Summer 8-11-2020

Deep Functional Mapping For Predicting Cancer Outcome

A.K.M. Kamrul Islam

DEEP FUNCTIONAL MAPPING FOR PREDICTING CANCER OUTCOME

by

A.K.M. KAMRUL ISLAM

Under the Direction of Saeid Belkasim, PhD

ABSTRACT

The effective understanding of the biological behavior and prognosis of cancer subtypes is becoming very important in-patient administration. Cancer is a diverse disorder in which a significant medical progression and diagnosis for each subtype can be observed and characterized. Computer-aided diagnosis for early detection and diagnosis of many kinds of diseases has evolved in the last decade. In this research, we address challenges associated with multi-organ disease diagnosis and recommend numerous models for enhanced analysis. We concentrate on evaluating the Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Positron Emission Tomography (PET) for brain, lung, and breast scans to detect, segment, and classify types of cancer from biomedical images. Moreover, histopathological, and genomic classification of cancer prognosis has been considered for multi-organ disease diagnosis and biomarker recommendation.

We considered multi-modal, multi-class classification during this study. We are proposing implementing deep learning technique based on Convolutional Neural Network and Generative Adversarial Network.

In our proposed research we plan to demonstrate ways to increase the performance of the disease diagnosis by focusing on a combined diagnosis of histology, image processing, and genomics. It has been observed that the combination of medical imaging and gene expression can effectively handle the cancer detection situation with higher diagnostic rate rather than considering the individual disease diagnosis. This research puts forward a blockchain-based system that facilitates interpretations and enhancements pertaining to automated biomedical systems. In this scheme, a secured sharing of the biomedical images and gene expression has been established. To maintain the secured sharing of the biomedical contents in a distributed system or among the hospitals, a blockchain based algorithm is considered that generate a secure sequence to identity a hash key. This adaptive feature enables the algorithm to use multiple data types and combines various biomedical images and text records. All data related to patients, including identity, pathological records are encrypted using a private key cryptography based on blockchain architecture to maintain data privacy and secured sharing of the biomedical contents.

INDEX WORDS: Medical Imaging, Machine Learning, Deep Learning, CT Scans, Histology Image, Microarray Gene Expression, Convolutional Neural Network, Generative Adversarial Networks, Computer-aided-diagnosis, Blockchain.

.

DEEP FUNCTIONAL MAPPING FOR PREDICTING CANCER OUTCOME

by

A.K.M. KAMRUL ISLAM

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2020

DEEP FUNCTIONAL MAPPING FOR PREDICTING CANCER OUTCOME

by

A.K.M. KAMRUL ISLAM

Committee Chair: Saeid Belkasim

Committee: Yanqing Zhang

Pavel Skums

Marina Arav

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

August 2020

DEDICATION

To my beloved brother Ifte Rubel who passed away during my PhD study and whom I like more than my life.

ACKNOWLEDGEMENTS

Writing acknowledgements to all who have contributed to this work is practically impossible. But I would like to explicitly thank a few of the people who helped me along the way.

First and foremost, my appreciation and heartfelt gratitude to my thesis supervisor and mentor, Dr. Saeid Belkasim, Georgia State University, for his constant endeavour, step by step guidance and the time he devoted for this research work.

I would also like to express my deepest gratitude to the Computer Science department for their funding of this project in Georgia State University, as without this it is unlikely that a successful progress of this project would have been achieved. In addition, I would like to thank Prof. Yanqing Zhang, Dr. Pavel Skums and Dr. Marina Arav for their continuous guidance and help throughout.

I am very thankful to the Dean, Faculty of Arts & Sciences, for the facilities and the work environment provided to me. These facilities enable me to enrich my knowledge and carry out my research. Many thanks go to all the staff at Georgia State University, my friends and colleagues who have helped me with this research project.

In addition, I am indebted to my dearest parents for their endless love.

A.K.M. Kamrul Islam

August 2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- **SVM ....** Support Vector Machine
- **SGD ....** Stochastic Gradient Descent
- **ANN ....** Artificial Neural Network
- **DNN ....** Deep Neural Network
- **MLP ....** Multi-Layer Perceptron
- **CNN ....** Convolutional Neural Network
- **Resnet ....** Residual Neural Network
- **FCN ....** Fully Convolutional Neural Network
- **FCRN ....** Fully Convolutional Regression Network
- **STN ....** Spatial Transformer Network
- **RNN ....** Recurrent Neural Network
- **LSTM ....** Long-Short Term Memory
- **GRU ....** Gated Recurrent Unit
- **CT ....** Computed Tomography
- **MRI ....** Magnetic Resonance Imaging

1 Introduction

In this chapter, we introduce complete outline to this thesis intended to permit a quick assessment of its focus, objectives, contributions, challenges, and structure.

1.1 Background and Motivations

Cancer is a heterogeneous cluster of disorders exemplified by unrestrained expansion of the cells. Cancers are mostly categorized by the nature of cells or tissue considering their origination. Since malignant expansion can arise in almost all locations of the body, around 100 distinct categories of cancers are obvious. Cancer is an enormously complicated and varied disease. Those characteristics, known as traits of cancer, are an integrated set of skills gained during tumor genesis.

In 2012, 14.1 million individuals identified as cancer and 8.2 million people died of that in the world. World Health Organization (WHO) estimated that, around 21.7 million people will be detected as cancer patient and 13 million individuals will die due to cancer by 2030 [1]. Recently, cancer has turn out to be the second prominent source of fatality in the earth. The developing countries are facing the large number expansion of cancer due to several factors like heavy smoking, poor diet, physical inactivity, and environmental pollutions. Now-a-days, commonly identified cancer occurred worldwide are ten different categories. These are lung (13%), breast (11.9%), colorectum (9.7%), prostate (7.9%), stomach (6.8%), liver (5.6%), cervix uteri (3.7%), and bladder (3.1%) [2].

Advances in cancer research have been obvious last a few years, but still remarkable gaps can be observed in this field due to the existence of several subtypes of cancer in the world. Understanding and knowledge gathering would be essential steps to identify and diagnosis of the cancer disease. Researchers now focusing on several distinct subtypes that exist in cancer with distinct roots, various threat aspects, unique genomic alterations, unusual biological activities, and dissimilar predictions, and scenarios need to be discovered. Early detection, therapy, and supervision of cancers can reduce the risk of the disease [3]. However, failure to detect the cancer effectively would occur high chances of mortality. Several key factors that distract the accurate identification of the disease is given below:

- Absence of robust research on precise disease subtypes.
- Lack of knowledge of genomic probability factors.
- Scarcity of effective and automated disease detection tools.
- Inadequate indication-based customized medicine tactics designed to the disease subtypes and tumor attributes.
- Lack of awareness given to research on supportive issues for long term care of patients to observe the disease progression

Machine Learning (ML) has yielded significant improvement in modern years. ML is overly used now-a-days in different sectors to achieve incredible performance with high accuracy rate in every steps of work. ML models can analyse different types of data like text, image, audio, video, etc. to extract information from the data and helps to predict the data sequence if new sets of data appear [4]. In this thesis, we aim to use machine learning and deep learning models for solving various problems related to biomedical imaging and genome sequencing in terms of giving a robust disease diagnostic method.

In recent years, computer-aided diagnosis (CAD) have been overly used for disease diagnosis as it is tedious task for human beings to identify disease instantly from the tons of biomedical images and gene expression sequence. Machine learning especially deep learning models can play a great role for disease diagnosis by accurately identifying the abnormal regions from the medical contents. CAD systems considers those deep learning models reduced the time and task of human being with better disease diagnosis. Moreover, vast amount of data can be handled smoothly and in a faster way comparing to human being.

1.2 Aims

The research work aims to develop more accurate and effective cancer detection model by comparing to current disease diagnosis model with the purpose of achieving better performance. Therefore, the research approach has followed by involving five main steps

1. Understanding the existing individual, hybrid disease detection model and their applications in different domain
2. To identify the key issues that can enhance the performance and accuracy of the model
3. Design and implementation of a new feature extraction technique and incorporate it into the model that can be evaluated through performance measurement methods
4. Investigating a set of deep learning model and consider how the models can be extended and which characteristics are necessary to achieve better accuracy
5. To understand the medical contents collected from internet of medical things (IoMT) devices and sharing technique in a distributed environment using blockchain technique. Performance analysis of the medical data access, sharing, modification in a centralized or distributed system.

This research aims to address the following questions:

1. Can deep learning techniques outperform conventional machine learning algorithms in detecting and classifying cancer from biomedical images as well as genome sequences?

2. How effective the feature extraction techniques in terms of attaining beneficial biomarkers from both image and genome perspective?
3. How can deep learning methods perform finest in terms of limited biomedical data? Which deep neural network architecture is appropriate for disease diagnosis?
4. How to supply security and police investigation for patient medical health records? How clinicians/physicians will access the stored data?

1.3 Objectives

The objective of this study is-

1. To develop a computer-aided diagnosis (CAD) system that can process the data effectively in terms of identifying abnormality in the medical contents.
2. To evaluate the performance of the proposed deep learning models with existing real-life datasets.
3. To develop an effective deep learning model for disease diagnosis by considering feature extraction technique, segmentation technique, data augmentation technique etc. and compare the performance to the state-of-the-art method.
4. To propose an ensemble classification model for cancer detection. Pre-trained transfer learning models can be considered if the dataset is of limited size.
5. To develop a model with two stage association study of the medical contents from biomedical images and genome expression for better identifying the disease.
6. To protect the personal details of a patient and to produce confidentiality to patient medical records victimization using “Blockchain Technology”.

1.4 Challenges

The central focus of this research is to classify and segment biomedical images for detecting cancer biomarker from different human organs like lung, breast, eye, etc. Moreover, the gene expression analysis using microarray gene expression data of different organs is considered for prediction of metastasis. The association study is also a key part for multiway analysis for detecting cancer outcome. The main challenges that we faced in this research are as follows: a) Biomedical dataset for detecting cancer biomarker is limited in size as the dataset is confidential and patient's personal information. Moreover, the publicly available dataset is limited in sizes and samples. We applied augmentation technique to enhance the dataset to fed it into the deep neural network. b) Due to limited size overfitting issue can come in biomedical image analysis. Transfer learning with fine tuning would be an answer but sometimes there has some performance issues that we cannot rely on. c) CNN model is overly used for image classification, but the performance is not satisfactory, therefore, ensemble model could be a good choice for cancer biomarker detection. d) In terms of microarray gene expression data analysis, there are huge number of features to deal with. Extracting and selecting significant feature would be a great challenge for analyzing the data for predicting metastasis. e) Research showed that association study of radiological data, histological data, and genome sequences provide significant performances for disease diagnosis. It is challenging to collect biomedical image, genome sequence, histology image dataset for the same sample. f) Secured sharing of image and genomic information in a distributed system is a great challenge when analyzing the confidential patient data. Blockchain technology with privacy preserving technique can be applied.

1.5 Contribution

The main impacts of this research are:

1. An investigation of existing techniques for cancer detection and classification from biomedical images and genome sequences.
2. A novel automatic segmentation and classification method using deep learning. Transfer learning approach is used when the limited biomedical dataset attained.
3. Performance measurement techniques applied to show the robustness of the model comparing state-of-the-art methods.

1.6 Outline of the Whole Report

The rest of the thesis is structured as follows:

- **Chapter 1** discusses the thesis objectives and highlights the research contributions and challenges.
- **Chapter 2** provides a brief study of biomedical imaging modalities, classification, segmentation, and gene expression analysis with blockchain technology.
- **Chapter 3** introduces a fusion based approach for lung nodule identification and classification.
- **Chapter 4** proposes an ensemble of convolutional neural network for microscopic image classification based on breast histological image and cervical cancer image.
- **Chapter 5** establishes a segmentation method using deformable convolutional w-net based on three different datasets.
- **Chapter 6** presents an ensemble approach for microarray gene expression classification for breast and colorectal cancer.
- **Chapter 7** proposes an association study for radio-genomics and histo-genomics analysis for predicting disease biomarkers.

- **Chapter 8** describes the blockchain based framework for secured sharing of biomedical contents in a distributed environment.
- **Chapter 9** recapitulates the research findings and delivers the objectives of future research with concluding remarks.

2 RELETED WORK

In this chapter, we contemplate an outline of certain background notions of the subsequent segments of this thesis. We provide a summary of medical imaging techniques with the theoretical background are vital for diagnosis of cancer.

2.1 Biomedical Imaging

Biomedical imaging plays an essential role for the diagnosis of cancer stages and hence it is very crucial in healthcare system now-a-days. Medical images can be obtained from different imaging modalities (or techniques) which provide a reliable and non-invasive assessment of diverse cancers [5]. Medical image analysis can extract the meaningful information about the different aspects of diseases conditions. The images are a collection of numerical values as the picture elements known as *pixels*. In an image, each dimension consists of the number of pixels define the *resolution*. In an image, 8,294,400 number of pixels are in a resolution of 8.3 megapixels that consists of 3840 × 2160 pixels. *Volumetric image* of a medical image is the sampled 2D images form 3D image in three dimensions lengthways [6]-[7]. The *voxels* refer as a 3D volumetric image that translate the three-dimensional associations among 3D pixels. Using image processing techniques both 2D and 3D images encoded information can be extracted and interpreted.

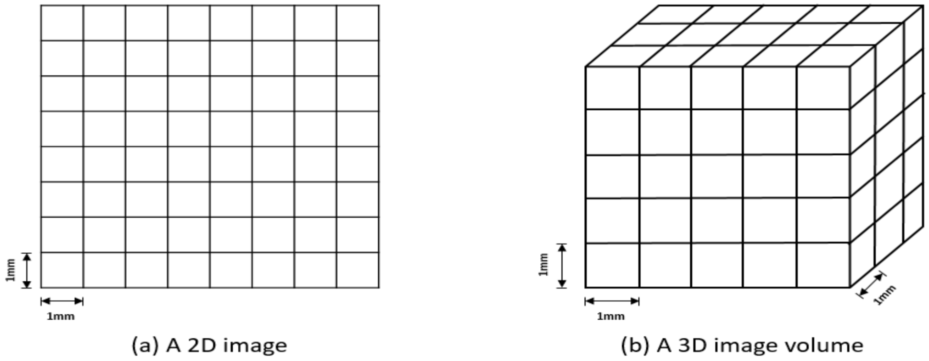


Figure 2.1 Image Pixels, Voxels, and Resolutions.

Contrast resolution is a set of concentration colors of red, green, and blue (RGB) that is possible to separable in both colored and greyscale image. In Figure 2.1 shows a 2D and 3D volumetric image. Pixel and voxel are pictured in two dimension and three-dimension arrays of grid.

2.2 Biomedical Image Modalities

Different types of imaging modalities are used by modern healthcare for cancer diagnostic and for treatment by attaining different features of a human body. These techniques are split into two classes based on approaches and process in for visualizing diverse features of disease. These are Anatomical and Physiological imaging. Normally Anatomical image captures and pictures the structures of anatomy of the range of interest (ROI) in two or three dimensions [8]. These images assist doctors to understand and assess conditions of disease for diagnosis. These images also assist to identify the response for a specific treatment evolution. For capturing the metabolic condition of the ROI is responsible for the functional imaging. This will help the physicians to evaluate the physiological conditions of patients to recognize the structural anomalies like tumors.

Normally an anatomical clinical modality uses X-ray image, computed tomography (CT) and magnetic resonance imaging (MRI). Single-photon emission computed tomography (SPECT) and positron emission tomography (PET) are included in functional imaging [9]. These imaging techniques generate a single image or generate image volumes. This called single-modality medical imaging.

Now a day for cancer diagnosis multi-modality imaging like PET/CT and SPECT/CT becomes popular diagnosis procedure in clinical studies and in medical research.

2.3 Magnetic Resonance Image (MRI)

This is the most used imaging technique in cancer segmentation, diagnosis, and prediction in modern clinical conditions. MRI uses blend of several X-ray to create tomographic images. In the Figure 2.2 shows a CT image of the brain.

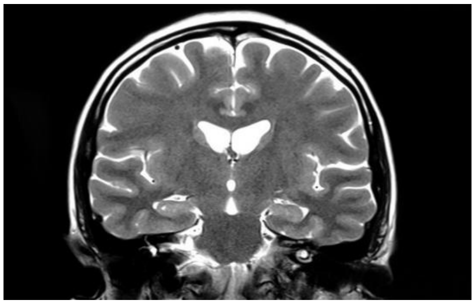


Figure 2.2 Brain MRI Image of NSCLC Dataset

2.4 Computed Tomography (CT)

It is one of the most exploited imaging modalities for detecting cancer, diagnosis, and prediction. It makes also use blend of numerous X-ray to generate tomographic images. Figure 2.3 shows CT image of the lung field of Non-Small Cell Lung Cancer (NSCLC).

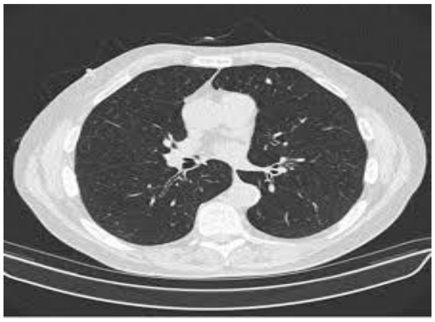


Figure 2.3 NSCLC CT Scan Dataset

2.5 Histological Images

Histopathological imaging is the golden measurement for diagnosis cancer. This is the result from biopsies of tumor tissue. The samples of tumor tissue are marked, then collected and then mounted onto glass slides for graphical examination. This assist to picturing the cellular structures ROIs of tumor. This also gives information on types of mutated cell and therapeutic insights.

