National College of Ireland

Lung Cancer Detection Using Classification Algorithms

MSc Research Project
Data Analytics

Sumit Jadhav
Student ID: 18129633

School of Computing
National College of Ireland

Supervisor: Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Sumit Jadhav
Student ID:	18129633
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Muhammad Iqbal
Submission Due Date:	12/12/2019
Project Title:	Lung Cancer Detection Using Classification Algorithms
Word Count:	6817
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Lung Cancer Detection Using Classification Algorithms

Sumit Jadhav
18129633

Abstract

Diagnosing lung cancer with high accuracy is most critical to make a significant change in survival rate. For diagnosing lung cancer different imaging techniques are used by radiologists such as Magnetic Resonance Imaging (MRI), Computer tomography (CT) and X-ray. These techniques help to detect the benign or malignant nodules present in the lungs but with certain limitations to the human eye. This study proposes to build a classification system that can identify the benign and malignant nodules and provide better accuracy for lung cancer detection. A Kaggle dataset of 6691 images of CT scans is used in this research. In preprocessing, these images were split into individual file and then resized in 64x64 resolution for data consistency along with antialiasing to smooth the edges of the nodules for better detection. In this study five classification models were applied – Convolution Neural Network (CNN), Random Forest classification algorithm (RF), Support Vector Machine (SVM) and boosting techniques such as Xtreme Gradient Boost (XGBoost) and Adaptive Boost (ADABoost) to classify the malignant and benign lung nodules and finally all the obtained results were compared. This study finds the accuracy of all the applied classifier models. From all these algorithms CNN outperforms with an accuracy of 89%. Support Vector Machine classifier accuracy was observed to be improved with RBF kernel. Random forest accuracy was observed to be 83%. This approach is successful for identifying the malignant and benign lung nodules with a greater number of images than previous studies.

1 Introduction

1.1 Background

Cancer is the most treacherous disease for human beings. Lung cancer is responsible for more deaths than combined death count of colon, prostate, ovarian and breast cancer (Narmada et al.; 2019). Lung cancer is a serious health concern for humans and alone in the United States of America with a count of 225,000 people each year (Rossetto and Zhou; 2017). The main factor causing lung cancer is smoking and the duration of smoking is directly proportional to the person getting affected with cancer. To detect lung cancer manually is a very tedious and risky job even for specialists. To gain deeper insights and identification of lung cancer in early stages, different machine leaning methods are used in image classification. By applying deep learning techniques such as CNN and other classification algorithms, an automated system can be built which can perform with higher accuracy rate and helps in accurate classification.

1.2 Motivation

Figure 1 represents the deaths that occurred in 2018 with the types of cancer in the worldwide population. This death count is increasing even if the patient is treated with proper healthcare. According to (Paul et al.; 2018) the lung cancer has only 18% , 5-year survival rate which is leading cause of death around the world and number two in United States of America (USA). (Antonelli and Yang; 2007) states that, if the lung nodules malignancy to be detected in earlier stage then the survival rate of the cancer patient could be increased drastically. Cancer can be properly handled only in primary stages and the diagnosis of cancer in primary stage is difficult. In this case the machine learning techniques such as convolution neural network and classification algorithms can be used to classify the benign and malignant lung nodules. Deaths occurred by type of cancer is explained in the figure 1.

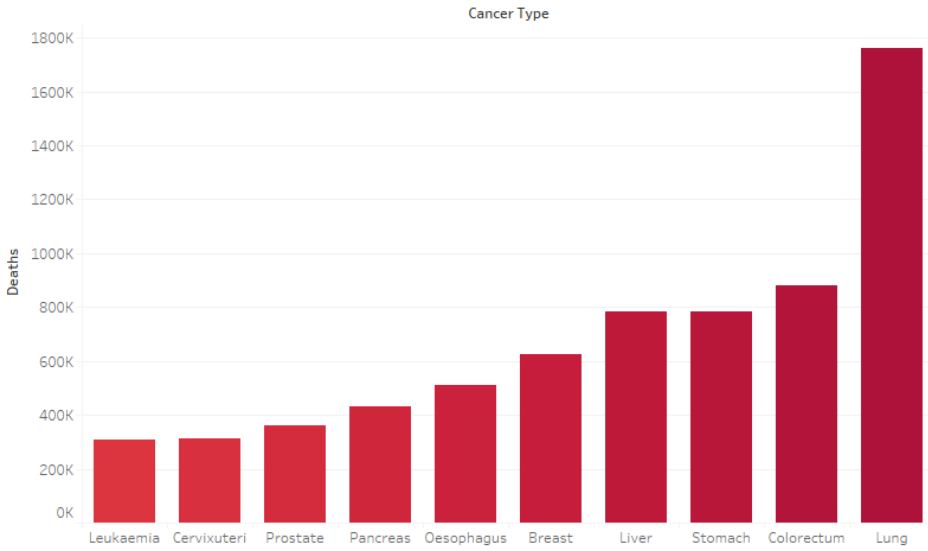

Figure 1: Statistics for deaths in cancer 2018 (Source: Statista)

1.3 Research Objective

The focus of this research is to build an efficient classification model with minimum configuration which gives a higher accuracy output. Thus, the deep learning model such as CCN is considered prime in this research and the output of other classification models is compared with the same. The aim of the research is as follows:

- Preprocessing the image dataset: The present dataset is stacked in on .tiff file which is split into individual images. After the split, these images are converted into same dimension for data consistency for CNN and another classification algorithm.
- Applying classification algorithms: Convolution Neural Network (CNN), Random Forest classification algorithm (RF), Support Vector Machine (SVM) and boosting techniques such as Xtreme Gradient Boost (XGBoost) and Adaptive Boost (ADA-Boost)

- Comparison and evaluation of the applied algorithm with the best performing model.

1.4 Research Question

This research addresses the problem faced in achieving higher accuracy for classifying the lung nodules from malignant to benign and which will help professionals for better decision making.

Research Question: *How well the lung nodules can be classified from malignant to benign using deep learning techniques and classification algorithms with limited computation power for achieving higher accuracy?*

The rest of the structure of the paper is organized as follows: Section 2 consists of the critical review of the related work done in this field. Section 3 explains Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology used and how the model is used to answer the research question. Section 4 explains the design overview of the architecture applied for this research. Section 5 explains the overall implementation of the models for the research. Section 6 results of the experiment carried out in this research. Finally, the conclusion of this research is discussed in Section 7. Section 8 will consist of acknowledgment followed by references.

2 Related Work

Identification of lung nodules and classifying them from benign to malignant is a very crucial job and time-consuming even for skilled professionals. As a product of this, much research is done in this field for developing an automated system that will take image as an input and classify the lung nodules which would be highly reliable and time-efficient which can achieve higher accuracy. Due to the recent applications and advancement in this area, multiple approaches are considered in this study to tackle this problem providing some quality results.

