.

16. Yuan, R., P. M. Vos, and P. L. Cooperberg, "Computer-aided detection in screening CT for pulmonary nodules," *American Journal of Roentgenology*, Vol. 186, pp. 1280–1287, 2006.

17. Hiram, M. O., O. O. Villegas, G. G. C. Sánchez, O. Domínguez, and A. Nandayapa, "Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine," *BioMedical Engineering Online*, Vol. 14, pp. 1–20, 2014.

18. Firmino, M., R. M. Mendoça, M. R. Dantas, A. H. Morais, R. Valentim, and H. R. Hekis, "Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects," *BioMedical Engineering OnLine*, Vol. 13, p. 41, 2014.

19. Han, F., G. Zhang, H. Wang, B. Song, H. Lu, D. Zhao, H. Zhao, and Z. Liang, "A texture feature analysis for diagnosis of pulmonary nodules using lidc-idri database," in *2013 IEEE International Conference on Medical Imaging Physics and Engineering*, pp. 14–18, IEEE, 2013.

20. Kaya, A., and A. B. Can, "A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics," *Journal of Biomedical Informatics*, Vol. 56, pp. 69–79, May 2015.

21. Chen, H., Y. Xu, Y. Ma, and B. Ma, "Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images," *Academic Radiology*, Vol. 17, pp. 595–602, 2010.

22. Liu, X., F. Hou, H. Qin, and A. Hao, "A CADe system for nodule detection in thoracic CT images based on artificial neural network," *Science China Information Sciences*, Vol. 60, pp. 1–15, 2017.

23. Setio, A. A., C. Jacobs, J. Gelderblom, and B. V. Ginneken, "Automatic detection of large pulmonary solid nodules in thoracic CT images," *Medical Physics*, Vol. 42, pp. 5642–5653, 2015.

24. Zinovev, D., J. Furst, and D. Raicu, "Building an ensemble of probabilistic classifiers for lung nodule interpretation," *IEEE*, Vol. 2, pp. 155–161, 2011.

25. Wiemker, R., M. Bergtholdt, E. Dharaiya, S. Kabus, and M. C. Lee, "Agreement of CAD features with expert observer ratings for characterization of pulmonary nodules in CT using the LIDC-IDRI database," *Medical Imaging 2009: Computer-Aided Diagnosis*, Vol. 7260, 2009.

26. Akram, S., M. Y. Javed, A. Hussain, F. Riaz, and M. U. Akram, "Intensity-based statistical features for classification of lungs CT scan nodules using artificial intelligence techniques," *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 27, pp. 737–751, 2015.

27. Demir, Ö., and A. Y. Çamurcu, "Computer-aided detection of lung nodules using outer surface features," *Bio-Medical Materials and Engineering*, Vol. 26, pp. 1213–1222, 2015.

28. Matsuki, Y., K. Nakamura, H. Watanabe, T. Aoki, H. Nakata, S. Katsuragawa, and K. Doi, "Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT," *American Journal of Roentgenology*, Vol. 178, pp. 657–663, 2013.

29. Wang, W., L. Jiawei, Y. Xuedong, and L. Hongli, "Data analysis of the lung imaging database consortium and image database resource initiative," *Academic Radiology*, Vol. 22, pp. 488–495, 2015.

30. McNitt-Gray, M. F., G. ArmatoIII, C. R. Meyer, A. P. Reeves, G. McLennan, R. C. Pais, J. Freymann, M. S. Brown, R. M. Engelmann, and P. H. Bland, "The lung image database consortium data collection process for nodule detection and annotation," *Academic Radiology*, Vol. 14, pp. 1464–1474, 2007.

31. Haralick, R. M., and L. G. Shapiro, *Computer and robot vision*, Vol. 1, Boston: Addison-Wesley Reading, 1992.

32. Haralick, R. M., and K. Shanmugam, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 610–621, 1973.

33. Zernike, F., "Diffraction theory of the knife-edge test and its improved form, the phase-contrast method," *Monthly Notices of the Royal Astronomical Society*, Vol. 94, pp. 377–384, 1934.

34. Rejani, A., Y. Ireaneus, and S. Thamarai, "Early detection of breast cancer using SVM classifier technique," *International Journal on Computer Science and Engineering*, Vol. 1, 2009.

35. Vapnik, V., *Statistical Learning Theory*, New York: John Wiley and Sons, 1998.

36. Girosi, F., M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Computation*, Vol. 7, pp. 219–269, 1995.

37. Smola, A. J., B. Schölkopf, and K. R. Müller, "The connection between regularization operators and support vector kernels," *Neural Networks*, Vol. 11, pp. 637–649, 1998.

38. Huang, Y. L., "Computer-aided diagnosis using neural networks and support vector machines for breast ultrasonography," *Journal of Medical Ultrasound*, Vol. 17, pp. 17–24, 2009.

39. Rizzi, M. M., D. Matteo, C. Guaragnella, and B. Castagnolo, "Health care improvement: Comparative analysis of two CAD systems in mammographic screening," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 42, pp. 1385–1395, 2012.

40. Breiman, L., "Random forests," *Machine Learning*, Vol. 45, pp. 5–32, 2001.

41. Corbett, D. M., "Stream-gaging procedure, a manual describing methods and practices of the geological survey," tech. rep., US Govt., 1943.

42. Rosenblatt, F., "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, Vol. 65, p. 386, 1958.

43. Bishop, C. M., *Pattern Recognition and Machine Learning*, Cambridge: Springer, 2006.

44. Leema, N., H. K. Nehemiah, and A. Kannan, "Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets," *Applied Soft Computing*, Vol. 49, pp. 834–844, 2016.

45. Haykin, S. S., *Neural Networks and Learning Machines*, New York: Prentice Hall, 2009.

46. Fausett, L., *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*, New Jersey: Prentice-Hall, Inc., 1994.

47. Khehra, B. S., and A. P. S. Pharwaha, "Classification of clustered microcalcifications using MLFFBP-ANN and SVM," *Egyptian Informatics Journal*, Vol. 17, pp. 11–20, 2016.

48. Demuth, H., M. Beale, and M. Hagan, "Neural network toolbox tm 6 user's guide matlab," *The MathWorks*, 2013.

49. Liu, J. K., H. Y. Jiang, C. G. He, Y. Wang, P. Wang, and H. M. others, "An assisted diagnosis system for detection of early pulmonary nodule in computed tomography images," *Journal of Medical Systems*, Vol. 41, p. 30, 2017.

50. Basheer, I. A., and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of Microbiological Methods*, Vol. 43, pp. 3–31, 2000.

51. Sauter, J. N., V. M. LaBarre, J. D. Furst, and D. S. Raicu, "An evaluation of consensus techniques for diagnostic interpretation," in *Medical Imaging 2018: Computer-Aided Diagnosis*, Vol. 10575, pp. 1057538–10, International Society for Optics and Photonics, 2018.