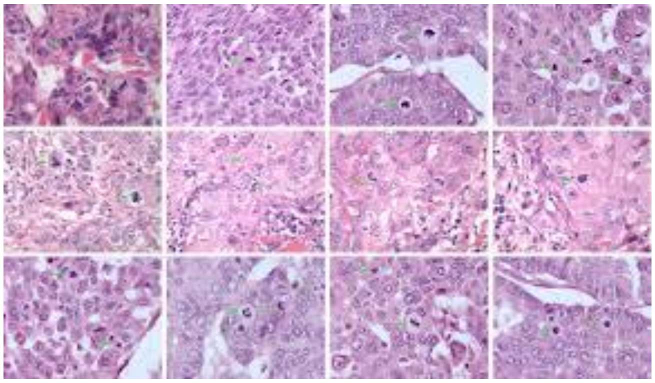


Figure 2.4 Breast Histology Image of BreKHis Dataset

High-resolution whole-slide-imaging (WSI) in histopathology that facilitated the numerical evaluation of tumor histomorphometry as well as its involvement with clinical methods for cancer diagnosis [10, 11]. Similar to CT images, digital WSI enables medical image analysis techniques to be applied to extract image features of abnormal cells. Image traits can then be exploited to correlate to genetic profiles of the tumor. Figure 2.4 shows the presence of sentinel lymph node in breast cancer metastases.

2.6 Medical Image Classification

This part represents different methods of medical image classification that shows the authentication of our claims. It is shown here that the existing automated image classification approaches to associate with the manual image classification to assist the experts and reflect their distinct advantages.

Image classification is said to be the most essential phase of digital image processing. It is appealing for a *"pretty picture"* or an image with the magnitude of coolers signifying numerous features. It is surely useless without knowing the colors meaning. The main two classification approaches are Supervised Classification and Unsupervised Classification.

Supervised classification is responsible in identifying the classes of Information in the image. They are called *"training sites"*. For every classification class, to develop a statistical characterization the image processing software is used. It is known as "*signature analysis"*. For each information class the statistical characterization is created. Then the image is classified through reflection of examine for each pixel. It helps to make verdict about the signatures that look like most.

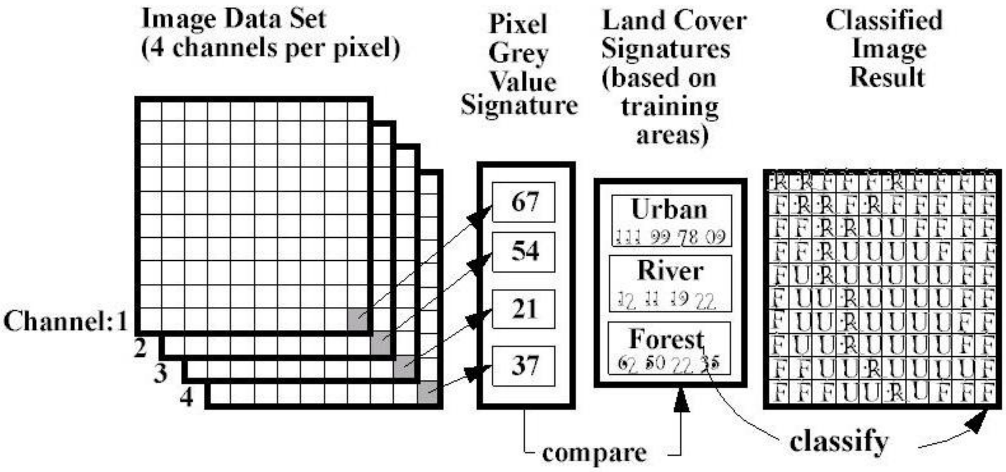


Figure 2.5 Supervised Classification [10]

In Unsupervised classification method a huge number of unidentified pixels are examined. All the pixels are divided into several classed and these are depending on the natural groupings, that is depending on values of the pixel. This kind of classification does not rely on analyst quantified training data as supervised classification. The pixels within a given region type should not be diverse to each other in the same dimension. However, different classes' data should be separated well.

2.7 Medical Image Segmentation

This part describes different methods for segmentation of medical image. In the area of medical imagine, segmentation is a process of image dividing. It separates an image in different part where it holds a group of pixels that collectedly signify an ROI. It eliminates unconnected image areas, thus lessen the complexity.

In maximum medical applications, segmentation of a medical image is described by an expert physician is count as golden standard. For cancer patients medical image segmentation are trusts upon the visual examination of images and the manual description of tumor ROI. Figure 2.6 gives the ROIs in different section of medical image modalities. These are defined by qualified physicians.

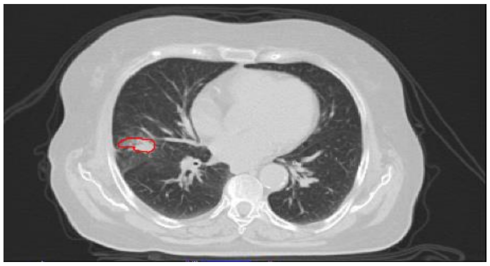


Figure 2.6 Region of Interest (Lung Cancer) in a CT image

Automated segmentation of medical images helps practitioner immensely to identify lesion in medical images by gathering useful information that may lead to take necessary action. Several architectures have been proposed for medical image lesion detection and segmentation tasks. Semantic segmentation has been overly used last couple of years that incorporating various structure like U-net, V-net, Y-net, and W-net. Figure 2.7 depicted a diagram on W-net for medical image segmentation and analysis tasks.

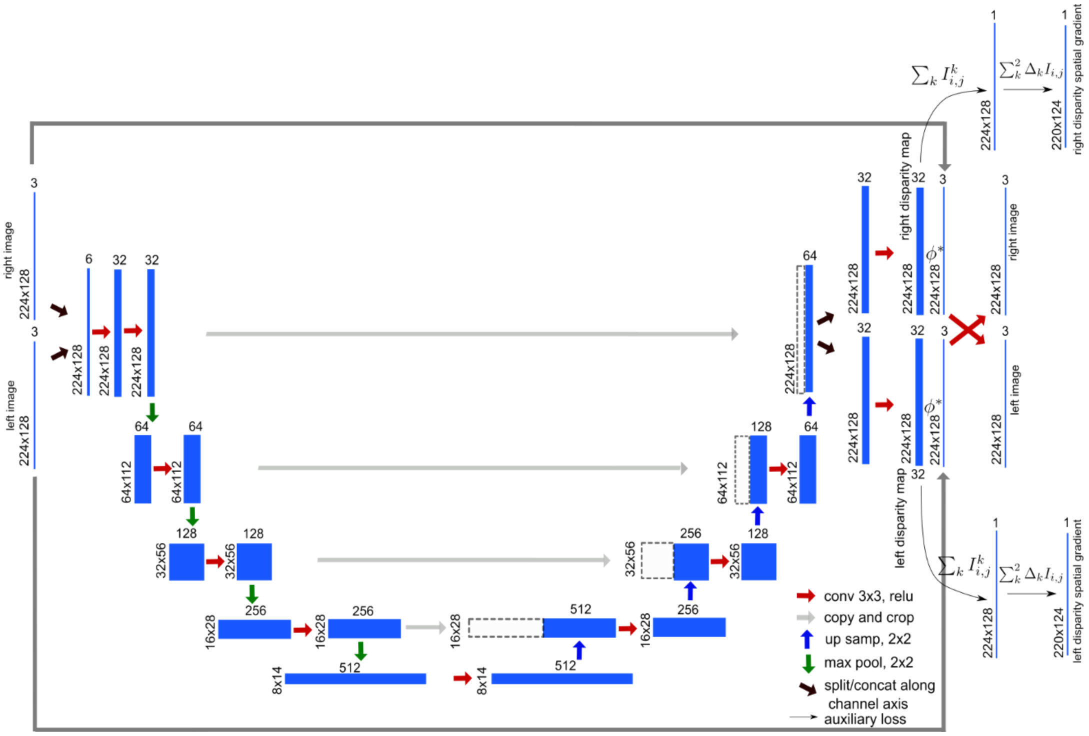


Figure 2-7 W-net for Image Segmentation [11]

2.8 Gene Expression Profiling

The visual examination is measured only the diagnosis scheme of tumor histopathology. But sometimes the diagnosis error is found in tumor classification due to graphical unclear morphological properties. A quantitative perception can be obtained as Gene expression profiling facilitates tumor identification and predicting for the future clinical diagnosis for the cancer [12]. This subsection is mainly designed for genes core concept, its expressions and application in medical studies.

A definite region of the deoxyribonucleic acid (DNA) strands for a gene. It can achieve ribonucleic acid (RNA) using genetic codes through transcription. Synthesizing proteins is the characteristics of RNAs to encode biological functions through translation. DNA change or damage is very common as the genetic codes allow mutation and permanent alteration of genetic elements. However, the change in DNA, mutations could be resulted in changes in functions and behavior of genes. The multiple prognosis of cancer genes is the result of mutation according to the evidence of Clinical studies. Gene expression profiling is responsible for the patient’s genetic information and the application of individual precision medicine.

Gene expression profiling depends on the invasive surgical procedures. So, the use of gene expression profiling is controlled although its significance in the diagnosis and prediction of cancer patients. Unfortunately, strong phenotypic and genetic heterogeneity is seemed to allow the cancer growth for an individual manifesting at multiple sites. The different characteristics of treatment depends on the varieties in gene expressions across multiple sites.

2.9 Learning Algorithms

In medical image processing the machine learning has added a great vale. The Machine learning techniques increase the ability to develop the automated algorithm in clinical studies. To develop any computer aided system the expertise in this field is the demand of time. One of the subbranch of machine learning the class of Deep learning which requires to train multiple train data level. For example, the convolutional neural network is the implementation of deep layers which can extract the potential image feature [13]. This section emphasizes the principles of deep learning models especially the CNNs architecture because they present the current sophisticated image in recognition of the object of the image. The relevant architecture of CNN model is given bellow.

2.9.1 Artificial Neural Network

ANN introduces the machine according to the biological principle of human brain structure [14]. It forms nodes said as "artificial neurons" and builds a graph structure of neural network. The machine learning uses to extract the nonlinear relationship of experiment data by ANN. The artificial neuron processes and transmits signals using the network of ANN as same as biological synapse. ANN needs to achieve an optimal architecture to harmonize network weights and bias to get the appropriate outcome.

In ANNs the basic building blocks are the neuron. Using Figure 2.8 it is clear how the neuron takes multiple input values and produce a single output value. Here the input vector is X of n elements, the weight vector is W for the same number of elements, equation 2.3.1 summarizes the whole scenarios:

$$\mathrm{L}\ (W, X) = \Sigma_{\,X=0}{}^{n} f\ (xi, wi) + \text{bias} \tag{2.1}$$

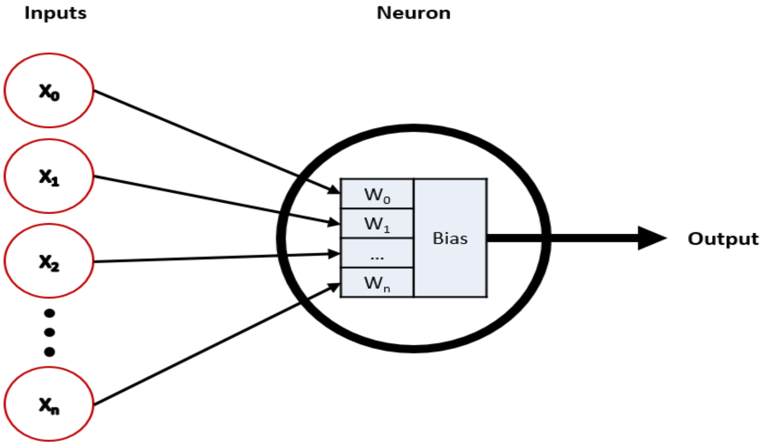


Figure 2.8 Single Layer Neuron [12]

In single model of neuron depicted in Figure 2.8 contains ANN training processes. To transmit the component of training data it uses the form of feature vector. The classification is conducted in two parts; each neuron is multiplied by its internal weight, then performs on the input vector whether it is above or below the threshold value.

2.9.2 Machine Learning

The modern healthcare system depends much of the principle of machine learning. Main principle of machine learning is to recognize the pattern using intelligence mathematical model. The goal is achieved by classification and retrieval of image, in tumor image section. The basic machine learning techniques as ruled-based system is developed the models of artificial intelligence. It is essential to "teach" to identify the distinctive features of patterns in machine learning for the data in order to produce outputs of money. Typically, learning methods are divided as: supervised, unsupervised and reinforcement learning.

While multiple disciplines of observable and reinforcement learning demonstrate strong potential. The marked medical imaging data is used to discover region-specific genetic associations. Since the outcome to this work are primarily related to supervisory methods, we only need them here.

- Training Data: is the repository of data to train a model used in machine learning. The training data presentation and its future relationship is obtained by machine learning model.
- Validity data: is a distinct data set to monitor the training process. This data set is made compare the output data with the predicted one by the training data labels. The performance of unseen data is approximate by the model parameter.
- Test data: is a collection of recorded data to measure the accuracy of the model after the training process is completed. The test data shows the effectiveness of the trained model. Test data is not including in the process of training.

2.9.3 Deep Learning

In-depth learning is classified as a section of machine learning techniques; this allows multiple processing-level calculation models to learn multiple levels of abstraction from the internal representation of input data. It performs a limited task of traditional thematic machine learning method to process ordinary data in its raw form depth. Deep learning solves this problem by solving abstract representations from high-level understanding of input datasets, to low-level learning and interpreting multiple processing levels. This is accomplished by continuously giving raw information through a continuous multilayer architecture, where the deeper layers have learned abstract representations from the representations of the previous layers.

The multilayer model for teaching under the supervision of in-depth teaching strategies employs error propagation or "backpropagation" for training [15]. Back proposition calculates the gradient of the ANN's weight-related error function and moves the gradient backwards through the neural network. The backflow of error gradients allows for efficient calculation of gradients for ANNs.

GPUs are used to advance the effectiveness of the training process from 10 to 20 times compared to usual therapeutic training methods in standard CPUs through the application of deep learning techniques.

2.9.3.1 Convolutional Neural Networks

It is a special type of ANN that is designed to work with input data in multidimensional arrays form, such as color two dimension that contain 2D arrays for each RGB channel. CNN training is easier and generalized. It is organized in different phases, and each phase has a special layer with unique functions. CNN's building blocks have three specialized layers: convoluted, pooling, and activation.

The convoluted layers contain units that organized in feature maps’ form. Its each unit is linked to local patches via a set of weights from the previous layer referred to as the filter bank shown in Figure 2.9. The sum of the results of the local filter bank is conveyed over a linear activation layer. Having same kind of feature map is shared with the same filter bank. This value allows evolutionary layers to identify local combinations of local properties from the preceding layer. Because of highly connected probability of local values and swap locations at the input of an image.

Pooling layers are for merging features in three-dimensional closeness that has semantic similarities. The principle behind the pooling layers is to locate the themes which are usually formed by highly connected features through a thick granular method.

CNNs are stimulated by the transmission of biological signals over cells with various complexities and different functions in visual neuroscience. CNNs use compositional orders in ordinary signal processing.

The high-level features output from a combination of low-level features. In the images, objects combine edges, motifs, and parts. CNN has made great strides in the sophisticated industry in multiple branches. CNN has been used mainly in medical image examination to perform multiple tasks. These include image segmentation, classification, and identification of diseases.

Google net, a deep CNN architecture submitted in 2014 as part of the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). It has 22 levels. Google Net is created to performance as a classifier for images of nature.

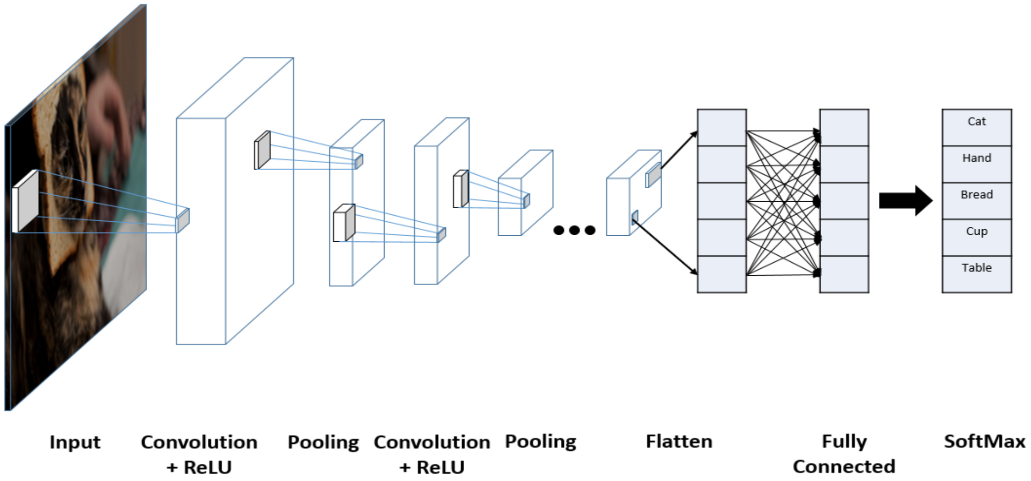


Figure 2.9 CNN with Convolution Layers [15].