2.1 Convolution Neural Network (CNN)

According to(Rao et al.; 2016) CNN has special properties such as spatial invariance and ability to extract multiple features from the image. Therefore, CNN is suitable for CT scan images. This study has a limitation for dataset containing 71 patients record achieving the accuracy of 76% whereas in the study of (Rossetto and Zhou; 2017) the dataset consists of 150,000 individual slices of CT scans for 1500 patients and the archived accuracy was 97% this implies that deep learning model learns better with more data and gives better accuracy. In the study of (Cai et al.; 2019), proposed architecture of CNNH with a hashing function which consists of identical branches of share weights and parameters with 3 convolutions. Comparing this to the study of (Narmada et al.; 2019) the same architecture is applied but instead of hashing layer softmax layer is used. Comparing the outputs of these two models it can be stated that with softmax layer accuracy of the model seems to improve up to 95%. Ponnada and Srinivasu (2019) proposes modified CNN architecture which is also known as EFFI-CNN which consists of unique combination of 7 layers of convolution, max pool, fully connected layer, and softmax layer. This

study was done with the comparison of the models like ICDSSPLD-CNN and EASPLD-CNN and the EFFI-CNN model achieved best results comparatively.

In the research of (Sharma et al.; 2019) the proposed model has a sequential approach where the preprocessing of the image is followed by segmentation, feature extraction to classification. In this study multiple filters such as Gaussian filter, Gabor filter, and median filter were used to enhance the image to achieve better accuracy. For feature extraction multiple factors were also considered such as Gray Level Co-occurrence Matrix (GLCM), energy, entropy, contrast, mean, absolute mean and standard deviation. This method leads to achieving an accuracy of 98.08% using MATLAB platform. Whereas (Paul et al.; 2018) use an approach of extraction with the help of ensemble classifier which enhances the accuracy. This research is done on comparatively small-scaled dataset and to overcome this transfer learning approach is suggested. After performing the tests, the outcome of the built model was 76.79% which is a quality result. Whereas a different approach suggested by (Matsubara and Nacher; 2018) where spectral CNN was developed which is based on protein interaction network data and gene expression profile to classify lung cancer achieves an accuracy of 81%.

2.2 Support Vector Machine (SVM)

In the research of (Mousa and Khan; 2002) it was stated that radiologists fail to diagnose small lung nodules as much as 30%. This study follows an approach of image enhancement, feature extraction, and classification. Using SVM, polynomial of degree 10, 75% accuracy was achieved. The research of (Guo et al.; 2015) proposed a novel approach of intra-tumor homogeneity which has promising results for staging of lung cancer. Research done by (Mousa and Khan; 2002) and (Guo et al.; 2015) has done a process of feature extraction but they both considered different features while implementing the proposed models. Limitations of (Guo et al.; 2015) study were having a limited data set so that limited number of features were considered. In the study of (Rahane; 2018) it is stated that SVM model is very time efficient as compared to other classification models such as decision tree, K-nearest neighbor, LASSO regression, and Artificial Neural Network (ANN). For pre-processing, median filter and segmentation of the images provides an accurate result according to(Rahane; 2018). This study also approaches a unique combination of technologies such as Amazon Web Services (AWS), MySQL and JAVA but fails to specify the performance of the built model.

According to (Lobo; 2018) SVM is a supervised learning model and the aim of the implemented model was not only to classify the severity of the lung nodules but also to valid and easier analysis. Therefore, the large dataset is extracted for features having distinctive property which helps to differentiate categories into pattern. For feature extraction, contrast, correlation, energy, sum variance, maximum probability, dissimilarity factors were considered. The applied model gave an accuracy of 79.16%. On the other research of (Günaydin; 2019) are based on classifying lung cancer without performing any preprocessing on the images. In this research multiple classification algorithms were considered and finally the outputs were compared with before and after applying Principal Component Analysis (PCA) with considered models. Results for SVM before and after PCA were 50% and 55.4% respectively which are quite low hence it can be stated that for better accuracy preprocessing of the data is much needed.

In the research of (Senthil and Ayshwarya; 2018) the standard approach of image preprocessing, segmentation and feature extraction were followed before the SVM classifier but there the feature selection process was done by firefly algorithm which is a metaheuristic algorithm. The resultant output for the applied model was found out to be 92% accurate for classify lung nodules in binary sense of normal and abnormal.

2.3 RANDOM FOREST (RF)

The research done by (Kouzani et al.; 2008) mentions the classification of lung nodules using ensembles classification. This research was based on imaging of 32 patients with a large dataset on which 3 experiments were carried out with different tests and train size with comparative study of classification models like random forest, decision tree and support machine vector and finally the comparisons were drawn out. This study also concludes that when the dataset is divided into 80% of the dataset was used for training and 20% of the set was used to test the random forest classifier performed to be the best. According to this study it is also mentioned that for random forest classifier the classification errors appeared to be much higher but the execution time for random forest was much lower than SVM and decision tree. In the research of (El-askary and Salem; 2019) random forest classifier is used to detect lung anomaly in which the pixel is either with nodules or non-nodules. This study considers the dataset having multiple imaging source such as CT, Computer Radiography (CR) and Digital Radiography (DX) of 1010 patients. For the preprocessing CT scan images are converted into double then convoluted with its binary mask for removing the lung boundaries. For solving the unbalanced data problem (El-askary and Salem; 2019) proposes under-sampling of the majority class and removing zero intensity instance as well as undefined feature values. As the training and testing of data were done on increment of 20% at each phase till 100% the accuracy seemed to drop from 90.23% to 69.13%. in conclusion this research specifies that accuracy could be increased with the increasing number of trees. The study of (Kouzani et al.; 2008) and (El-askary and Salem; 2019) have similar outputs of accuracy in the primary stages.

Research done by (Hu and Nie; 2017) studied a dataset of 603 pulmonary lung nodules in which 288 were benign and 315 were malignant nodules. For pulmonary lung nodules classification, multiple latent types have to be considered like solid, part solid and soft nodules which makes a difficult job for classifier. In this study the followed approach of Improved RF classifier explains the efficient working of random forest so well due to the reason of ensemble classifier is able to composite weights from base classifier and randomize stages decrease the correlation between base learner classifier. To improve classification performance of RF (Hu and Nie; 2017) considered two ways, one is class decomposition in which homogeneous and heterogeneous clustering can be done and the other way is to tune the number of decision trees. After implementing the model, results were compared by the other classification models and the improved RF outperforms SVM, RF and Linear Discriminant Analysis (LDA) with an accuracy of 92.37%.