It included the fundamental concept of increasing width and depth. Between this work it is keeping calculation costs constant. In the sophisticated region of time, Google net showed significant improvements.

2.10 Performance Evaluation Metrics

In image retrieval and classification, performance evaluation can be measured with a variety of metric accuracy and retrieval metrics commonly used in these domains. In image retrieval, accuracy refers to the percentage of true positive images and this percentage of true positive images regained in all relevant images in the datasets. This describe as below:

$$Precision = tp \,/\, tp + fp \tag{2.2}$$

$$Recall = tp/\, tp + fn \tag{2.3}$$

Where true is the number of positive or relevant retrieved images, false positive or irrelevant retrieved images and false negative or relevant retrieved images. Therefore, the accuracy measures the accuracy of the recovery, where the recovery measures the ability to recover relevant items from the database. Another metric that can be used for evaluation purposes is the F-score metric, which is matched with accuracy and can vary in varying degrees depending on the use score:

$$Fscore = (1 + score^2)\, Precision\ .\ Recall\ /\ score^2\ .\ Precision + Recall \qquad (2.4)$$

Therefore, increasing the score may increase the weight of the recall metric from accuracy and vice versa. In the image classification, the accuracy of a classifier at the test stage is a true positive number compared to the total classified examples.

2.11 Blockchain

During this age of digitalization, the health care sector needs to more secure way to transfer data between different stakeholders. The number of health care data is also is big enough to handle with great sensitiveness. Medical data has different types of formatting and representation. These data include health records, images, sensor data, genome sequences, bills, and payments and many more. Collecting and combining all these data can help to get run machine learning and different analysis to get inside information. This information will be very valuable for different sector from medical center to patient and between them many other actors.

Making a common intelligent system for medical information management, it is very important to bring all format of data in a common format that they can understand each other. This system data is increasing every second. So, volume of the dataset also a big challenge to handle in this system.

The GDPR and HIPAA are two regulations in Europe and United states as advocate patient's privacy. This describe the law about the rights of a patients to give access to their medical data [16]. It is given the data protection of a patient. To build a total electronic health care system, it is crucial to keep information confidential from different level of stakeholders according to the permission of the patient. In the electronic health care system, maintaining the data, giving permission to different stakeholders, formatting the data for all platform, all these increase the system complexity and through a big challenge.

Health care eco system ensuring the security of the health record and share this data in secure way to national and international wide. For this need a common architecture for share with different actors in the system.

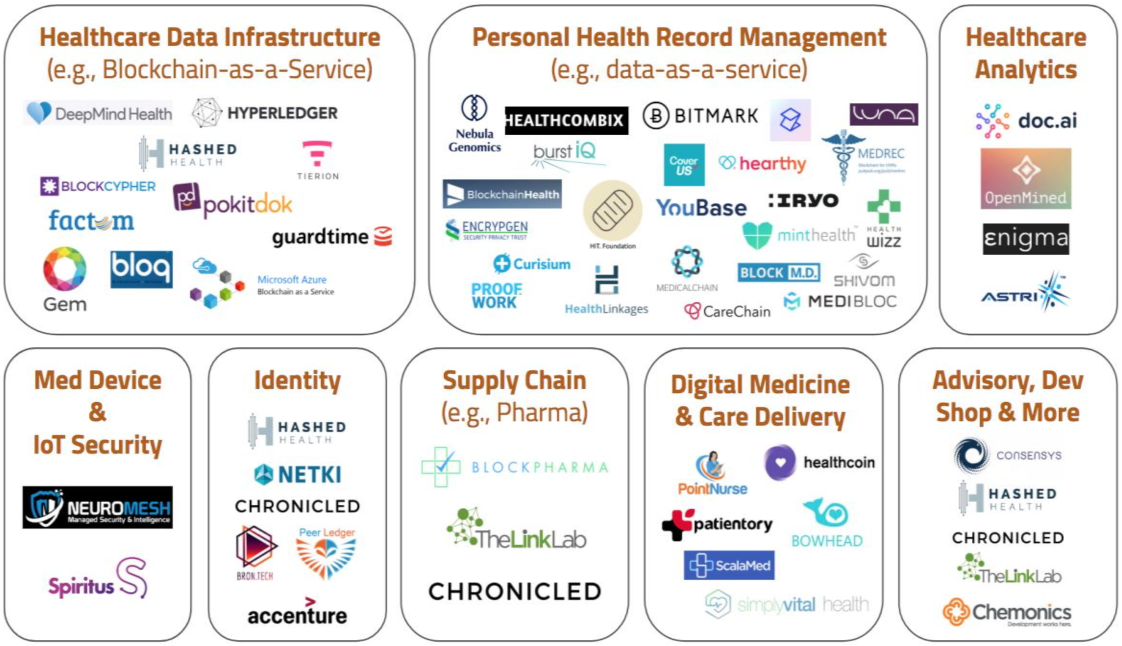


Figure 2.10 Healthcare Related Blockchain Project [16]

Blockchain could be a game changer in this filed with high security and in the distributed form. This technology gets the attention of both academia and industry. This technology employs complete control to a single user with giving fully controlled from a central point. For sharing health care, this technology enhances the privacy-preserving data. In the scenario of chronic diseases, data need to share between different physician to get multiple medication. For getting better treatment of chronic diseases, may need to go different places for different types of medication. Using blockchain technology the data could share in a secure way. This technic optimized the supply chain process. It is now crucial to implement the blockchain in the medical information eco system to share the information in secure process and give patient a optimized solution for his/her treatment.

3 LUNG CANCER DETECTION AND CLASSIFICATION IN CT IMAGERY USING HYDRA-NET WITH FEATURE FUSION

In this chapter, we contemplate an outline of certain background notions of the subsequent segments of this thesis. We provide a summary of medical imaging techniques with the theoretical background that are crucial for cancer diagnosis.

3.1 Introduction

Deep learning is a new branch of machine learning techniques that involve with high level data abstractions by considering a series of non-linear transformations. The complex model architectures followed by these algorithms to achieve better performances from large-scale unlabelled data and it passes through several hidden layers to identify and extract the complex features that finally produce an output. The deep learning uses previous layer computations in terms of identifying features and that is a power of the technique [17]. In recent years, deep learning has great impact on medical data processing, like lung cancer detection, skin cancer detection, etc. Lung cancer is the most dangerous cancer for a human being that increasing in an alarming rate with $12 billion in health care costs yearly in the US. Doctors are fully depending on the CT scans but could not be able to identify the disease. Therefore, computer aided support is needed for the doctors to make their job easy and that can reduce the death rate due to lung cancer. There are several machine learning classification approaches used in past few years by the researchers like SVM, RNN etc [18]. However, to effectively handle the complex structure SVM, KNN techniques are not good enough. Therefore, Deep learning technique is a good one to handle the complex situation without considering a huge number of nodes that is used in traditional machine learning technique like SVM and KNN [19]. It is now growing rapidly in different health sectors like bioinformatics, brain image analysis, retinal image analysis etc.

In this research, two deep learning approaches has been applied to identify the lung cancer from CT scan images. Firstly, an autoencoder approach is used with data pre-processing with u-net architecture is used for nodule detection. Secondly, a 3D Descent is used for lung nodule detection and classification. Finally, a hydra net is used for feature fusion of the two afore-mentioned model. The proposed hydra model shows significant performance comparing with state-of-the-art methods. Several machine learning and statistical techniques was used to detect lung cancer nodules with better accuracy rate. The rest of the paper is organized as follows: section 3.2 considered the datasets. Section 3.3 considered the pre-processing, 3.4 considered the methodology and, 3.5 described the experimental results. Finally, the whole paper terminates with a summary in section 3.6.

3.2 Datasets

In this research, Kaggle datasets used first that enclosed with CT scan images of different patients labeled with no cancer and cancer. For the training set, CT scan images from 1397 number of patients has been taken and 198 patients CT scan images is considered as test set. Moreover, the images are gray scale of size 512 x 512 that is very large, and the images are already labeled. Though each patient's datasets are already labeled, there is no information about the location of the nodules.

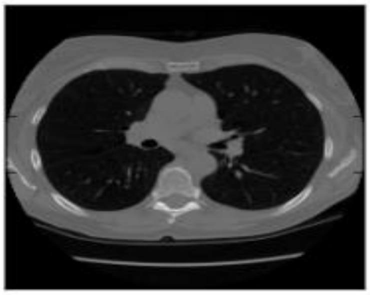


Figure 3.1 Kaggle CT Scan Image

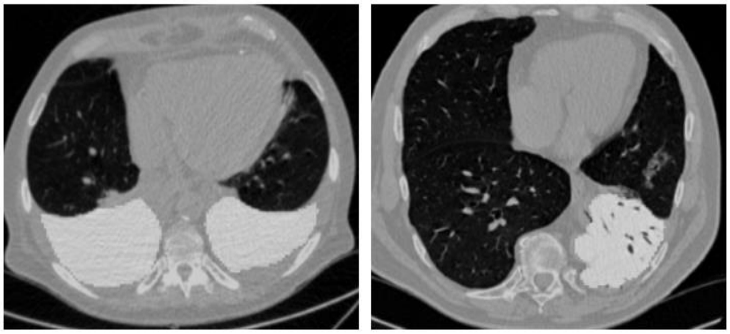


Figure 3.2 LUNA16 CT Scan Image

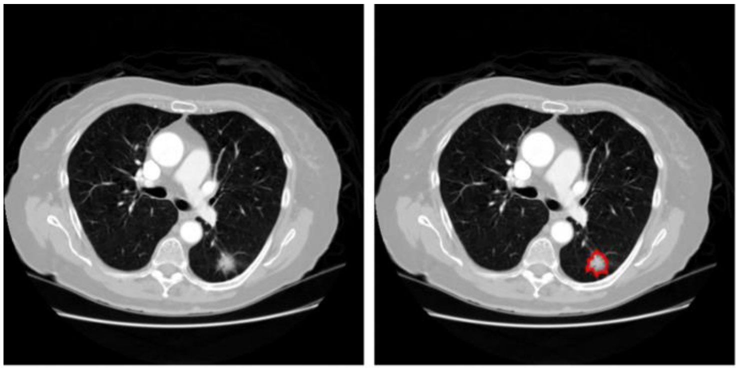


Figure 3.3 LIDC-IDRI CT Scan Image

In terms of LUNA16 datasets [2][4], the CT scan images of the patients is considered with labeled datasets. Moreover, there are some annotations that make it more useful to identify the location and classification of nodules. Kaggle dataset contained 25000 gray scale CT scan images and the LUNA16 contained 888 CT scan images. Moreover, LIDC-IDRI dataset consists of 686 lung nodule samples with 1010 CT scans. The sample image for the three different datasets is mentioned in figure 3.1, 3.2, and 3.3.

3.3 Methodology

In this section, the pre-processing stages employed in the experiments is explained along with the approaches to address the problem.

3.3.1 Pre-processing

The pre-processing steps covers the following areas like segmentation of nodule from lung, z-score normalization and finally nodule detection using u-net architecture

3.3.2 Lung Segmentation

Lung segmentation is the key task to work with lung images to identify the lung disease. The candidates of the lung nodules regions are considered from the CT scan images. False positives reduction has been performed by applying thresholding value calculation technique depending on the pixels intensity that separate the lungs pixels from the image. Kaggle and LUNA datasets is considered for the segmentation purpose along with the filtering method like erosion and dilation to calculate the patches [20].

3.3.2.1 Z-Score Normalization

There are several normalization technique available now-a-days. Max-min normalization and z-score normalization technique is overly used in data mining and machine learning arena. In this research, z-score normalization technique is used for the CT scan images that subtracted the mean to centralize the images. Then the standard deviation is considered to apply division technique to complete the normalization process.

3.3.2.2 Nodule Detection

Nodule detection is very important to identify the lung cancer regions from an image. U-net [8] structure of neural network performs significantly well to segment the images and properly detect the nodules. CT scan images is considered from both the dataset like Kaggle and LUNA for the segmentation purpose and after carefully segmented the images using U-net architecture, the dataset is fed to the deep neural network for classification purpose.

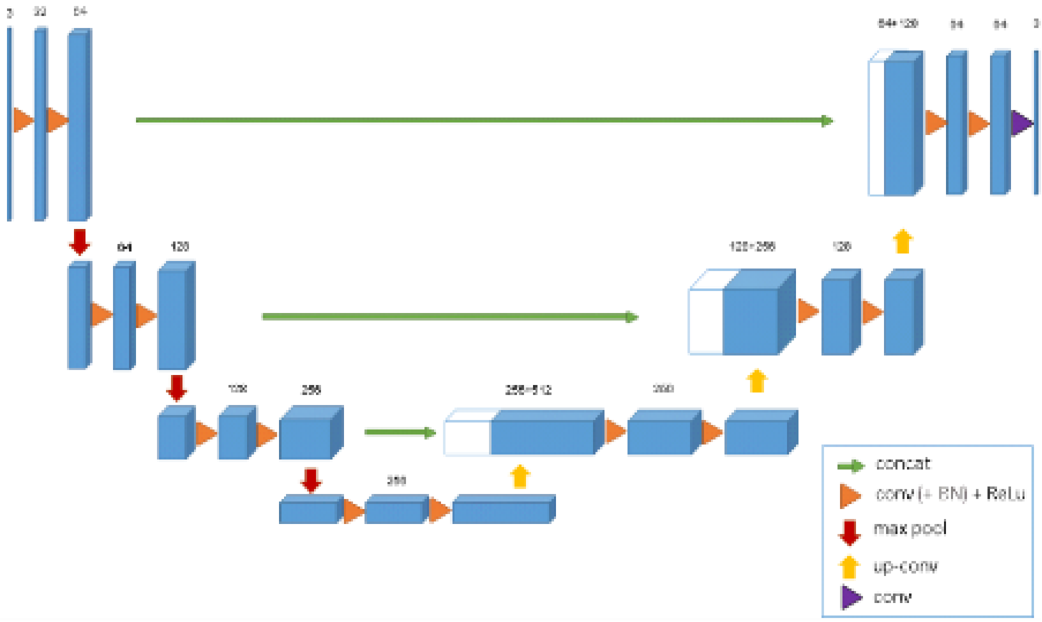


Figure 3.4 U-net Structure for Nodule Detection [21]

Binary mask output can be obtained after all the processes that represents the specific location of the nodule in an image. The U-net architecture for segmenting the images has been mentioned in Figure 3.4. The quantity of nodules for specific patient also obtained from this process with cancer and no cancer chances labeled mentioned in figure 3.5 and 3.6.

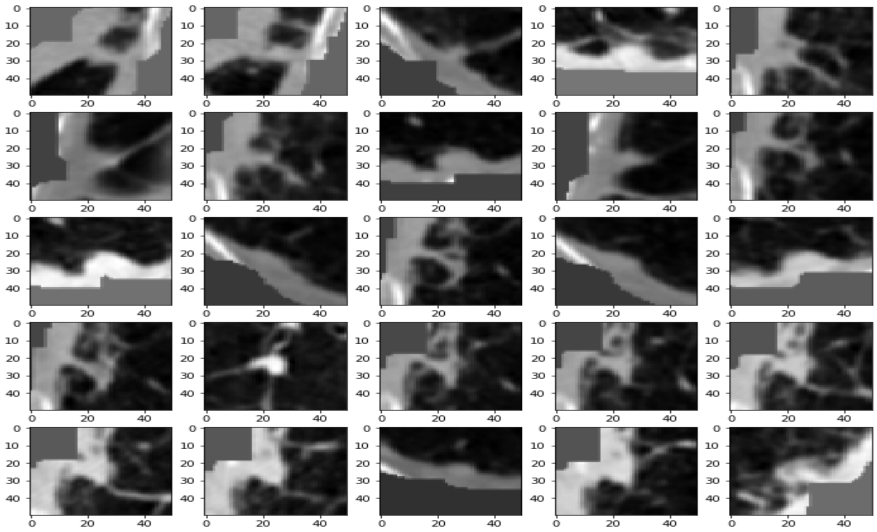


Figure 3.5 Quantity of Nodules Identified for a Patient Label Zero

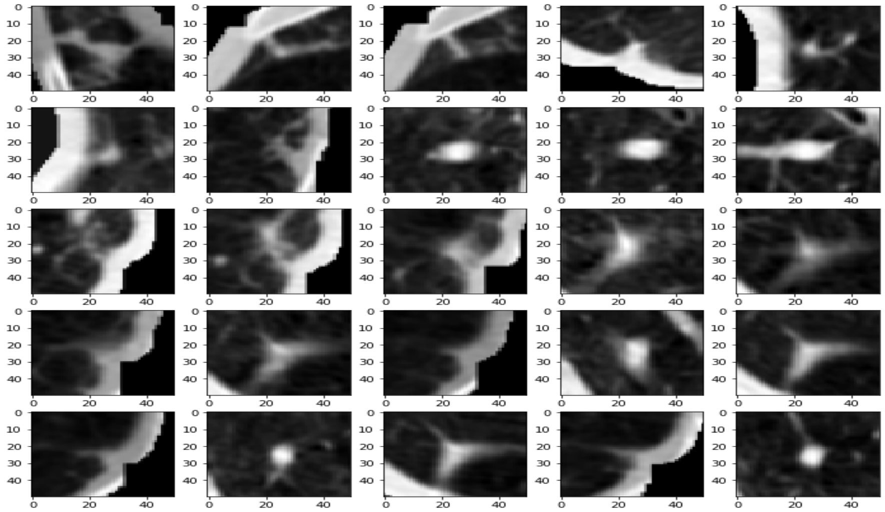


Figure 3.6 Quantity of Nodules Identified for a Patient Label One

3.3.2.3 Convolutional Autoencoder Approach

Convolutional autoencoder approach is very popular for lung nodule detection from CT scan images where the image patches are considered from the lung input images. After detecting the patches from with a view to identify the nodules in the images, encoding operation is performed to represent all the patches in a feature vector. Finally, SVM is used for the classification purpose. U-net architecture mentioned before is used for nodule detection.