Whereas research of (Baboo and Iyyapparaj; 2018) explains the importance of contextual clustering for segmentation phase by which GLCM and LBP feature were extracted. Later these extracted features were used by RF, SVM, and K-Nearest Neighbor (KNN) classifiers. After implementing the model RF performs exceptionally well with accuracy

of 98%. The study of (Paing; 2018) proposed a system where sampling and three feature selection criteria were applied which are Relief, Genetic Algorithm and Particle Swarm Optimization which improves the performance of classifier. Research of (Hu and Nie; 2017) and (Paing; 2018) has a similar approach of Improved as (Hu and Nie; 2017) but with a slight modification of sampling and major voting. Comparing the performance of original RF with feature selection Improved RF it is observed that performance in terms of accuracy is highest in Genetic Algorithm's feature selection RF with 89.9% accuracy. Another research of (Roy and Banik; 2019) proposes a model where it can be able to identify cancerous portion of the lungs. In which MATLAB platform for implementation was used where SVM and RF classifier's performance was compared. For segmentation of the image watershed algorithm was used and for feature extraction Speed Up Robust Features (SURF) algorithm was used. The final result of the implementation concludes that due to SVM having ability of grouping mechanism, it can categorize negative and positive specimens of cancer. Hence SVM outperforming RF with accuracy of 94.5%.

2.4 ADAPTIVE BOOST (ADABoost)

According to (Safiyari; 2017) states that ADABoost is one of the most popular ensemble methods. In this research comparative study was done with other ensembles such as ADABoost, Bagging, Dagging, MultiBoosting, Random Subspace. For balancing the unbalanced data, under-sampling operation was performed and for validation and test and train split of the data, 10-fold cross-validation was done. Finally, in the results it was observed that ADABoost performed comparatively better than other ensembles considered in this study. Research done by (Zinovev et al.; 2011) states that ensemble-based machine learning is aimed to improve the classification performance of the classifiers. Whereas ADABoost techniques appear to be majorly used in medical field. In this proposed model the features were extracted for the particular nodule and interpreted so that it can be further used in classification process. ADABoost is known for creating a combination of weak learners which together can act as a strong learner. Also, in this research the ADABoost is considered for creation of probabilistic classifier using belief decision tree as a base classifier.

In the research (Zhu et al.; 2008) ADABoost is used to select bag of features and was used to build a two-level classifier to solve multiclass classification problems. The main approach behind ADABoost is that a strong classifier can be created by combining many weak classifiers as seen in the research (Zinovev et al.; 2011). Also (Zhu et al.; 2008) states that ADABoost focuses on samples with more weights which can be seen as harder for the weak classifier. To create a distance between positive and negative data ADABoost uses one bag feature at each iteration and will sum up the sample weights for negative and positive samples. After implementing the model MCMI-ADABoost observed to be highest accuracy of 89.3% among SVM, J48 and K-NN algorithm. In the research of (Liu et al.; 2015) ADABoost was used to classify the region of interest in lung CT images of CISL categories and to evaluate the built model fivefold cross-validation method was used. ADABoost was used in the research of (Tanaka et al.; 2014) for automated identification of lung nodules. This model uses ADABoost for detection of false positives using rule-based method whereas another application of ADABoost is done in the research of (Li; 2018) to generate a white nodule-likeness map. As a comparison drawn out with strong learners like SVM it was observed that ADABoost requires much less

execution time to achieve same level of accuracy. As the research also concludes that modest ADABoost also outperforms the real and gentle ADABoost.

2.5 EXTREME GRADIENT BOOST (XGBoost)

In the research of (Turki; 2018) the XGBoost algorithm was used to evaluate the performance of the predictions generated. As XGBoost is a scalable end to end tree system it is used to evaluate the area under the curve and accuracy on the real clinical data related to thyroid cancer, colon cancer, and liver cancer. For evaluation of these algorithms the dataset was partitioned into five-fold for cross-validation where the dataset is partitioned randomly and the performance of these algorithms is reported. The average mean area under curve (AUC) is used to score the good performance of these algorithms. After the implementation of the experiment, XGBOOST yielded highest AMAUC score of 81.10% for thyroid nodules for colon cancer it scored second highest with the accuracy of 87.20% and for liver cancer it scored 79.70%. The other study of (Jia et al.; 2018) for XGBoost is used to classify the pulmonary lung nodules on CT scan images as it is more optimized version of gradient boost algorithm. The proposed models implement segmentation by K-means then applying median filtering. By the AUC curve of the proposed model scored 93% for ensemble model of RF and XGBoost. The research of (Fu et al.; 2019) proposes MP4Ei model which is based on statistical theory and gradient boosting decision tree. 23 selected features are considered as input variables for XGBoost with Bayesian parameter tuning and cross-validation for finding out the most simplified model. After implementation of the MP4Ei model given an accuracy of 84.51%.

To detect cervical cancer in an early stage (Deng et al.; 2018) XGBoost, SVM and RF were used to classify the diagnostic result of four target variables which are Hinselmann, Schiller, Cytology and Biopsy. One of the advantages of applying XGBoost is that it adds the regular term of cost function which is able to control the complexity of the model which is used to avoid overfitting. The final result for accuracy of XGBoost for Hinselmann 96.34%, Schiller 95.59%, Cytology 96.30%, and Biopsy 95.59%. After the completion of result it was stated that XGBoost performed better to classify malignancy of the nodules better than the SVM classifier. The research of (Pham and Tan; 2019) proposes an automated model that minimizes the levels of uncertainty and subjectivity of human assessment in melanoma in which XGBoost is used as a classifier for image classification. In this implementation 2 experiments were carried out such as setting the values of hyperparameter to default and other approach was to several hyperparameters. In the results of this experiment it was found that 44 out of 100 images were correctly identified with melanoma as 44.44% of recall score and 69.36% of specificity.

2.6 Conclusion

Considering all the related work above mentioned it can be observed that building an efficient and effective classification system different machine learning algorithm have been studied to achieve optimum results and high accuracy rate. Classification models have different approach based on the relevant and the size of the data. It can be seen that deep learning model such as CNN needs to have data in large volume to perform well whereas models like RF and SVM tend to perform better if the feature extraction is done well. Similarly, as per the boosting techniques such as ADABoost and XGBoost are mostly

applied with an ensemble approach. As per the classification approach, this research aims to compare the performance in sense of high accuracy and preprocessing for data consistency and image enhancement but without performing any feature extraction.

3 Methodology

This research is based on building a system that can efficiently classify the malignant and benign lung nodules for that designing a flow that works efficiently is very important. After assessing the requirements of the research CRISP-DM is considered as best approach for this research due to being reliable, less costly and efficient. The most important benefit of this technique is we can modify it according to the required research. Considering the requirement of the research some modifications are made in this technique as shown in figure 2. The detailed outline of each step is explained further.