3.3.2.4 Autoencoders

Autoencoder approaches have been considered to identify the lung cancer and eventually compared all the processes to come up with the best approach. In the first approach, the patches are identified from the CT scans images after detecting the nodules with appropriate pixel values. 48 x 48 patches extracted from the image and three techniques like autoencoder, local binary patterns evaluation, flattening is applied to encode the patches. The feature vector is identified from all the patches that generate 64-dimensional vector and eventually fed into the SVM classifier.

The architecture of the autoencoder used here is shown in Fig. 3.7. The convolutional autoencoder used in this approach.

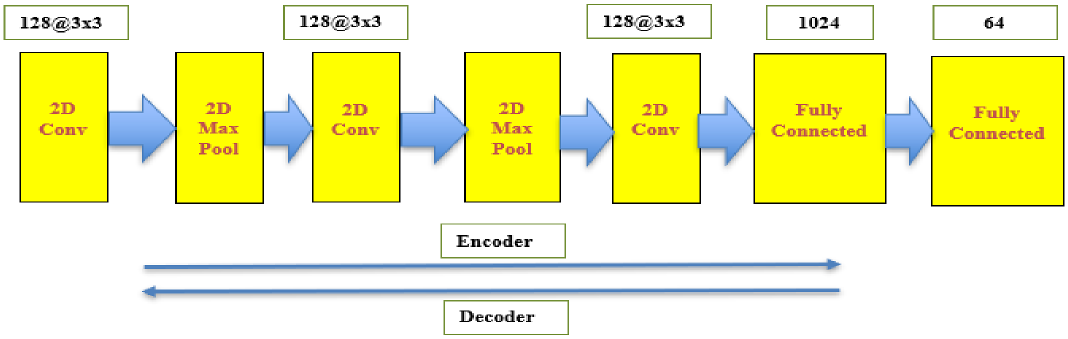
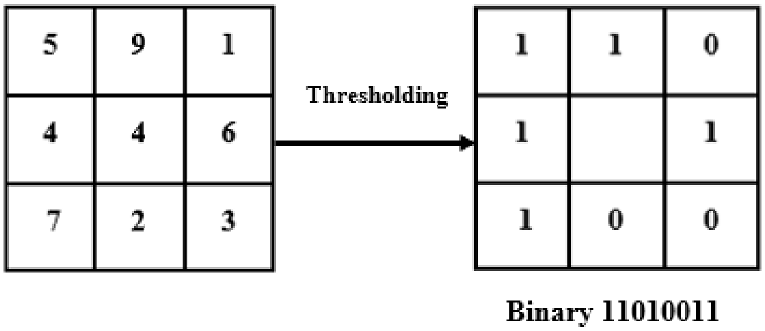


Figure 3.7 Convolutional Autoencoder

3.3.3 Local Binary Patterns (LBP)

LBP is a popular technique also used termed as local binary patterns that generate the binary code from the image by calculating the value of the neighboring pixels and update the current pixel value. The feature vector is generated from the n bit number of neighbors and can be used to express an image with the binary bit patterns mentioned in figure 3.8. Flattening is the method where the raw pixels of the patches is used for encoding purpose and represented using a vector [22]. The whole process mentioned in figure 3.8 where a convolutional neural network is used.



5	9	1
4	4	6
7	2	3

1	1	0
1		1
1	0	0

Figure 3.8 Local Binary Pattern for an Image Patch

3.3.3.1 Flattening

For encoding the candidate patches raw pixel patches is used to form a vector. In here 48 x 48 patches is simply used as a vector. In terms of patient level representation, a feature vector is needed for each patient to describe whether the patient experienced lung cancer disease or not. Finally, the average value of encoding is considered based on all the patches to represent the final feature vector.

3.3.4 3D DenseNet Approach

The DenseNet architecture consists of 3D convolutional neural network with sequence of Dense block and transition layer. The performance of 3D DenseNet is exceedingly high comparing with other state-of-the-art deep learning model as the overall distance in between input to the output is less. Due to the architecture with shorter distance, the better optimization result can be obtained as the vanishing gradient would be enhanced. The overall architecture of 3D CNN is mentioned in figure 3.8 where the 3D volumetric image is used as the input and passed through the 3D convolutional layer and the sequence of Dense block is considered along with transition layer and finally fully connected layer is used with softmax to classify the value. In this approach, lung cancer 3D CT scans is used as input, rather than 2D CT scans. In this model, 3D volumes of shape 120 x 512 x 512 is considered with the batch size of 32.

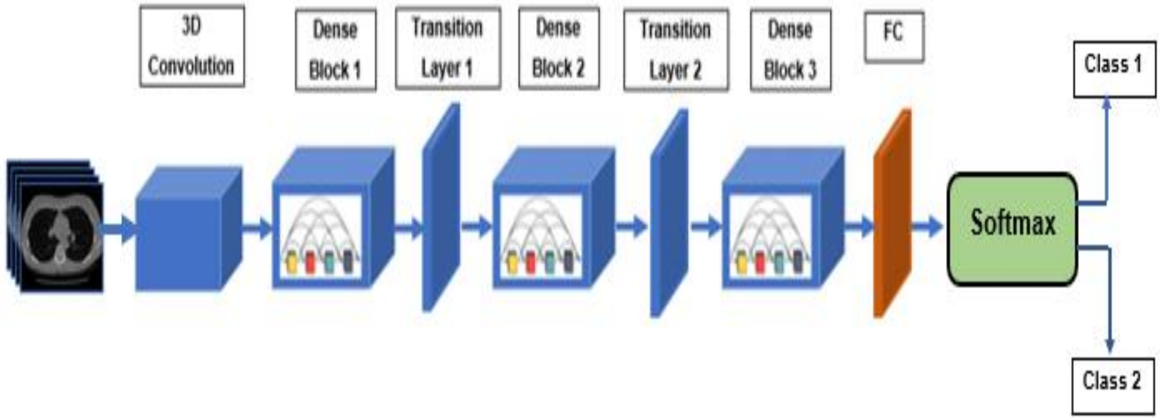


Figure 3.9 3D DenseNet.

3.4 Hydra Net Approach

An ensemble of deep neural network is considered for forming a hydra net architecture. The training set, initial weights, and the number of layers is varying tremendously to form the hydra net architecture. In our proposed approach, we considered the 2D CNN and 3D DenseNet of different architecture to combine and form the hydra net structure. The important thing of this kind of network is it is complex and time consuming to deal with different weight from initial to final layer with varying structure. Hydra architecture considers various transformation based on geometry in terms of the training samples of the network.

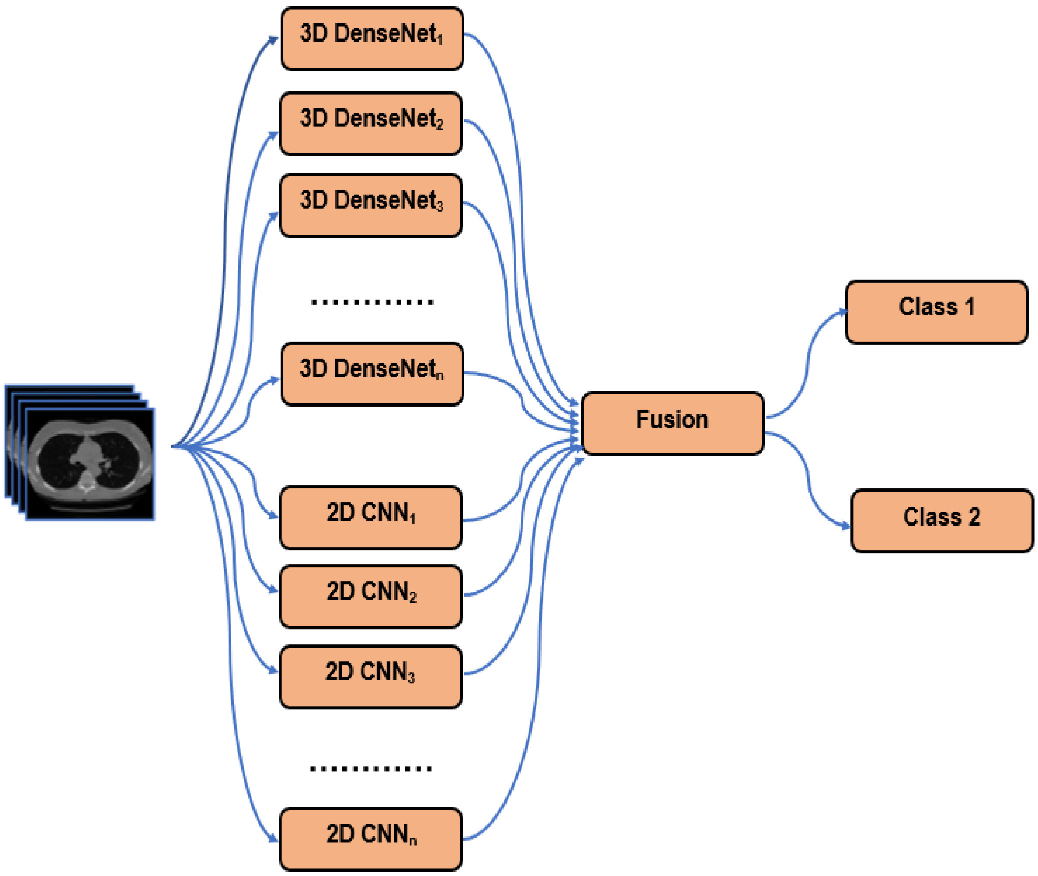


Figure 3.10 Hydra Net

Moreover, hydra net calculates the results by combining the values of the heads with score vector that represent the classification. Therefore, each produces a score vector result and all the head score values is considered. Finally, the values with highest majority voting would be considered for identifying the final label. Depending on the number of heads and number of votes, we can determine whether there are false positives or not.

3.5 Experimental Results

All the methods that were used in this research was implemented in python with other essential libraries like NumPy, Scikit-Learn, TensorFlow, Keras. In terms of performance evaluation, the U-net architecture is considered for detecting the nodules with true positive rate of

85 percent and the 65 percent for training and test datasets where the LUNA16 was considered. The hardware and software system specification are mentioned in Table 3.1

Table 3.1 Hardware and Software Description

Hardware	**Software**
Processor: i7-6000, 2.80 gigahertz	OS: 64-bit Windows 10
Primary Memory: 16 gigabytes RAM	API: Keras [50]
GPU: NVIDIA GeForce GTX 770	Backend: Tensorflow [51]
Storage: Solid State, 250 gigabytes	Language: Python 3.7.3 [52]

This experiment evaluates the computation result using the parameters, such as accuracy, specificity, sensitivity, and F-score. These parameters are achieved using True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) derived from the confusion matrix. Equation (3.1) to Equation (3.5) represents the computing formula of these five performance parameters from the value of confusion metrics.

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.1}$$

$$Specificity\ (SP) = \frac{TN}{TN+FP} \tag{3.2}$$

$$Sensitivity\ (SE) = \frac{TP}{TP+FN} \tag{3.3}$$

$$Precision\ (PR) = \frac{TP}{TP+FP} \tag{3.4}$$

$$F - score\ (Fs) = 2 \times \frac{SE \times PR}{SE+PR} \tag{3.5}$$

To evaluate the performance of the U-Net, we assumed that a nodule was detected correctly if the U-Net detected a nodule within 10 pixels of its correct location in the annotated mask. The U-Net described in 3.2.1 achieved a true positive rate of 85% and 65% on train and test sets, respectively, of the LUNA16 dataset, and an average false positive rate of 0.14 and 0.4 per slice on the same, respectively.

The log loss can be formulated as:

$$LogLoss = -\frac{1}{n} \sum_{i=1}^{n} [\, y_i \log(\widehat{y_i}) + (1 - \; y_i) \log(1 - \widehat{y_i}) \,] \tag{3.6}$$

Where, n is the number of patients in the test set

^y_i is the predicted probability of the image belonging to a patient with cancer

y_i is 1 if the diagnosis is cancer, 0 otherwise

The Kaggle dataset contains CT scans from 198 patients that are not labelled, these patients form the test set for the submission on which we are scored.

Table 3.2 Model Comparison

Method	Accuracy	Log Loss
2D CNN	81.21%	0.43257
3D DenseNet	84.74%	0.42865

From Table 3.2 it can be observed that, 3D DenseNet achieved better result compared with 2D CNN method where the log loss is near about 0.42. In 2D CNN, U-net architecture was used along with three auto encoder techniques that performs significantly better than using raw datasets directly into the classifier and heatmap techniques. 3D DenseNet performance is not better than 2D CNN. The accuracy of the dataset is achieved where 20028 samples is considered for the training set and the 9866 samples is used for the validation purpose with different epoch values.

Table 3.3 Comparison with the State-of-the-art Method

Method	Accuracy
Multi-Scale CNN [10]	86.84%
Vanilla 3D CNN [11]	87.40%
Hydra Net	91.75%

From Table 3.3 it can be observed that, multi-scale CNN shows the accuracy result of

86.84. The other deep learning models is compared with the proposed hydra net. The performance of the hydra net shows better accuracy results comparing with the other model as the different architecture with different weights is considered in hydra net. Finally, the majority voting scheme with strengthen vanishing gradient enhance the model performance.

3.6 Summary

Lung cancer detection has been a key research area last couple of years. The researcher has applied several methods to improve the performance of the computer aided process. Classification methods play key role to identify the nodules from the images with better prediction rate. In this research, two methods are considered to evaluate the performance. U-net architecture is used for nodule detection purpose. Convolutional neural network is used as the classification purpose and have shown greater performance along with autoencoder, binary patterns and flattening. The performance of the 3D DenseNet is superior than the 2D CNN with better log loss value. Moreover, the ensemble hydra net performs better than the single deep neural network model.

4 MICROSCOPIC IMAGE CLASSIFICATION USING ENSEMBLE OF DEEP CONVOLUTIONAL NEURAL NETWORK

4.1 Introduction

Microscopic image study has great impact for cancer detection and classification and the practitioner is fully focusing on the images to attain valuable information. Early detection and diagnosis can boost the possibilities of endurance rate of the cancer. Cancer is extremely spreading in the world and become the prominent cause of mortality of the mankind. Mammography and biopsy are the two overly applied methods for cancer disease diagnosis. Radiologist is fully responsible for early detection of the cancer in mammography whereas tissue is considered for biopsy in terms of detecting cancer. It is important to identify the cell where cancer occurred based on the shape, size, degree of malignancy, and distribution of tissue. However, identifying cancerous cell by visually inspecting it is time consuming and laborious task. Moreover, expert pathologist needed to go through the process and sometimes it is hard to identify cancerous cells from bunch of images. Therefore, effective identification of cancerous cell may lead to biopsy and may reduce the chance of infection and mortality [24].

Latest advancements in machine learning and image processing have facilitated the improvement of computer-aided diagnosis (CAD) systems for identifying and analyzing breast cancer and cervical cancer from the microscopic images more smoothly and faster way with higher accuracy results. The CAD system analyzes the microscopic images of the sample tissue, and finds the patterns corresponding to the cancerous and non-cancerous condition and classifies the images respectively into benign and malignant class. The major challenges associated with the classification of breast and cervical cancer images include the inherent complexity in microscopic images such as cell overlapping, subtle differences between images and uneven color distribution.

The objective of this study is to develop an accurate and reliable solution for breast cancer and cervical cancer classification.

In this study, we have analytically examined the deep learning models for automated identification of breast cancer. Key features of this work are as following:

- Deep convolutional neural network with transfer learning is considered to detect breast cancer and cervical cancer from microscopic images
- Data augmentation technique including rotation, cropping, flipping, and resizing has performed for image augmentation.
- Examined the performances of deep ensemble model with the state-of-the-art breast histology and cervical cancer image classification methods.

This chapter is organized as follows: Section 4.2 describes the proposed method for classification of benign and malignant histopathological images and cervical images. Section 4.3 describes the dataset and data augmentation method. Section 4.4 presents the evaluation metrics along with the experimental results. Finally, the conclusion is drawn in Section 4.5.

4.2 Methodology

In this research, we introduce a novel deep neural network architecture for microscopic image classification using transfer learning. In the proposed architecture, three state-of-the-art CNNs are used to extract features using transfer learning and fine tuning. The pretrained CNN models such as Xception [25], Resnet [25], and Inception-Resnet-v2 [25] would be considered as feature extractor for the proposed ensemble model for microscopic image classification. The overly applied environmental image dataset known as ImageNet dataset is employed with these CNNs. ImageNet consists of millions of natural images can be applied for biomedical images as the dataset is small. The extracted features from each pre-trained CNN would be combined in fully connected layer for microscopic image classification.

Google’s inception CNN model is extended to develop Xception [25] model by incorporating detpthwise separable convolutions and it has gained significant performance and have been applied for biomedical image analysis and applications. As different sized convolutional filters used in this model it reduces the number of hyperparameters as well as computational complexity. The Xception module is robust, stronger than the Inception module, and major composition inside the pretrained Xception architecture, as demonstrated in figure 4.1.