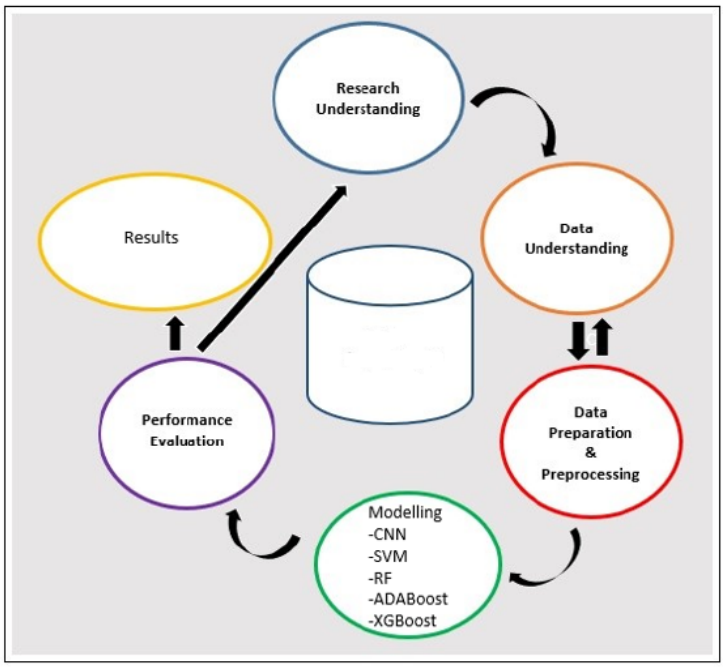

Figure 2: CRISP-DM

3.1 Research Understanding

The first phase of CRISP-DM is understanding the research objectives. The gathered knowledge is converted to a machine learning problem. In this phase, overall planning of the research is considered and aim of the research is well discussed. As the main objective of this research is to build a classification system in which minimum resources are used to achieve highest accuracy. The built system will be able to classify the benign and malignant lung nodules which can help to diagnose lung cancer resulting in the possibility of survival chances of the lung cancer patient up to 5 years.

3.2 Data Understanding

Collecting the data is the first task of performing research. For this, data availability becomes an important factor. As data collected for this research is an image dataset from Kaggle.com with labels for malignant and benign. This dataset is completely available on the public domain named Kaggle and was created by Kevin Mader. This dataset has a total of 6,691 images in which 4,165 images are labeled as benign and 2,562 images are labeled as malignant. This dataset was extracted from The Cancer Imaging Archive (TCIA) and was converted into a multipage Tagged Image File Format (TIFF) format. This image can be viewed with a specific software specified by the author which are ImageJ or KNIME.

3.3 Data Pre-Processing

As the dataset was present in a multipage tiff it was split into an individual image and converted to a Joint Photographic Experts Group (JPEG) format. This was done using TIFF Splitter tool. As all the images should be in same dimensions the reshaping of the images was done to the dimension of 64x64 by using Python Jupyter and further anti-aliasing filter was applied. Figure 3 explains the steps taken in data preprocessing.

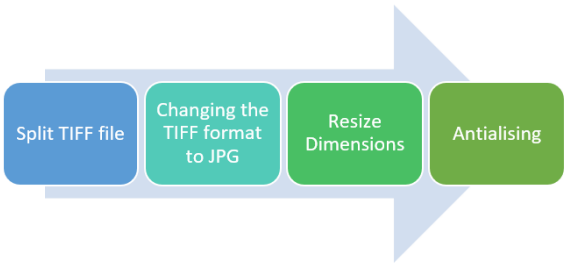

Figure 3: Image Processing

3.4 Modeling

The processed data can be used as an input to the model which is ready to use, five classification modeling techniques have been used for further implementation. As this research follows a straight forward approach of classification without any feature extraction it was possible to build a time-efficient system that can classify lung nodules based on the respective severity with minimum configuration and which can produce high accuracy and quality results. For the classification of lung nodules following models are used: CNN, SVM, RF, ADABoost and XGBoost. As discussed by Rao et al. (2016),Mousa and Khan (2002), Hu and Nie (2017),Zhu et al. (2008) and Fu et al. (2019) are simple for the application and provide high accuracy

3.4.1 Convolution Neural Network (CNN)

One of the most efficient deep leaning methods is Convolution Neural Network which is commonly used for imaging data. As deep learning method it is very efficient and

specifically used for imaging data. CNN can also be used for classification of imaging which it can predict the binary results which is essential for this research in which 0 signifies benign nodules and 1 signifies malignant nodules.

3.4.2 Support Vector Machine (SVM)

(Rahane; 2018) explains that SVM is a margin classifier that separates the two groups by a hyperplane which makes the SVM identity as non-probabilistic binary classifier. The training data points which is closest to the nearest classifier is also known as Support Vector. The distance between the malignant nodules of cancer and the hyperplane could be far as possible due to which it can be used to classify the nature of the lung nodule.

3.4.3 Random Forest(RF)

According to (Kouzani et al.; 2008) Ensemble learning is referred to an algorithm which generates a collection of classifier which learn the classification by training individual learner and fusing their respective predictions. In an ensemble learning method, which can generate many classification trees in forest, each tree generates a classification. The forest selects the classification that has most of the votes. Random forest works well due to the randomized stages decreasing correlation between distinctive learners in the ensemble.

3.4.4 Adaptive Boost (ADABoost)

As mentioned in the research of (Zhu et al.; 2008) ADABoost is a strong classifier and could be created by combining the weak classifiers. The main function carried out by ADABoost is that it focuses on samples that have more weights which is harder for weak classifiers. Hence, for this research any feature extraction is not included so for the classification ADABoost can be used as a strong classifier.

3.4.5 Extreme Gradient Boost (XGBoost):

XGBoost is an advance version of gradient boosted decision tree which is known for speed and performance. In comparison to another gradient boosting the algorithm, XGBoost is considered a benchmark. In gradient boosting new models are created that can predict the residual error of prior models then all added together to the final prediction. This approach can be used for both regression and classification.

3.5 Performance Evaluation

The primary metrics for performance evaluation of this research are accuracy, precision, recall and F1 score and total time taken by then model for complete execution. These evaluations are done on five different classification models such as CNN, SVM, RF, XG-Boost, ADABoost.

3.6 Deployment

Deployment is the final stage in CRISP-DM methodology which consisting of deployment of the built project in business. This stage also includes the maintaining and monitoring of the developed application if it becomes part of day-to-day business.

4 Design Specification

To develop an efficient classification system which help in classifying the lung nodules based on their respective severity an architecture design is developed. This architecture represents the techniques and tools that were used to build this system. The designed architecture follows a three-tier design architecture where it consists of interface, business logic layer and data persistent layer. In which data persistence layer describes the source of the data and business logic layer explains preprocessing of the data, training and evaluation of the classification models. The detailed architecture used to classify lung nodules is explained in figure 4.

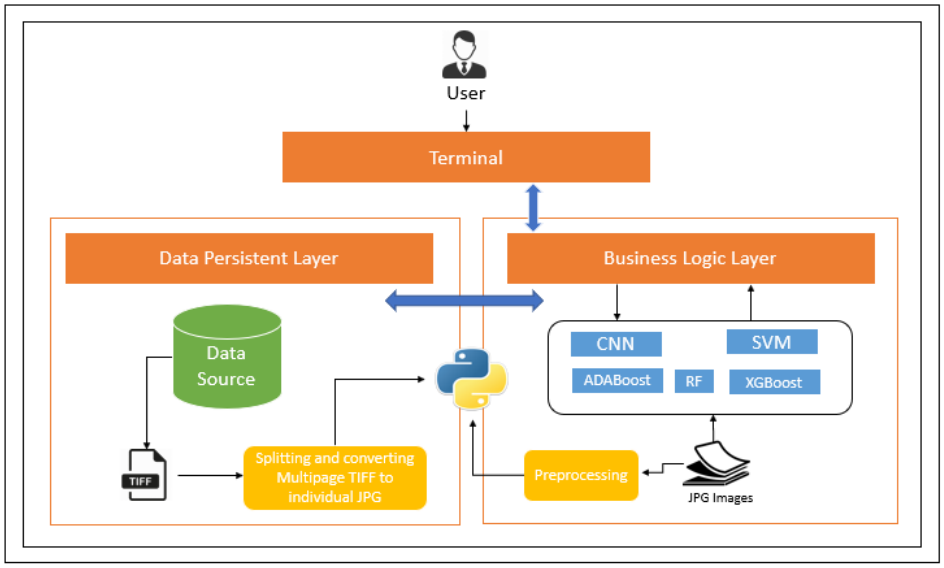

Figure 4: 3 Tier Architecture for Lung Nodule Classification

5 Implementation

In this section overall implementation of the system is discussed. This section also explains all the tasks carried out in order to successfully classify the lung nodules. First phase of this stage is environment setup which given an information about the tools used for this research and required environment. Further data cleaning, pre-processing and transformation is done and then detailed information about the applied model architecture and working is explained.

5.1 Environment Setup

Implementation of this research was performed on a 64-bit Windows 10 operating machine with RAM of 8GB and as a programming language Python was used. As IDE Jupyter Notebook was used to build this classification system which is provided by Anaconda framework. Overall implementation of this research was done on of Python 3.7.5 and Jupyter Notebook 6.0.2.

5.2 Data Preprocessing and transformation

The present data is in multipage TIFF format so then splitting of these images was done using image splitter and conversion of format from TIFF to JPG was done as well. As the TIFF file does not supports compression so it would increase the total time taken to run the model and as JPG image format supports compression. Further the present images were resized in 64x64 dimension for data consistency especially as a requirement for neural network and then anti-aliasing was turned on for more understanding of border detection of lung nodules. For CNN to make sure all the images are in grayscale from cv2 library of python cv2.IMREAD_GRAYSCALE function was used as well as cv2.resize function was used to resize the image.

5.3 Classification Algorithms

5.3.1 Convolution Neural Network (CNN)

CNN consists of many layers such as input layer, output layer, and multiple hidden layers. These hidden layers may consist of a sequence of multiple convolution layers. The detailed architecture of applied CNN model is illustrated in figure 5. Further there can be multiple layers of activation layer, pooling layer, fully connected layer, and normalization layer. The proposed CNN model is a sequential model which is a simplified model in Keras which allows constructing a model layer by layer. Each layer consists of weights that can correspond to the layer which is followed by the previous layer. For further understanding of layer used in building of proposed CNN is explained below mentioned by (Rao, Pereira and Srinivasan, 2016) and (Ponnada and Srinivasu, 2019).

Convolution Layer: The purpose of the convolution layer is to abstract certain features from the input image. Convolution deals with spatial relationships between pixels and it can be done by learning image features using small squares of input image. For the implementation of this CNN 4 convolutions were applied.
Max Pooling Layer: Max Pool Layer is another type of spatial pooling. By the use of this layer dimensionality of the feature, a map can be reduced and critical information can be preserved. For this purpose, with the convolution phase 4 max pool layer were added.
Flatten Layer: Flatten layer flattens the output of the convolution layer which is responsible for the decision making of classification. Flattening is considered a key step in all Convolution neural networks. For this neural network one layer of flatten() was used.
Dense Layer: The dense layer is generally put before the classification layer. It sums up the information which could be extracted from the previous layer of the feature maps. One dense layer where argument of (64) was given and Dense(1) was passed before compiling the model.
ReLU Activation Function: ReLU is most commonly known as Rectified Linear Unit. ReLU is linear for all positive values and zeroes for all negative values which imply that the computation cost can be considered as zero.
Sigmoid Activation Function: Sigmoid function is used in the model due to the existence of it between the prediction of probability. Since the probability of lung cancer exists between malignant and benign which is 0 and 1. Therefore, the use of sigmoid is advised and Adam optimizer was used within to measure accuracy of the model Accuracy metrics was used.

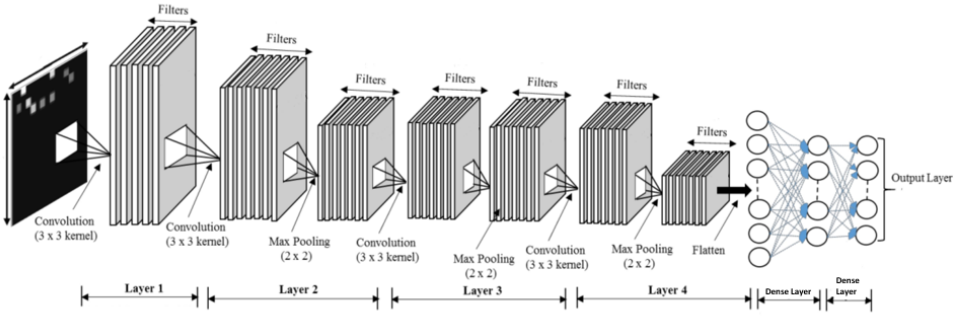

Figure 5: Convolution CNN Architecture

In this CNN architecture, after testing the multiple iterations of convolutions from 2, 4, 6, 8 the 4 convolution layers performed best and used from 16x16 up to 128x128 convolutions and 1 flatten and 2 dense layers. In this CNN model sequential type of model is used due which gives an ability to build the model layer by layer. As well as in final phase sigmoid activation function is used because it lies between 0 and 1 therefore it can be beneficial for the model to predict binary output.

5.3.2 Support Machine Vector (SVM)

As the classification of the lung nodules is a binary classification which is benign or malignant SVM classifier uses a hyperplane which was used to separate these values. According to (Rahane, 2018) the malignant nodules would be as far as possible from this hyperplane. In implementation of this classifier Linear and RBF kernel were tested for better accuracy and with setting the gamma on auto RBF kernel provides better accuracy. To apply SVM classification model svc.SVM function was used from sklearn framework. For making the classification predictions from the classifier the svc.predict() function was used. After testing this model with 10-fold cross validation no significant change in the accuracy was observed.