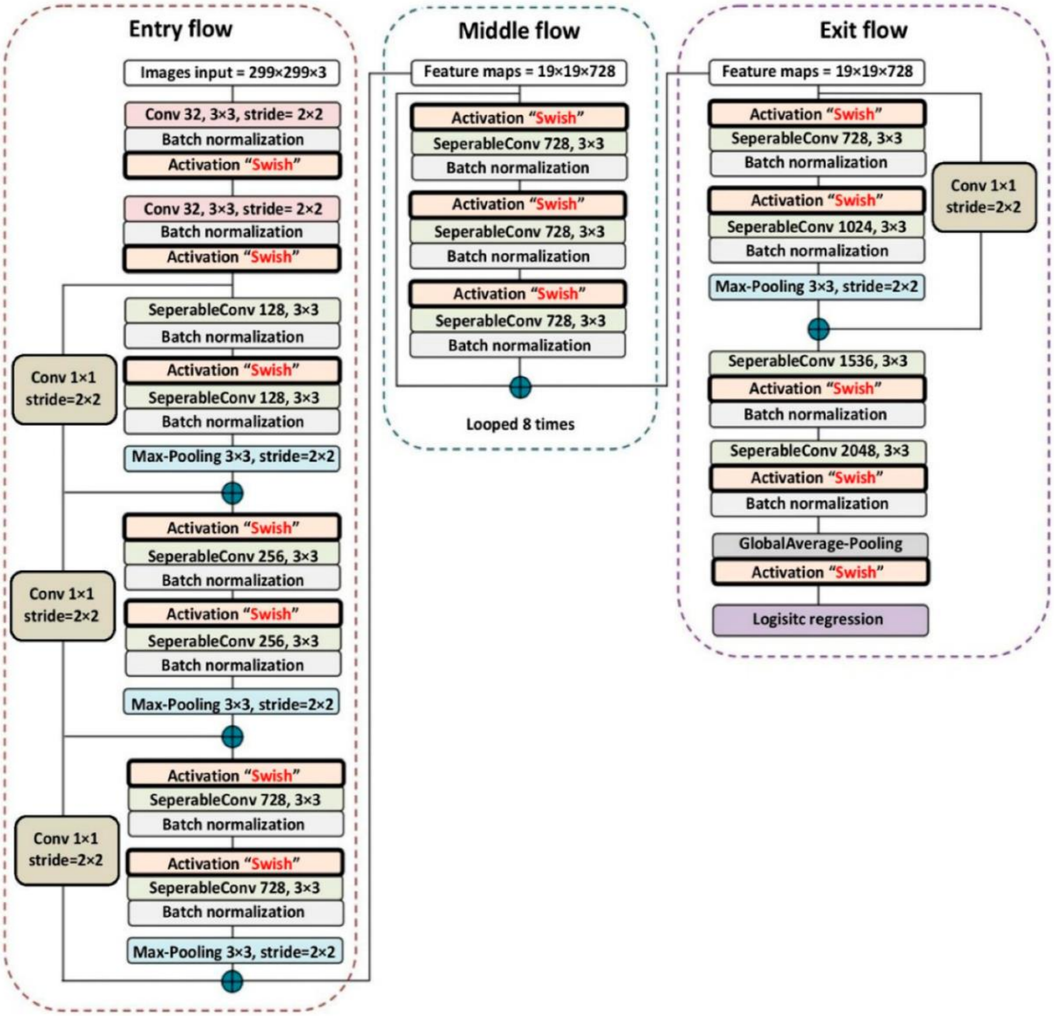


Figure 4.1 Pre-trained Xception [25]

Resnet is a deep residual network that consists of multiple neural networks with sequence of convolutional connection. In here we considered Resnet152, thought there are other Resnet version with different depth. Resnet152 is overly used for biomedical image classification and shows significant performance. It reduces the running time and computational complexity by decreasing the number of parameters. Figure 4.2 illustrates the basic architecture of Resnet152.

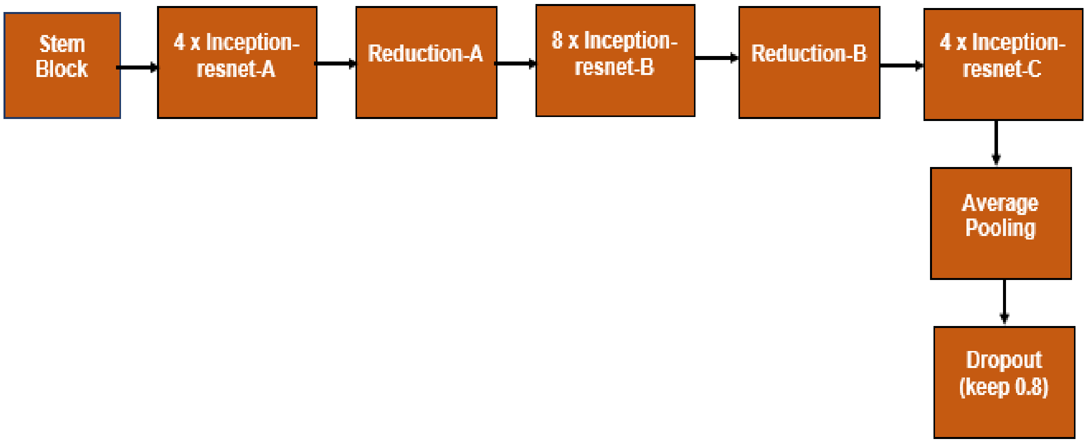


Figure 4.2 Pre-trained ResNet-152

Inception-Resnet-v2 [25] consider the blended of Inception CNN structure and Residual connection to form the new architecture. A sequence of convolutional layer with max pooling operation is considered in Inception-ResNet-v2 structure with the bunch of residual connections that decreases the running time of the network. Figure 4.3 shows the basic network architecture of Inception-Resnet-v2.

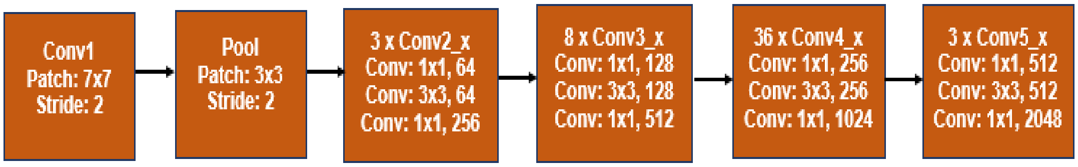


Figure 4.3 Pre-trained Inceptoin-ResNet-V2

4.3 The Proposed Ensemble Model

The proposed ensemble model is the combination of three pre-trained CNN model mentioned above. ImageNet dataset is used consists of millions of nature images that is used to train the three pre-trained CNN model by incorporating the transfer learning approach. Deep CNNs are proficient to discover basic image features that are appropriate to further image datasets without training from scratch. Figure 4.4 illustrates the transfer learning structure for a single CNN with fine tuning. The pretrained networks function as a feature extractor for generic image features and the two last layers are fully connected layers for classification. The details of the features generated by the pretrained deep CNNs are summarized as follows.

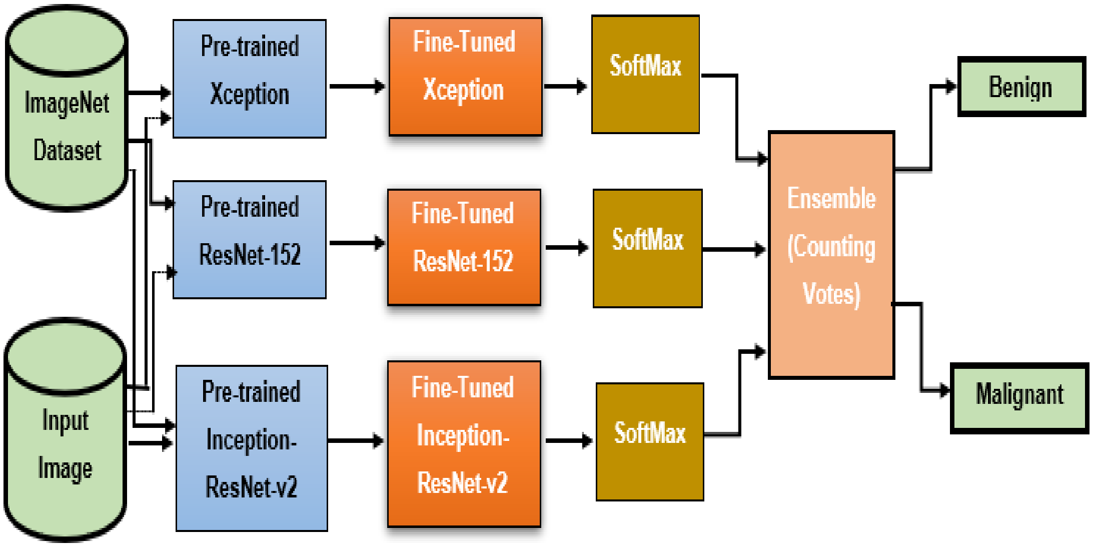


Figure 4.4 Ensemble Model for Breast Histology Image Classification

In proposed ensemble model, the voting approach is used. The output of different pre-trained CNN model majority voting is used and the class with most votes is considered as decisive label for test sample.

4.4 Datasets

Microscopic image dataset of breast cancer and cervical cancer is considered in this research. Breast cancer datasets with sub-types were collected from BreaKHis (The Breast Cancer Histopathological Images) [26]. BreaKHis consists of 7,909 pathological breast cancer images. The dataset with different magnification of 40X, 100X, 200X, and 400X from 82 patients were selected for sub-types classification. The number of benign images is 2,480 and the number of malignant images is 5,429 malignant. The dataset contained four distinct histological sub-types of benign breast tumors: adenosis, fibroadenoma, phyllodes tumor, and tubular adenona; as well as four malignant tumors: ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma.

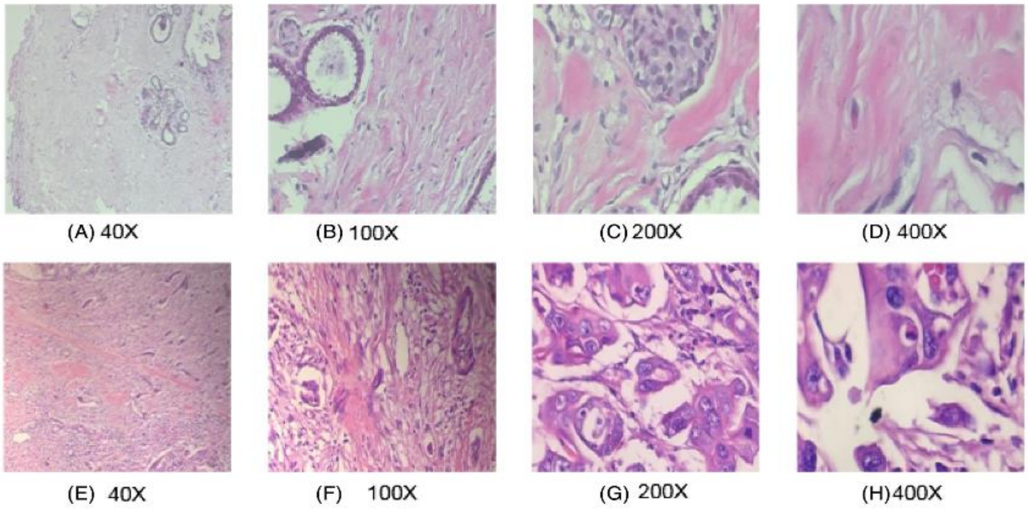


Figure 4.5 BreaKHis Dataset for Histopathological Images

In Table 4.1, BreaKHis dataset with different magnification factor is mentioned for different number of patients with benign and malignant classes. The dataset is split into training and test set. Approximately 80% of available data were chosen for the training set and remaining 20% of the data were used for performance evaluation.

Table 4.1 Number of Patients with Magnification Factor for BreaKHis Dataset.

	Magnification Factor				
Class	**40X**	**100X**	**200X**	**400X**	**Number of Patient**
Benign	625	644	623	588	24
Malignant	1370	1437	1390	1232	58
Total	1995	2081	2013	1820	82

In terms of cervical cancer, two overly used benchmark dataset is considered named PAP-smear dataset and 2D-Hela dataset. PAP-smear contains 917 microscopic images with benign and malignant. The 2D-Hela dataset contains microscopic images of 862 with 10 different categories [27]. In Figure 4.6, the sample of 2D-Hela dataset is mentioned and in Figure 4.7 representing PAP-smear dataset.

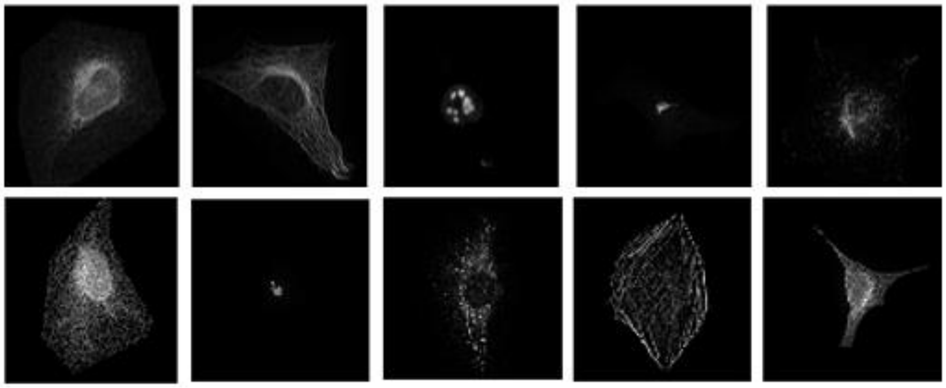


Figure 4.6   Cervical Cancer 2D-Hela Dataset

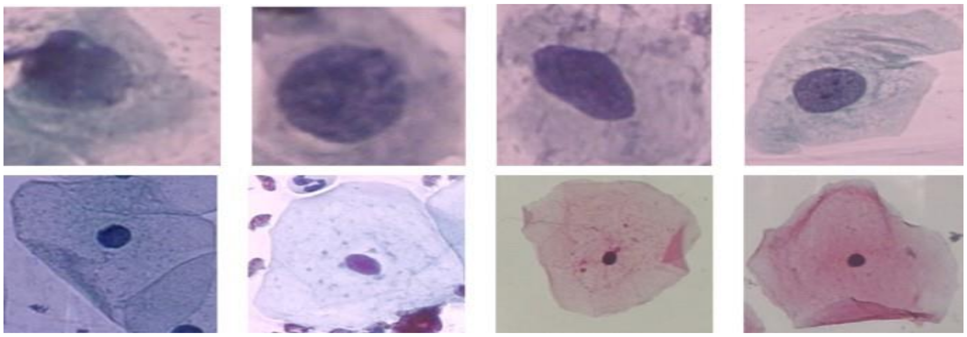


Figure 4.7   Cervical Cancer PAP-smear Dataset

4.4.1 Data Augmentation

Data augmentation is a crucial phase to consider when the dataset is too small to have sufficient samples to be considered to train a deep neural network with a view to perform features extraction process. The necessity of image augmentation is crucial to achieve better performance in deep learning classification model. The data augmentation process is applied for microscopic images both for breast histopathological images and cervical cancer images. Augmentor Python library [27] is used to perform the data augmentation process including random resizing, rotating, cropping, and flipping methods depicted in Figure 4.8.

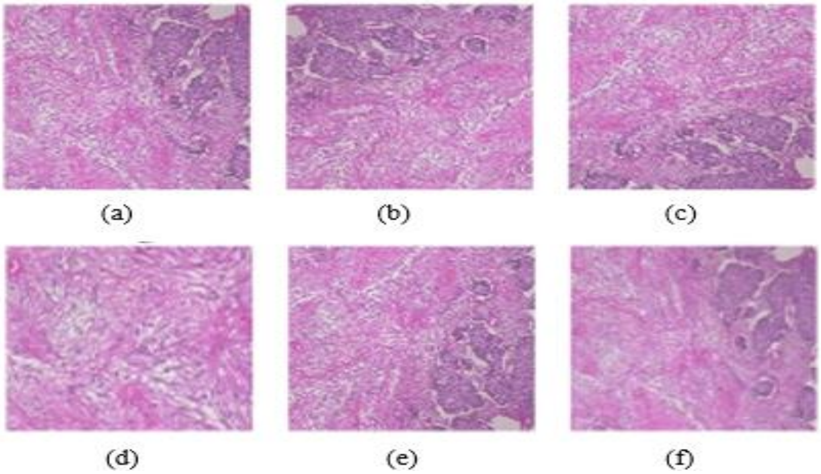


Figure 4.8 Data Augmentation Techniques including Rotating, Cropping, Flipping, and Resizing.

The input image (a) in the dataset is augment or rotated into 90-degrees (b) and 270 degrees (c). Then, image flipped (d) top bottom to right with 0.8 probability. Next image was cropped (e) with probability of 1 and percentage area of 0.5. Finally, input image was resized (f) with width=120 and height=120.

4.5 Results and Experiments

The proposed ensemble model with three pre-trained deep CNN model is considered for microscopic image classification. Technically, Python programming language with Keras and TensorFlow is used for the classification purpose [28]. The hardware and software system specification are mentioned in Table 4.2. The data augmentation technique is used to enhance the dataset as the breast cancer and cervical cancer dataset is too small to avoid overfitting. The dataset is split into training, validation, and testing set. The distribution of the dataset is 70% as training data, 10% as validation data and 20% as test data. The CNN model configuration is mentioned in Table 4.3 with batch size, optimizer, input size, and learning for each model.