5.3.3 Random Forest (RF)

Random forest selects random samples from the dataset and giving a tree size of default measure of estimators to 100. By doing so construction of decision tree is done for each sample and the classification is done from each decision tree. Once the classification is taken place voting is done at each classification. After following this process classification results are selected with the most votes. The train and test split of data was tested best for optimum accuracy was 70-30 ratio.

5.3.4 Adaptive Boost (ADABoost)

ADABoost assigns weights of training data sample in each iteration as it can ensure the accurate classification of each unusual observation. Here the unusual observation would be the malignant lung nodules. Here estimators are weak learners which are trained iteratively are set on the value of 50 and the learning rate is set to default to 1.0. Here

the train and test size for the classifier was taken as 90 – 10 and the random state was set to be on 109

5.3.5 Extreme Gradient Boost (XGBoost)

To apply XGBoost classification model XGBClassifier was used from scikit-learn framework framerwork. For classification obtaining predictions from the classifier the model. predict() function was used. As XGBoost is compatible to scale any form of data due to the complex structure of the algorithm which is self-learning over time. In this scenario XGBoost collate all the present data and compares all the information accurately for optimum classification result. Here the train and test size for the classifier was taken as 90 – 10 and the random state was set to be on 109. XGBoost classifier is a type of regularized model which helps to improve in a significant amount of error level which prevents overfitting.

6 Evaluation and Results

In evaluation section a detailed and comprehensive analysis of the evaluations and results which are achieved is done for this research. The models and architecture which were used for implementation for this research are selected from the models and parameters which performed best. For this study five models were selected for classifying the lung nodules are CNN, SVM, RF, ADABoost, XGBoost.

6.1 Evaluations

For evaluation of the implemented model following metrics are used and briefly explained:

Accuracy: Accuracy can be calculated by the number of correct predictions over the total number of predictions.

Precision: Precision is calculated as the number of correct predictions over the number of correct predictions and the number of predictions to which the model assumed to be correct but it was not. We can also say that it is a result of number of true positives divided by number of true positives and a number of false positive.

Recall: Recall can be calculated by the number of correct predictions over a number of correct predictions and prediction to which model assumed to be something else but it was not or it can be calculated as the number of true positives divided by the number of true positives and false negatives.

F1 Score: F1 score can be calculated by taking the mean of precision and recall.

Total Computation Time: It is the total time taken by the algorithm for its complete execution.

6.2 Results

The obtained results are explained in the form of experiments so that the best classifying model can be detected for lung cancer classification. These experiments are conducted on criteria like time efficiency, accuracy and comparison on measures of evaluation metrics.

6.2.1 Experiment 1: CNN with 4 Layer model for classification of lung nodules.

This model was built from scratch with four convolution layers and the respective filter size of 16, 32, 64, 128. Total 10 epoch were considered to be sufficient to gain a stable output. With this approach the model was successfully executed and gave a F1 Score of 83.61%. Obtained validation accuracy of 81.32%. Total time taken for 10 epochs was 3195 seconds which indicates that training a convolution neural network. Figure 6 explains the train – validation accuracy and loss for 10 epochs.

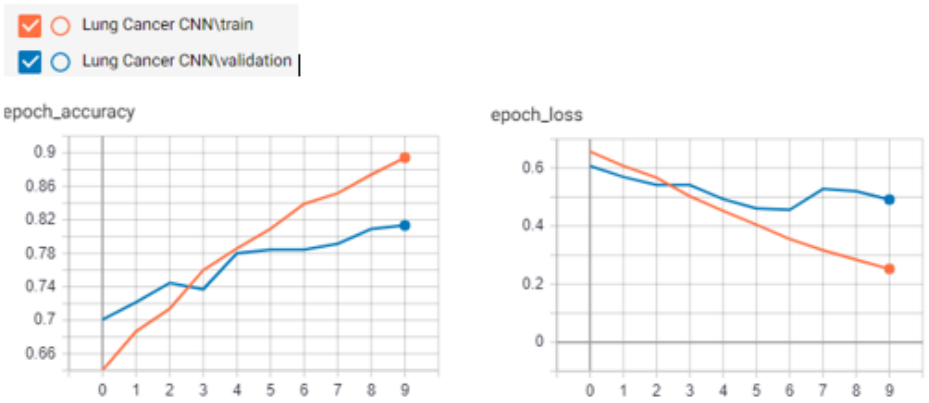

Figure 6: Tensorboard Graph for train- validation for accuracy and loss for 10 epochs

6.2.2 Experiment 2: Calculating the total time taken for each model for successful execution.

This experiment was carried out on each model and was calculated by time library present in python. Figure 7 explains the code used to calculated the time in terms of seconds.

```
In [ ]: import time
        start = time.process_time()
        # classification model code
        print(time.process_time() - start)
```

Figure 7: Python code for total time taken by each classification model.

In figure 9, it is observed that the CNN model has the most computational time taken

by 3195 seconds. And the lowest of the computational speed was observe by Random forest which was 65.56 seconds. For other classification models such as SVM, ADABoost and XGBoost the computational speed in terms of time was 568.39, 238.15 and 246.56 seconds.