Table 4.2 Hardware and Software Description

Hardware	**Software**
Processor: i7-6000, 2.80 gigahertz	OS: 64-bit Windows 10
Primary Memory: 16 gigabytes RAM	API: Keras [50]
GPU: NVIDIA GeForce GTX 770	Backend: TensorFlow [51]
Storage: Solid State, 250 gigabytes	Language: Python 3.7.3 [52]

Table 4.3 CNN Model Configuration

Model	**Input Size**	**Batch Size**	**Optimizer**	**Learning Rate**
Resnet152	299 x 299	12	RMSprop	0.045
Xception	299 x 299	12	SGD	0.045
Inception-ResNetv2	299 x 299	12	RMSprop	0.045

This experiment evaluates the computation result using the parameters, such as accuracy, specificity, sensitivity, and F-score. These parameters are achieved using True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) derived from the confusion matrix [29].

Equation (4.1) to Equation (4.5) represents the computing formula of these five performance parameters from the value of confusion metrics.

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.1}$$

$$Specificity\ (SP) = \frac{TN}{TN+FP} \tag{4.2}$$

$$Sensitivity\ (SE) = \frac{TP}{TP+FN} \tag{4.3}$$

$$Precision\ (PR) = \frac{TP}{TP+FP} \tag{4.4}$$

$$F-score\ (Fs) = 2 \times \frac{SE \times PR}{SE+PR} \tag{4.5}$$

The accuracy comparison of the three pre-trained method for breast cancer images is mentioned in Table 4.4 with other measurement technique such as precision recall, and F1-score. The accuracy of the individual CNN model along with the ensemble model result was compared with other state-of-the-art method.

Table 4.4 Comparative Study of Different CNN Model Breast Cancer

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Resnet152	94.12	0.93	0.95	0.94
Xception	95.75	0.95	0.97	0.96
Inception-Resnet v2	94.67	0.93	0.94	0.94
Ensemble	97.83	0.97	0.99	0.97

The accuracy comparison of the three pre-trained method for cervical cancer images is mentioned in Table 4.5 with other measurement technique such as precision recall, and F1-score. The accuracy and the standard deviation are combinedly represented in the table below. The individual CNN model along with the ensemble model accuracy result was compared with other state-of-the-art method.

Table 4.5 Comparative Study of Different CNN Model for Cervical Cancer

Model	**2D-Hela**	**PAP-smear**
Resnet152	89.72 ± 2.18	90.87 ± 1.48
Xception	90.72 ± 1.85	89.66 ± 1.89
InceptionResnetV2	92.00 ± 1.97	89.25 ± 2.23
Ensemble	96.72 ± 2.50	96.81 ± 1.75

(± std)

Results in Table 4.6 show the accuracy comparison with other methods. The last row of Table 4.6 also shows the results from our proposed ensemble method that has significantly improved in comparison to recent development study. Figure 4.9 represented the bar diagram of the accuracy comparison models.

Table 4.6 Accuracy Comparison of Breast Cancer

Method	**Accuracy (%)**
Nguyen [20]	92.63
Awan [21]	90.0
Kensert [22]	97.00
Vesal [23]	97.50
Khan [24]	97.52
Proposed Ensemble	97.83

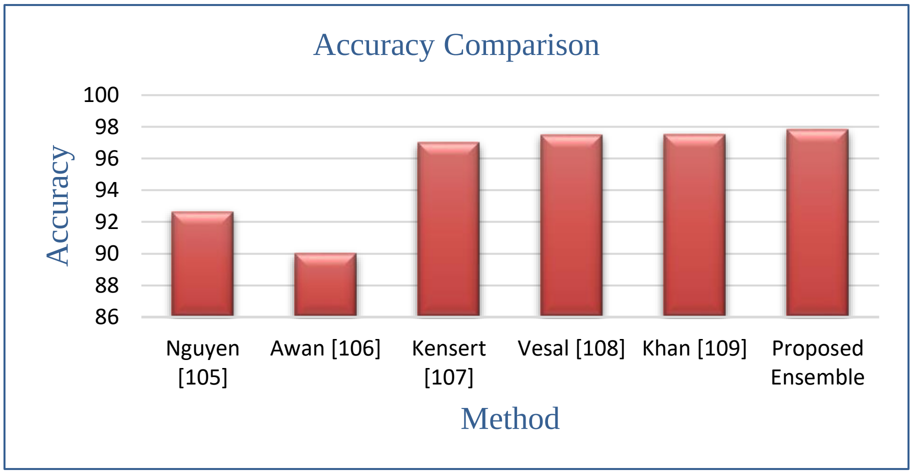


Figure 4.9 Accuracy Comparison of Breast Cancer Image

Results in Table 4.7 show the accuracy comparison with other methods. The last row of Table 4.6 also shows the results from our proposed ensemble method that has significantly improved in comparison to recent development study. Figure 4.10 represented the bar diagram of the accuracy comparison models.

Table 4.7 Accuracy Comparison of Cervical Cancer

Method	2D-Hela	PAP-smear
Lazebnik [30]	83.79 ± 2.5	84.03 ± 2.3
Ojala [31]	81.47 ± 2.1	81.43 ± 2.1
Liu [32]	86.20 ± 2.5	87.63 ± 2.1
Lin [32]	89.37 ± 1.5	89.96 ± 1.4
Long [33]	92.57 ± 2.46	92.63 ± 1.68
Proposed Ensemble	96.72 ± 2.5	96.81 ± 1.75

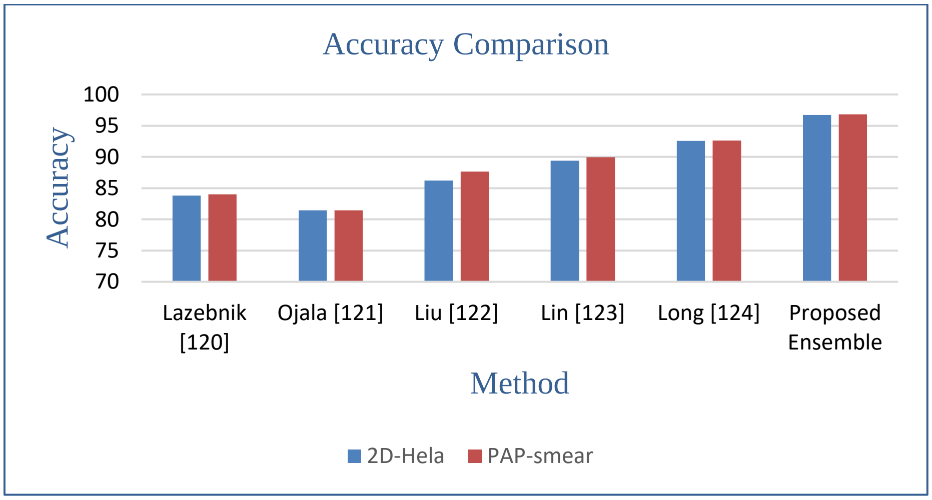


Figure 4.10 Accuracy Comparison Cervical Cancer Image

4.6 Summary

In this article, we proposed a novel deep learning framework for the detection and classification of breast cancer and cervical cancer using the concept of transfer learning. In this framework, features are extracting from microscopic images using three different CNN architectures which are combined using the concept of transfer learning for improving the accuracy of classification. The data augmentation technique is applied to enhance the volume of the images and that were fit for the CNN model to improve accuracy results. The performance of the proposed ensemble model is contrasted with state-of-the-art CNN model. It has been observed that the proposed ensemble model provides outstanding accuracy results without training from scratch that increases classification proficiency.

5 MULTI-ORGAN IMAGE SEGMENTATION BASED ON DEFORMABLE CASCADED W-NET

5.1 Introduction

Biomedical image analysis and segmentation has been a key task for investigating lesion in different parts of the body based on medical images with a view to collect useful information that might be helpful for doctors in terms of disease diagnosis. Automated analysis of medical images can reduce the chances of death by identifying and detecting abnormal region from an image [34]. Moreover, developments of machine learning and image processing technique have facilitated the improvement of computer-aided diagnosis (CAD) systems for incorporating segmentation techniques to detect different types of cancer in a faster way with higher accuracy results [35]. Extracting features from the images using machine learning techniques has been a key concern as the feature extraction processes from the images has enormous complexity. Semantic segmentation using deep learning technique has significant impact for lesion detection as it showed performances that can be comparable with state-of-the-art method. Semantic Segmentation technique has been applied for detecting lesion based on multi-organ segmentation methods of the images [36]. Deep learning techniques like U-net, V-net, Y-net, W-net has shown significant outcome for cancer lesion detection.

The objective of this study is to develop an accurate and reliable solution for cancer lesion detection and segmentation. We propose a deformable cascaded W-net based on U-net architecture. The proposed model shows significant performances for biomedical image segmentation [37].

In this study, we have analytically examined the deep learning models for automated identification of lesion with a noble segmentation technique. Key features of this work are as following:

- Deep learning technique using a deformable cascaded W-net is considered for semantic segmentation of medical images.
- Data analysis technique based on normalization and image patches generation process is applied. Generator network and discriminator network used to identify generator mask and annotated mask.
- Examined the performances of segmentation technique based on deep learning models with the state-of-the-art semantic segmentation techniques.

This chapter is organized as follows: Section 5.2 presents the problem statement. Section 5.3 describes the proposed method for medical image segmentation. Section 5.4 describes the datasets used in this research. Section 5.5 presents the evaluation metrics along with the experimental results. Finally, the conclusion is drawn in Section 5.6.

5.2 Problem Statement

Medical image segmentation has been a key issue last couple of decades. Unsupervised ways were overly used for effective segmentation of the images with pixelwise calculation and prediction for segmenting the images to detect objects Several techniques have been introduced like Markov Random Field method, Mean Shift method, etc [38]. However, supervised techniques get interest recent days in terms of image segmentation. Automated segmentation of the images can reduce the processing time and increase the performances of disease diagnosis. Machine learning and deep learning models within computer vision area has been shown significant outcome for image segmentation [39]. The deep learning models are evaluated based on fully convolutional networks to generate a pixelwise projection. The training of supervised learning models is utilized to discover filters to deliver segmentation results on medical images.

Due to different architecture of the deep learning model, it is not certain which framework would be good for semantic segmentation of the images. Deep learning techniques like U-net, V-net, Y-net, W-net has shown significant outcome for cancer lesion detection [40][41]. However, it is hard to play with the medical image dataset for semantic segmentation using deep learning techniques as the semantic segmentation models need a substantial volume of pixelwise labeled training data to carry out the segmentation task. Moreover, it is really challenging to accumulate the vast amount of labeled medical image data [42].

To solve the medical image segmentation problem, we proposed a deformable convoluted cascaded W-net for effective segmentation of the image. The objective of this study is to develop an accurate and reliable solution for cancer lesion detection and segmentation. The proposed deformable cascaded W-net is basically based on U-net architecture. The proposed model shows significant performances for biomedical image segmentation based on end-to-end and pixel-to-pixel manner [43].

5.3 Methodology

The primary goal of this work is to develop a deep learning models for accurate and reliable solution for cancer lesion segmentation on medical images. The architecture of the model is fully inspired by the overly used U-net architecture [44] and deformable convolutional network [45]. We proposed a deformable convoluted cascaded W-net for effective segmentation of the medical image for retinal vessel segmentation task, skin cancer lesion segmentation task, and lung cancer segmentation task. The proposed approach integrated the two architecture of Cascaded W-net and Deformable Convolutional Network.

Figure 5.1 and 5.2 showed the architecture of the two networks that integrated to form the novel deformable cascaded W-net for medical image segmentation purpose. The input raw medical images are pre-processed and generate patches to determine training and validation dataset. The patches generated from the three medical image datasets considered in this experiment based on patch sizes. The two networks like deformable convolution network and cascaded W-net are both end-to-end deep learning structures for semantic segmentation considered a 48x48 patch size. All outputs from the deep learning models are organized to produce a comprehensive segmentation map.

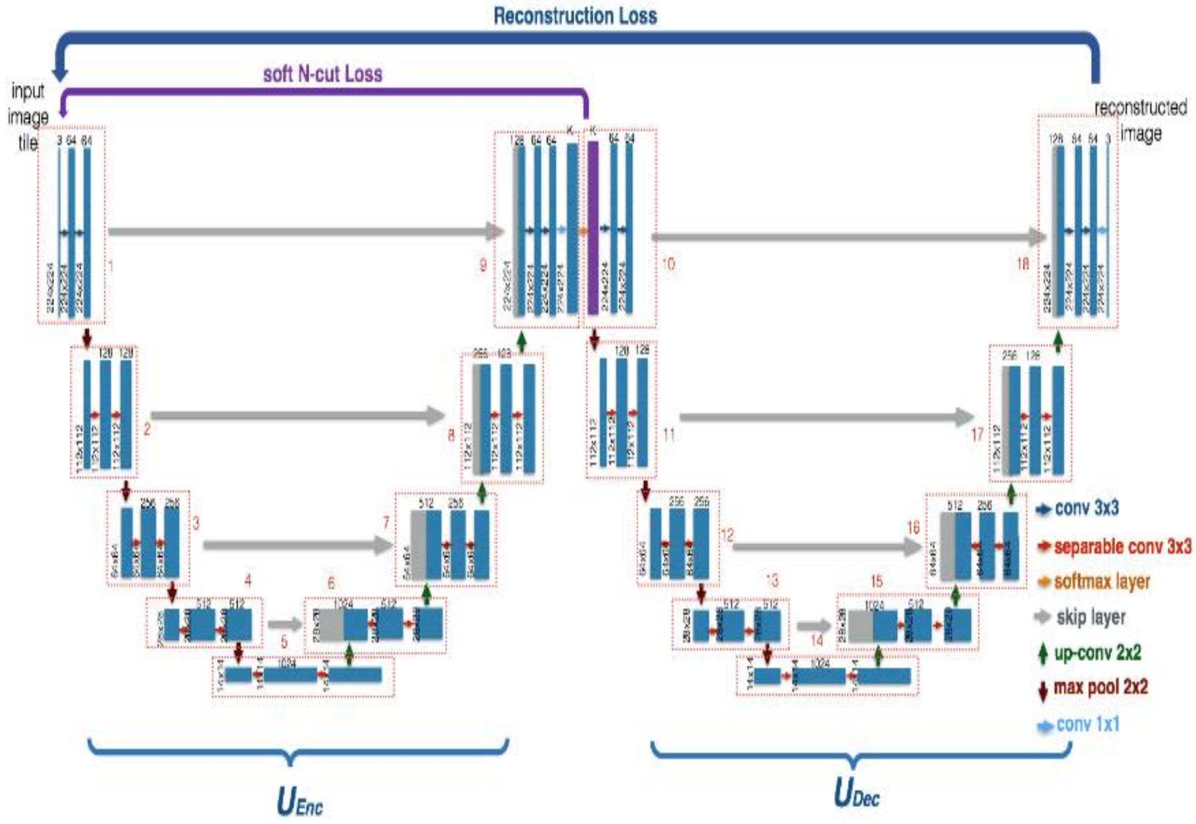


Figure 5.1 W-net Architecture [46]

Figure 5.1 depicted the cascaded W-net [46] architecture inspired by the architecture of the U-net [33] that contains encoder and decoder network with two sides for medical image segmentation. The convolutional layer of the deep learning network was replaced by the deformable convolutional block. The new model is trained to integrate the low-level feature with the high-level features, and the receptive field and sampling locations are trained to adaptive to vessels' scale and shape, both of which enable precise segmentation. Deformable convolutional network used the deformable convolutional block as encoding and decoding unit that integrated on top of cascaded W-net. The generator network and discriminator network used to identify generator mask and annotated mask by combining with the segmented output.

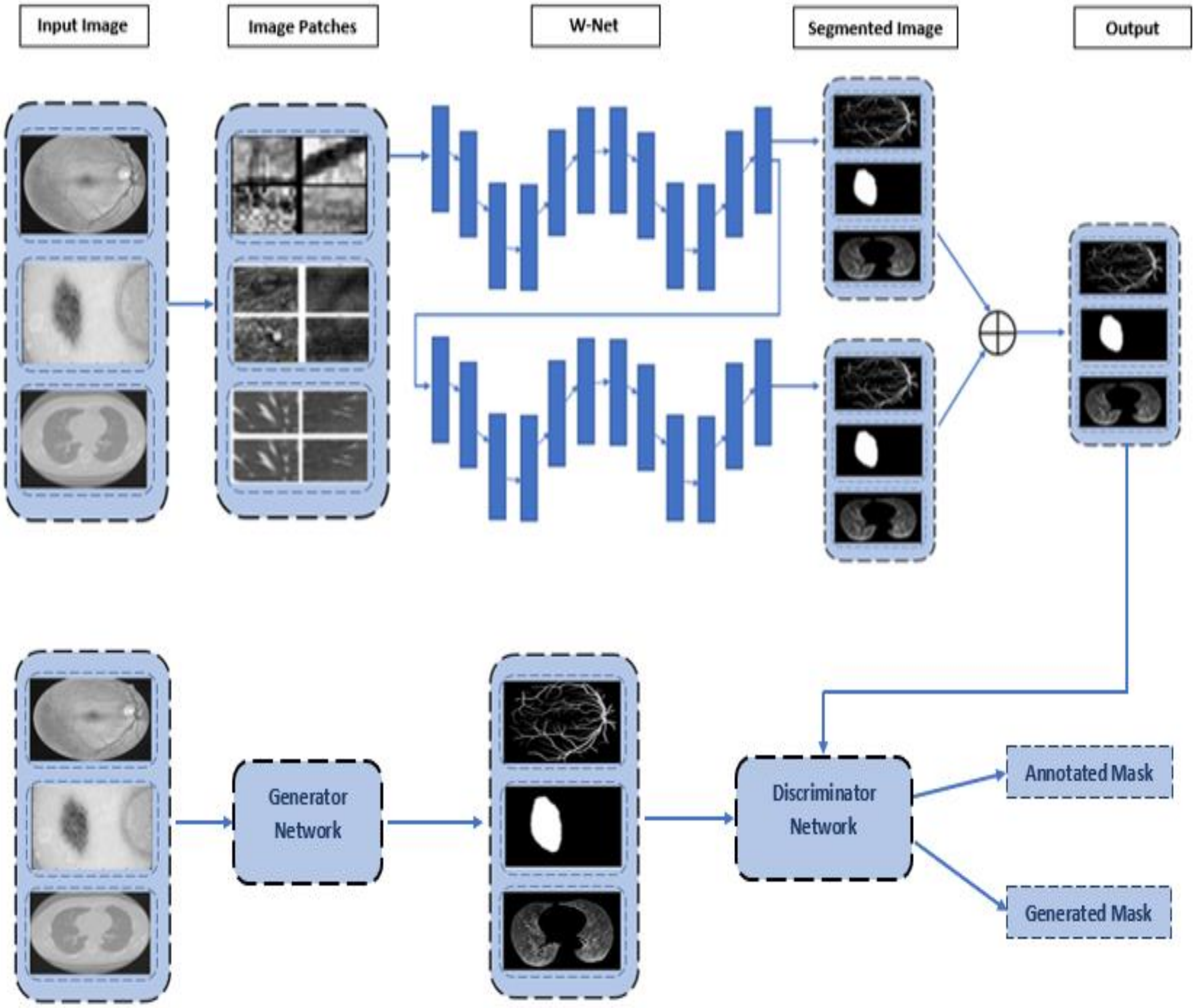


Figure 5.2 Proposed Deformable Cascaded W-net for Segmentation

Figure 5.2 illustrates the deformable cascaded W-net architecture consists of a convolutional encoder-decoder in a cascaded W-net framework. The deformable convolutional block comprises of a convolution layer with a batch normalization layer [47] and an activation layer that generate the offset. In terms of decoding, a typical convolution layer has been inputted to adapt filter numbers after merge operation. The features can learn on the deformable cascaded W-net to produce efficient segmentation results in terms of retinal blood vessel segmentation, skin cancer lesion segmentation, and lung cancer segmentation.

5.4 Datasets

5.4.1 Retinal Blood Vessel Dataset

Overly used DRIVE dataset is considered for blood vessel segmentation. The dataset comprised of color retinal images of 40 samples. From the dataset 50% of the samples is used as training dataset and another 50% of the samples is considered as test dataset. The images are of 565×584 pixels [48]. In terms of processing the images from DRIVE dataset, 220,000 patches were considered and split into training and testing patches of 180,000 patches and 40,000 patches, respectively.

5.4.2 Skin Cancer Dataset

Skin cancer dataset is considered from the Kaggle skin lesion segmentation competition of 2017. The dataset consists of 2000 samples and split into training and testing for model accuracy measurement. Out of 2000 samples, 1250 samples considered as training set and the rest of the 750 samples considered as testing and validation samples [49].

5.4.3 Lung Cancer Dataset

Lung cancer dataset is considered from the Kaggle Data Science Bowl of 2017 for detecting lung lesion from the CT images. The 2D and 3D CT lung images is used for the segmentation purpose. The size of the images is 512×512 and the dataset is split into training set and test set [49]. From the number of images, 80% is used as training set and 20% is used as test set.

5.5 Results and Experiments

The proposed model is applied for three different datasets for medical image segmentation. Technically, Python programming language with Keras and TensorFlow is used for the classification purpose. The hardware and software system specification are mentioned in Table 5.1. Batch size of 32 and number of epochs is considered 150 for all the experiment with optimizer, input size, and learning rate of the model.

Table 5.1 Hardware and Software Description

Hardware	Software
Processor: i7-6000, 2.80 gigahertz	OS: 64-bit Windows 10
Primary Memory: 16 gigabytes RAM	API: Keras [50]
GPU: NVIDIA GeForce GTX 770	Backend: TensorFlow [51]
Storage: Solid State, 250 gigabytes	Language: Python 3.7.3 [52]

This experiment evaluates the computation result using the parameters, such as accuracy, specificity, sensitivity, and F-score. These parameters are achieved using True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) derived from the confusion matrix. Equation (5.1) to Equation (5.5) represents the computing formula of these five performance parameters from the value of confusion metrics.

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.1}$$

$$Specificity\ (SP) = \frac{TN}{TN+FP} \tag{5.2}$$

$$Sensitivity\ (SE) = \frac{TP}{TP+FN} \tag{5.3}$$

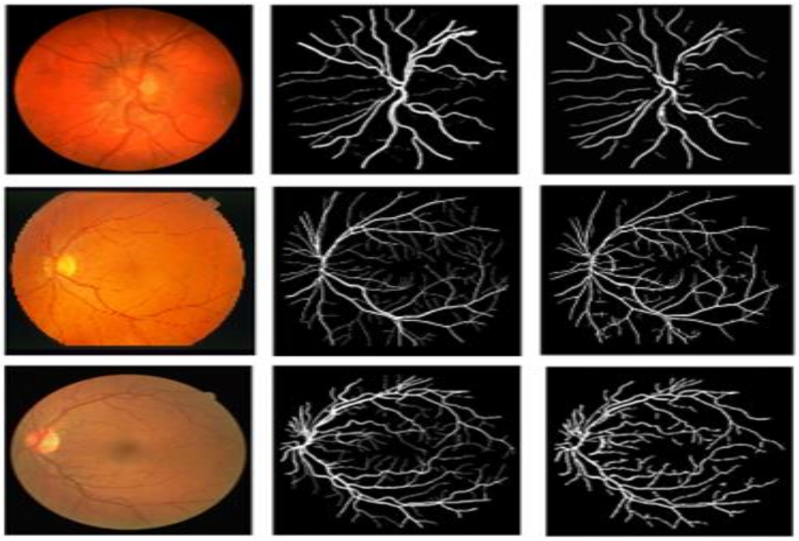
$$Precision\ (PR) = \frac{TP}{TP+FP} \tag{5.4}$$

$$F - score\ (Fs) = 2 \times \frac{SE \times PR}{SE+PR} \tag{5.5}$$

The accuracy comparison of the model for lesion detection and segmentation of different medical image datasets is mentioned in Table 5.2, 5.3, and 5.4 with other measurement technique such as precision recall, and F1-score. The accuracy of the model was compared with other state-of-the-art method and showed significant outcome.

5.5.1 Retina Blood Vessel Segmentation Results

Retinal blood vessel segmentation based on DRIVE dataset is considered here with the proposed segmentation technique using deformable cascaded W-net architecture. Figure 5.3 depicts the segmentation output of DRIVE dataset using three different columns. The retinal blood vessel input images represented in first column, the second column represented the ground truth, and the third column demonstrated the segmented image. It has been observed that the deformable cascaded W-net showed significant segmentation outcomes.



(a) Image (b) Ground Truth (c) Segmented Image

Figure 5.3 Segmentation Output of DRIVE Dataset

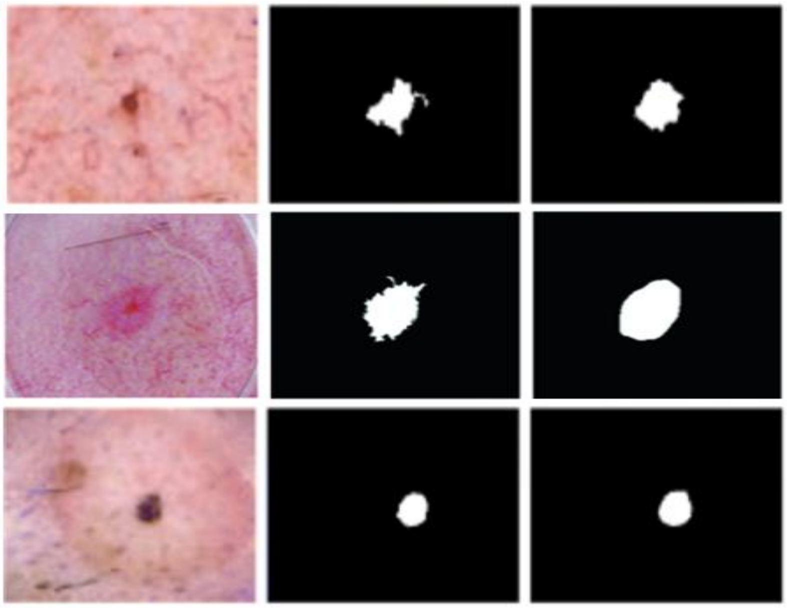
The proposed deformable cascaded W-net segmentation model’s performances have been compared with state-of-the-art segmentation techniques. It has observed in Table 5.2 that the model showed significant accuracy results with sensitivity, specificity, Recall, and AUC by comparing with other segmentation model. The proposed deformable cascaded W-net model provides a testing accuracy 95.68 with a AUC of 0.97.

Table 5.2 Blood Vessel Segmentation Model Comparison

Model	**Accuracy**	**Sensitivity**	**Recall**	**F1-Score**	**AUC**
Chen [35]	94.74	0.72	0.97	-	0.96
Liskowsk [36]	94.95	0.77	0.97	-	0.97
U-net [37]	95.31	0.75	0.98	0.81	0.97
Proposed Model	95.68	0.77	0.98	0.82	0.97

5.5.2 Skin Cancer Lesion Segmentation Results

Skin cancer lesion segmentation based on Kaggle dataset is considered here with the proposed segmentation technique using deformable cascaded W-net architecture. Figure 5.4 depicts the segmentation output of the Kaggle dataset using three different columns. The skin cancer input images represented in first column, the second column represented the ground truth, and the third column demonstrated the segmented image. It has been observed that the deformable cascaded W-net showed significant segmentation outcomes.



(a) Image (b) Ground Truth (c) Segmented Image

Figure 5.4 Segmentation Output of Skin Cancer Dataset

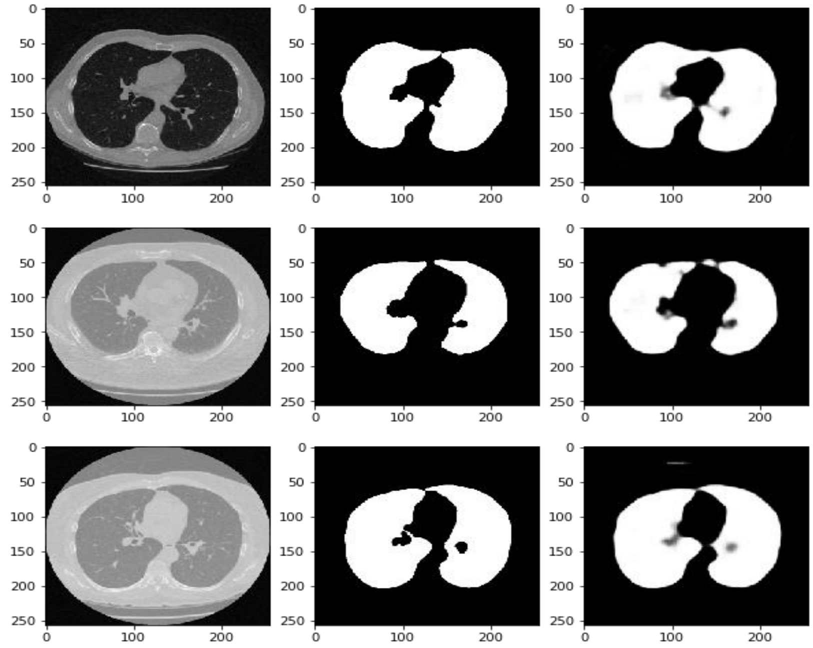
The proposed deformable cascaded W-net segmentation model’s performances have been compared with state-of-the-art segmentation techniques. It has observed in Table 5.3 that the model showed significant accuracy results with sensitivity, specificity, and Recall by comparing with other segmentation model. The proposed deformable cascaded W-net model provides a testing accuracy 99.24 with a F1-score of 0.98.

Table 5.3 Skin Cancer Segmentation Model Comparison

Model	**Accuracy**	**Sensitivity**	**Recall**	**F1-Score**
TDLS [45]	98.30	0.91	0.99	-
U-net [46]	93.14	0.95	0.93	0.86
Jafari [47]	98.50	0.95	0.98	-
Proposed Model	99.24	0.98	0.99	0.98

5.5.3 Lung Segmentation Results

Lung segmentation based on Kaggle Data Science Bowl 2017 dataset is considered here with the proposed segmentation technique using deformable cascaded W-net architecture. Figure 5.5 depicts the segmentation output of Kaggle dataset using three different columns. The lung input images represented in first column, the second column represented the ground truth, and the third column demonstrated the segmented image. It has been observed that the deformable cascaded W-net showed significant segmentation outcomes.



(a) Image (b) Ground Truth (c) Segmented Image

Figure 5.5 Segmentation Output of Lung Dataset

The proposed deformable cascaded W-net segmentation model's performances have been compared with state-of-the-art segmentation techniques. It has observed in Table 5.4 that the model showed significant accuracy results with sensitivity, specificity, Recall, and AUC by comparing with other segmentation model. The proposed deformable cascaded W-net model provides a testing accuracy 98.76 with F1-Score of 0.98.

Table 5.4 Lung Segmentation Model Comparison

Model	**Accuracy**	**Sensitivity**	**Recall**	**F1-Score**
U-net [48]	98.28	0.93	0.98	0.96
Hieu [49]	93.00	-	-	0.87
Proposed Model	98.76	0.98	0.99	0.98

5.6 Summary

We proposed an extension of the U-Net architecture using deformable cascaded W-net. The deformable convolutional network with cascaded W-net architecture is considered to form architecture of the model. These models were evaluated using three different applications in the field of medical imaging including retina blood vessel segmentation, skin cancer lesion segmentation, and lung segmentation. The experimental results showed that these proposed models ensure better performance with the testing phase by comparing with the state-of-the-art methods.

6 ENSEMBLE METHOD FOR PREDICTION OF CANCER METASTASIS USING MICROARRY GENE EXPRESSION DATA THROUGH MACHINE LEARNING

6.1 Introduction

Cancer causes highest number of deaths in the world now-a-days. It has been identified in many analysis that 13% of overall deaths caused by cancer. Cancer study has attained a significant success over last couple of years as scientists are currently concentrating on detection of cancer based on image, signals, and gene expression dataset. Moreover, automated identification and diagnosis can ensure better treatment for the patient as it reduce the processing time of cancer identification. The microarray technology has been a key interest of scientist and practitioner in bioinformatics related area [54]. The assessment and prediction of microarray data have immense influence in medical disease diagnosis last couple of decades as the microarray contain biomolecular evidence from tissue and cell samples correlated to various categories of cancer. Moreover, the microarray contains expressions of thousands of genes that helps to distinguish different tumor types based on revealing the patterns of normal and diseased cell genes.

The genomic scale pattern permits the scientist to supervise the genes with a view to classify and diagnosis cancer. To predict and classify cancer from gene expression dataset, the statistical and machine learning methodologies are crucial to effectively identify cancer outcome. Automatic identification and diagnosis using machine learning algorithms makes the classification, prediction, detection, and segmentation task easier and that helps to get desired output. Moreover, a precise and efficient disease estimation can decrease the threat of the cancer and increase the chances to get healed. Practitioners can rely on the dataset to take necessary action as the models can effectively identify the cancer.

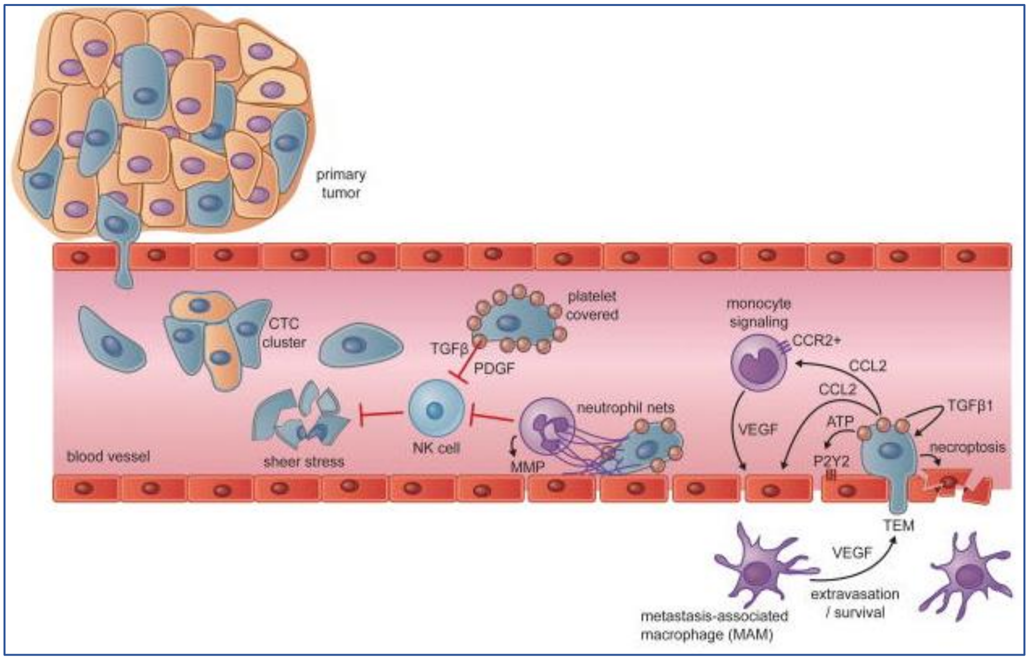


Figure 6.1 Metastasis Spreads from Organs to Lymph Nodes [55]

The transmission of metastasis from tissue to lymph nodes has demonstrated vividly in Figure 6.1 to detect tumor using genes and metastasis phases. To ensure better accuracy results from gene expression dataset, it is mandatory to analyze all the genes in terms of identify cancer. However, the large number of features in cancer gene expression dataset make the task challenging to consider most of the genes as the dataset is imbalanced with higher dimensional. To overcome the high dimensionality of the dataset data mining and machine learning method would be a good choice. Moreover, feature selection and feature extraction technique become the blessing for the practitioner to effectively identify cancer from the reduce number of genes. Several feature selection and feature extraction technique has been proposed in different literature for microarray gene expression data analysis.