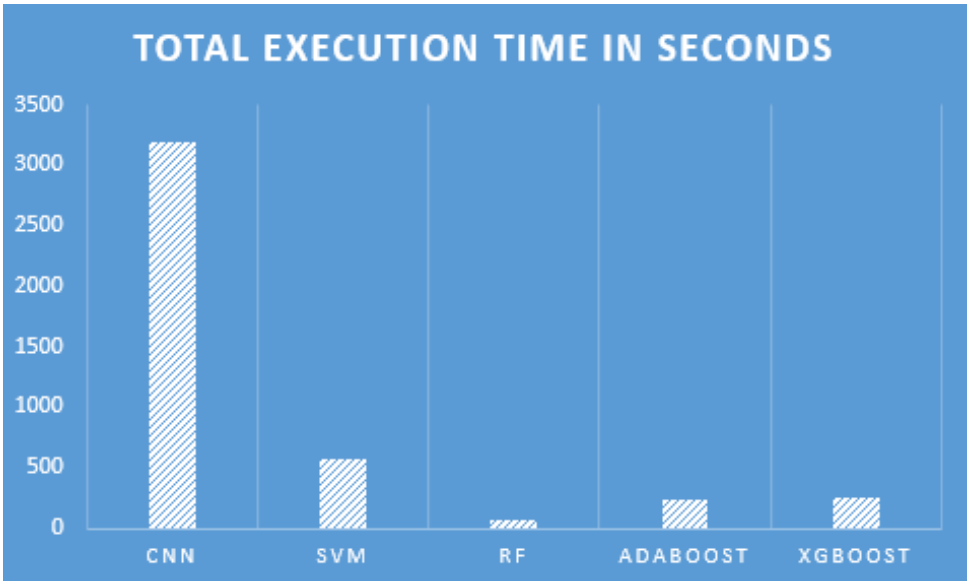

Figure 8: Computational time taken by each classification model

6.2.3 Experiment 3: Comparative study of classification models in terms of accuracy.

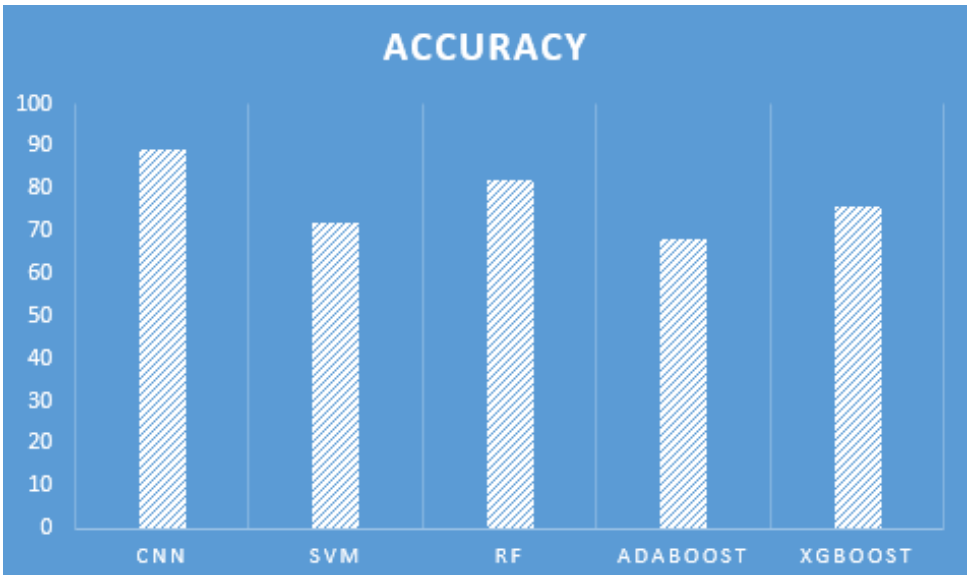

Figure 9: Accuracy of classification models.

It is observed that in figure 9, convolution neural network has achieved the highest accuracy as 89.43% as compared to other implemented models and the lowest score was

achieved by ADABoost as 68%. Other classification models such as Support Machine Vector, Random Forest and Extreme Gradient Boost performed well and gave quality results. CNN being a deep learning model and which is specially used for imaging data it can perform better than other models. For detection on lung cancer and to find malignant nodules, CNN model gives reliable results.

6.2.4 Experiment 4: Comparative study of classification models in terms of Precision, Recall and F1-score.

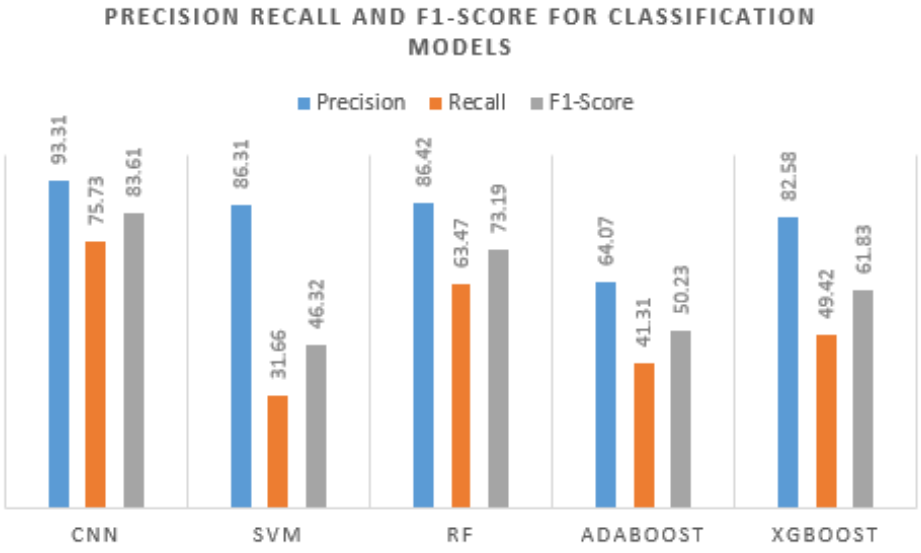

Figure 10: Classification models Evaluation Score

In this experiment it is observed that the precision score that states the positive identification actually were correct by the model. By comparing the precision score, it can be observed that CNN has the best precision score of 93.31% and ADABoost has the lowest score of 64.07%. Also, RF and SVM score are almost similar. Recall Score signifies the actual positives identified correctly. From figure 10 it can be observed that the SVM classification model appears to has lowest recall score of 31.66% whereas CNN and RF performed better than the other models. For the F1 Score, it signifies the balance between precision and recall. Observing the figure 10 it can be concluded that out of the 5 models CNN and RF are set to have the most precise balance of 75.73% and 63.47%.

7 Discussion

In this study, deep learning architecture like convolution neural network and other classification models like SVM, RF and boosting algorithms like ADABoost and XGBoost were used. These architectures were studied and based on the study; selection of classification models was done. As no feature extraction was done in this process the obtained results are preliminary product of classification achieving quality results. These models were trained on a total of 6691 images with which CNN scored the highest with highest computing cost. The precision and balance of the CNN model, so that the model can be

furthermore developed by people. Although the applied models successfully achieved the desired research goal but it can have more improvements. Testing of the model was done with the subset of same dataset which is widely used in machine learning community. Also, multiple iteration of testing was done to obtain more accurate results by splitting the data in various size of test and train with randomizing the data. The limitation in this research were attaining more set of CT images which could help to train the model more. The second limitation of this research was the field of study was limited to lung cancer, as the scope can be widened to other neglected cancers occurred in males in their later life such as testicular cancer. The field of study can be broadened to other cancer types and automation of the classification system

8 Conclusion and Future Work

To conclude this research the lung nodules were classified with high accuracy and with limited computation power. The preprocessing of the images was done efficiently which helped the model for less time consumption. In the end of the research comparative study was done to asses the quality of results. The CNN model obtaining the highest accuracy of 89.43% gives a quality result. As well as it was observed that Random Forest algorithm has the lowest computation cost achieving accuracy of 82%. This makes Random Forest an efficient model in terms of accuracy and computation cost. ADABoost boosting algorithm was found as a drawback with the approach of the research as the results obtained with this boosting algorithm were not promising.

In the future work the lung cancer detection can be done on other imaging format which can build a 4D image structure such as 4D MRI. This can used for accurate segmentation of the image and which can help to measure the size of lung nodules accurately. Second application of this research could be used for full scaled system for assistance to the radiologists and doctors for better decision making. Due to shortage of time and resources the extensive hyper parameter testing and tuning was not performed. In future work, more numbers of images and parameters should be taken into consideration which can benefit the classifiers. This study was based on complete supervised learning and for the future work unsupervised machine learning approach can benefit for identifying hidden patterns of lesions and nodules.

9 Acknowledgement

I would like to express my gratitude to my supervisor and guide Dr. Muhammad Iqbal for guiding me and kept me motivated throughout my research. It also helped me to gain a lot of knowledge while seeking his guidance for this research. If it wasn't for his constant feedback the results were achieved and I was able to present this research in meaningful and sophisticated manner would never been possible. As being a novice in this field, its his insights which allowed me to grasp the concepts of data analytics.
I would like to thank my sister and my parents for believing me and supporting me in this journey. I would like to thank my friends who always helped me in solving the difficulties I faced.

References

Antonelli, M. and Yang, G.-z. (2007). LUNG NODULE DETECTION USING EYE-TRACKING, pp. 457–460.

Baboo, S. S. and Iyyapparaj, E. (2018). A CLASSIFICATION AND ANALYSIS OF PULMONARY NODULES IN CT IMAGES USING, *2018 2nd International Conference on Inventive Systems and Control (ICISC)* (Icisc): 1226–1232.

Cai, Y., Li, Y., Qiu, C., Ma, J. I. E. and Gao, X. (2019). Medical Image Retrieval Based on Convolutional Neural Network and Supervised Hashing, *IEEE Access* **7**: 51877–51885.

Deng, X., Luo, Y. and Wang, C. (2018). Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods, *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* pp. 631–635.

El-askary, N. S. and Salem, M. A. (2019). Feature Extraction and Analysis for Lung Nodule Classification using Random Forest, pp. 248–252.

Fu, B., Liu, P., Lin, J., Deng, L., Hu, K. and Zheng, H. (2019). Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data, **66**(7): 2053–2064.

Günaydin, Ö. (2019). Comparison of Lung Cancer Detection Algorithms.

Guo, N., Yen, R.-f., Fakhri, G. E., Li, Q. and Hduo (2015). SVM Based Lung Cancer Diagnosis Using Multiple Image Features in PET/CT, **3**.

Hu, H. and Nie, S. (2017). Classification of Malignant-Benign Pulmonary Nodules in Lung CT Images using an Improved Random Forest (use style : paper title), *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* pp. 2285–2290.

Jia, F. A. J., Liu, B. J. and Gu, C. Y. (2018). Computer-aided diagnosis of pulmonary nodules on CT scan images, (Icmic): 2–4.

Kouzani, A. Z., Lee, S. L. A. and Hu, E. J. (2008). Lung Nodules Detection by Ensemble Classification, pp. 324–329.

Li, X. (2018). A Solitary Feature-Based Lung Nodule Detection Approach for Chest X-Ray Radiographs, *IEEE Journal of Biomedical and Health Informatics* **22**(2): 516–524.

Liu, X., Ma, L., Song, L., Zhao, Y., Zhao, X. and Zhou, C. (2015). Recognizing Common CT Imaging Signs of Lung Diseases Through a New Feature Selection Method Based on Fisher Criterion and Genetic Optimization, *IEEE Journal of Biomedical and Health Informatics* **19**(2): 635–647.

Lobo, P. (2018). Classification and Segmentation Techniques for Detection of Lung Cancer from CT Images, *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)* (Icirca): 1014–1019.

Matsubara, T. and Nacher, J. C. (2018). Convolutional Neural Network Approach to Lung Cancer Classification Integrating Protein Interaction Network and Gene Expression Profiles, *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)* pp. 151–154.

Mousa, W. A. H. and Khan, M. A. U. (2002). Lung nodule classification utilizing support vector machines, pp. 153–156.

Narmada, K., Prabakaran, G. and Mohan, S. (2019). Classification and Stage Prediction of Lung Cancer using Convolutional Neural Networks, (10): 993–998.

Paing, M. P. (2018). Improved Random Forest (RF) Classifier for Imbalanced Classification of Lung Nodules, *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)* (i): 1–4.

Paul, R., Hall, L., Goldgof, D., Schabath, M. and Gillies, R. (2018). Predicting Nodule Malignancy using a CNN Ensemble Approach, *2018 International Joint Conference on Neural Networks (IJCNN)* pp. 1–8.

Pham, H. N. and Tan, B. L. (2019). Lesion Segmentation and Automated Melanoma Detection using Deep Convolutional Neural Networks and XGBoost, *2019 International Conference on System Science and Engineering (ICSSE)* pp. 142–147.

Ponnada, V. T. and Srinivasu, S. V. N. (2019). Efficient CNN for Lung Cancer Detection, (2): 3499–3503.

Rahane, Wasudeo, H. D. Y. M. A. K. (2018). Lung Cancer Detection Using Image Processing and Machine Learning HealthCare, *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* pp. 1–5.

Rao, P., Pereira, N. A. and Srinivasan, R. (2016). Convolutional Neural Networks for Lung Cancer Screening in Computed Tomography (CT) Scans, *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* pp. 489–493.

Rossetto, A. M. and Zhou, W. (2017). Deep Learning for Categorization of Lung Cancer CT Images, pp. 10–11.

Roy, K. and Banik, R. (2019). A Comparative study of Lung Cancer detection using supervised neural network, *2019 International Conference on Opto-Electronics and Applied Optics (Optronix)* pp. 1–5.

Safiyari, A. (2017). Predicting Lung Cancer Survivability using Ensemble Learning Methods, (September): 684–688.

Senthil, S. and Ayshwarya, B. (2018). PREDICTING LUNG CANCER USING DATAMINING TECHNIQUES WITH THE, *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)* pp. 210–216.

Sharma, S., Kaur, M. and Saini, D. (2019). Lung Cancer Detection using Convolutional Neural Network, (6): 3256–3262.

Tanaka, S., Ikeda, Y., Kim, H. and Tan, J. K. (2014). Automatic Identification of Lung Candidate Nodules on Chest CT Images Based on Temporal Subtraction Images, *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)* pp. 1364–1368.

Turki, T. (2018). An Empirical Study of Machine Learning Algorithms for Cancer Identification, *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)* pp. 1–5.

Zhu, L., Zhao, B. and Gao, Y. (2008). Multi-Class Multi-Instance Learning for Lung Cancer Image Classification Based on Bag Feature Selection, pp. 487–492.

Zinovev, D., Furst, J. and Raicu, D. (2011). Building an Ensemble of Probabilistic Classifiers for Lung Nodule Interpretation, *2011 10th International Conference on Machine Learning and Applications and Workshops* **2**: 155–161.