The reduced feature can be considered to identify cancer based on machine learning algorithm and that surely decrease the processing time for doctors to identify cancer biomarkers. However, feature reduction is not an easy task and the research shown that none of the feature selection and feature extraction technique is sufficient for microarray dataset to identify cancer. Moreover, considering a limited number of dimensionality reduction approach could not represent the dataset appropriately. Therefore, all the feature selection and feature extraction technique need to be applied on benchmark dataset to identify cancer with better accuracy results.

In this research, we focused on colorectal cancer and breast cancer microarray dataset that consists of huge number of gene features. The microarray dataset is normalized by considering a range of value before fed into the machine learning model. Numerous machine learning techniques like logistic regression (LR), support vector machines (SVM), perceptron, gaussian naïve bias, k-nearest neighbor (KNN), random forest (RF), and artificial neural network (ANN) models are considered to detect and predict cancer from the colorectal and breast microarray dataset. The dataset split into training and testing phase with different set like (70-30) %, (60-40) %, and (80-20) % with 10-fold cross validation process and finally fed into the machine learning model to get accuracy results [56].

The major objective of this investigation is to detect the substantial features from the gene expression dataset to discover the baseline classifier with better accuracy result in terms of identifying cancer. The classifiers are compared with the prediction rate as well as other metrics such as F1 score, precision, and recall. The essential breakthroughs of this effort remain as follows:

- Numerous machine learning techniques like logistic regression (LR), support vector machines (SVM), perceptron, gaussian naïve bias, k-nearest neighbor (KNN), random forest (RF), and artificial neural network (ANN) models are considered in a Python Jupyter Notebook environment.
- Several feature extraction techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), t-distributed Stochastic Neighbor Embedding (t-SNE),

Independent Component Analysis (ICA), and Multidimensional Scaling (MDS) are applied both for colorectal cancer and breast cancer microarray dataset to reduce number of features. The classification accuracy is compared with and without feature extraction technique and the reduced features showed significant performances.

- A few feature selection techniques like Recursive Feature Elimination (RFE), Randomized Logistic Regression (RLR), Gradient Boosting Machines (GBM), Adaboost (Ada), Deep Neural Network (DNN), Minimum Redundancy Maximum Relevance (mRMR), Correlation Feature Selection (CFS), Hilbert-Schmidt Independence Criterion Lasso (HSIC-Lasso), and Relief-F applied both for colorectal cancer and breast cancer microarray dataset to select significant features. The classification accuracy of different machine learning model is compared with and without feature selection technique and the subset of features showed significant performances.
- Clustering techniques like K-means, Expectation Maximization (EM), Farthest Fast (FF), and Density Based (DB) clustering are applied for both the colorectal and breast cancer microarray dataset and compared the accuracy results.
- Ensemble machine learning method is applied and compared the performance with the individual machine learning model. It has been observed that ensemble classification method showed better performances.

The remainder of the chapter is organized as follows. In Section 5.2, the problem statement regarding this research has been considered. Section 5.3 investigated the microarray dataset used in this cancer classification and prediction. Section 5.3 describes about the methodology used for cancer detection and feature elimination with two microarray datasets. The comparison of different machine learning algorithm with and without feature selection and feature extraction technique is performed in section 5.4. Moreover, the accuracy, precision, and recall are measured. Section 5.5 concludes the paper with a few lines of future work.

6.2 Problem Statement

Cancer has spreading at an alarming rate recent day and occurring different parts of the body like brain, breast, lung, skin, etc. In each year more than 1 million women have breast biopsies in the United State, and over a few thousands experience colorectal cancer. Manual identification of cancer from image, genes, and histology images takes huge time to process the dataset. Automated detection of cancer from microarray gene expression dataset can reduce the processing time as well as give better detection rate.

However, classification of microarray dataset from huge gene expression can experience certain issues, like a) less number of samples exist for microarray gene expression data, b) huge number of features observed for particular sample and some of the features are totally irrelevant, therefore, feature selection and feature extraction technique needed to reduce the number of features, c) due to investigational difficulty there might be noise in the dataset, d) efficient classification and feature extraction technique needed as the dataset is computationally complex.

Several statistical and machine learning algorithms has been considered in recent study, but the accuracy and efficiency of the model is quite unsatisfactory. Moreover, feature selection and feature extraction technique needed to reduce the gene set with a view to get better accuracy results. Deficiency of contrast among the machine learning, feature extraction, and feature selection techniques to discover an improved framework for classification, clustering, and evaluation of microarray gene expression. Thus, dimensionality reduction of genes as well as feature selection approaches are crucial to enhance the accuracy and velocity of prognostication methods.

6.3 Methodology

In this section, a synopsis of applied classification algorithms is presented for microarray gene expression dataset for classification of cancer with and without feature selection techniques.

6.3.1 Dataset

Microarray is employed to characterize the representation of genes by utilizing numerous microscopic assessments to discover gene expression founded on light position in a sequence. In this study, two microarray cancer datasets are used, i.e. colorectal cancer and breast cancer dataset. The dataset is taken from an online Geo repository named NCBI Gene Expression Omnibus repository. Colorectal cancer dataset consists of 120 different patient information based on metastasis and non-metastasis classes with 16385 gene expression. Moreover, breast cancer dataset consists of 97 different patient information based on metastasis and non-metastasis classes with 24481 gene expression

In pre-processing stage, the microarray gene expression datasets need to be normalized by considering a range in between 0 to 10 for all the gene expression features. To efficiently process the dataset, the normalized data need to be transposed rather than fed into the machine learning algorithm. The datasets accumulated values in numerical way, organized by rows and columns. The patient number represented in a column and the gene numbers represented in a row. Table 6.1 lists the description of datasets, used in this paper, in terms of gene number, patient number, as well as the reference.

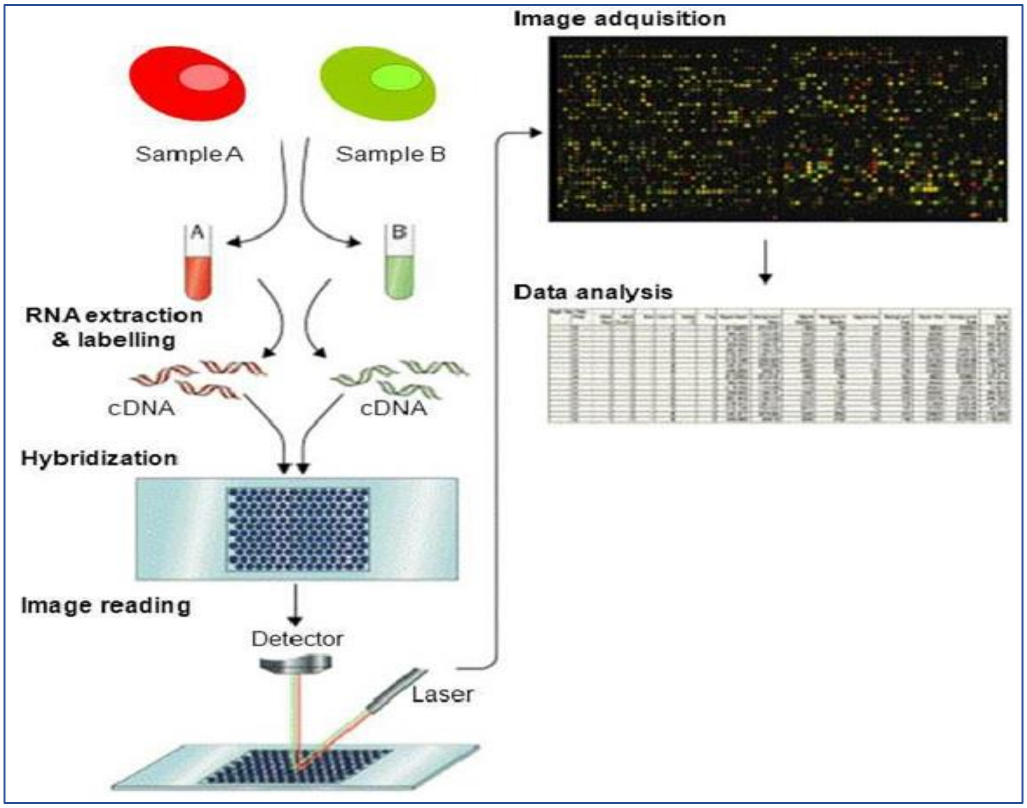
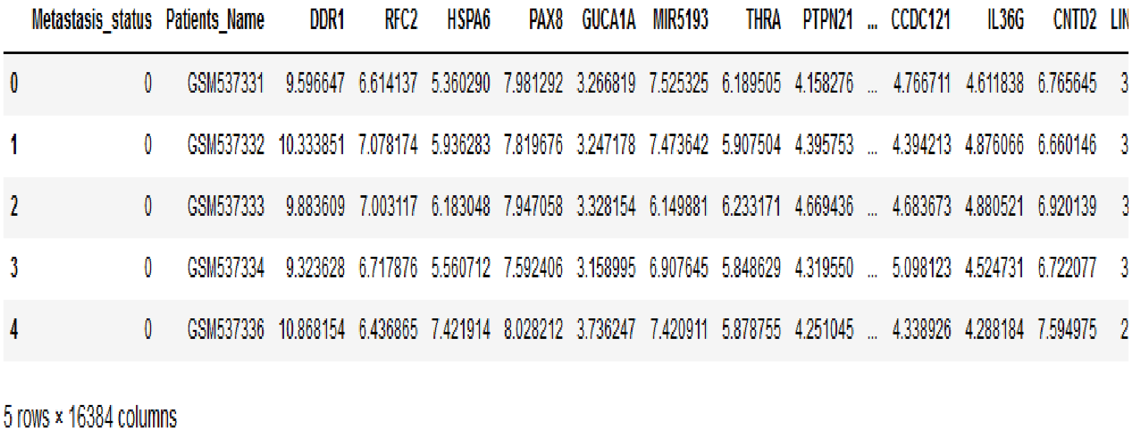


Figure 6.2 Microarray Data Processing [57].

In Figure 6.2, it has been observed that the microarray data collection process with several stages from the sample collection to microarray dataset generation along with normalization and pre-processing.

Table 6.1 Microarray Colorectal Cancer Dataset with 16385 Features and Limited Samples



	Metastasis_status	Patients_Name	DDR1	RFC2	HSPA6	PAX8	GUCA1A	MIR5193	THRA	PTPN21	...	CCDC121	IL36G	CNTD2	LIN
0	0	GSM537331	9.596647	6.614137	5.360290	7.981292	3.266819	7.525325	6.189505	4.158276	...	4.766711	4.611838	6.765645	3
1	0	GSM537332	10.333851	7.078174	5.936283	7.819676	3.247178	7.473642	5.907504	4.395753	...	4.394213	4.876066	6.660146	3
2	0	GSM537333	9.883609	7.003117	6.183048	7.947058	3.328154	6.149881	6.233171	4.669436	...	4.683673	4.880521	6.920139	3
3	0	GSM537334	9.323628	6.717876	5.560712	7.592406	3.158995	6.907645	5.848629	4.319550	...	5.098123	4.524731	6.722077	3
4	0	GSM537336	10.868154	6.436865	7.421914	8.028212	3.736247	7.420911	5.878755	4.251045	...	4.338926	4.288184	7.594975	2

5 rows × 16384 columns

In Table 6.1, the colorectal cancer dataset used in the research has been represented for 120 sample patients with 16385 gene expression features. Table 6.2 represented the two datasets used in this research with number of instances, total number of genes, and the classes for prediction.

Table 6.2 Microarray Dataset with Instances and Classes

Dataset	Total Genes	Instances	Classes
Colorectal Cancer	16384	120	2
Breast Cancer	24481	97	2

6.3.2 Proposed Method

In this study, we introduce two frameworks for microarray gene expression data classification. Microarray datasets with huge number of genes need to be reduced to get the better accuracy results as there are several irrelevant genes are in a microarray dataset. Feature selection and dimensionality reduction techniques are overly used for reducing feature. The block diagram in Figure 6.3 represented the dimensionality reduction techniques for feature extraction from the gene expression data with a view to identify the significant genes for cancer classification.

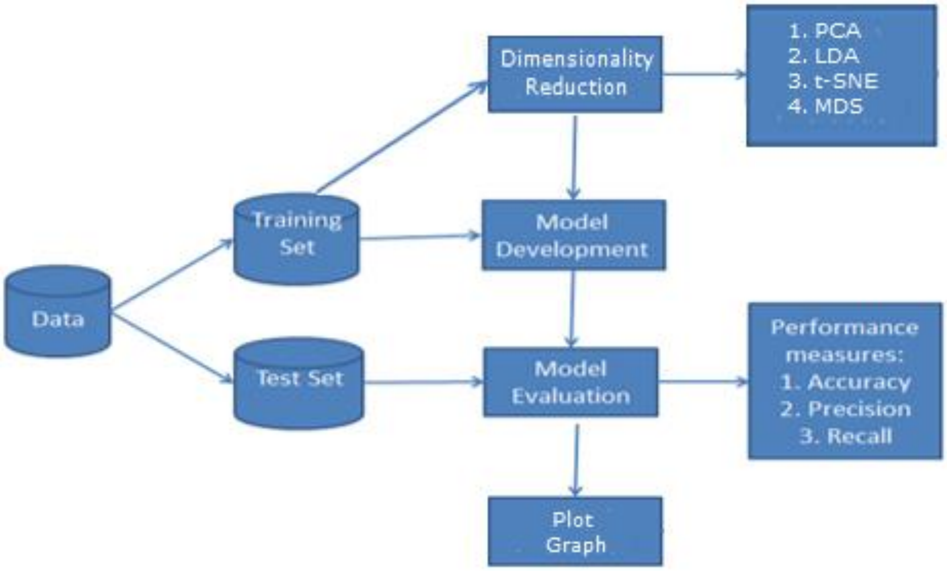


Figure 6.3 Block Diagram of Cancer Classification with Dimensionality Reduction.

The diagram applied for the classification of colorectal and breast cancer dataset with all the features based on popular machine learning algorithm like SVM, KNN, RF, ANN, and perceptron. The same algorithms are applied after reducing the number of genes by applying PCA, LDA, t-SNE, and MDS, etc [58]. The block diagram in Figure 6.4 represented the feature selection techniques to reduce features from the gene expression data with a view to identify the significant genes for cancer classification.

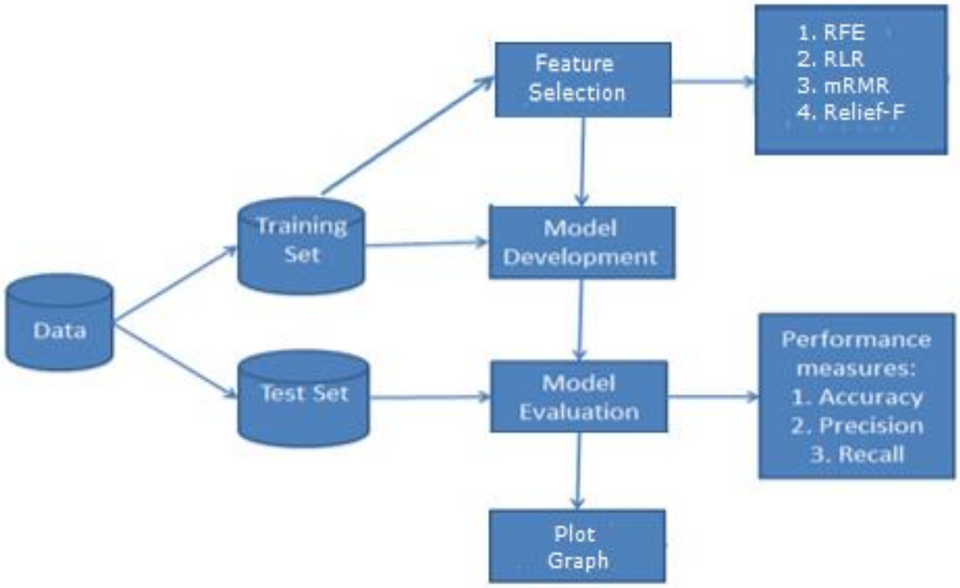


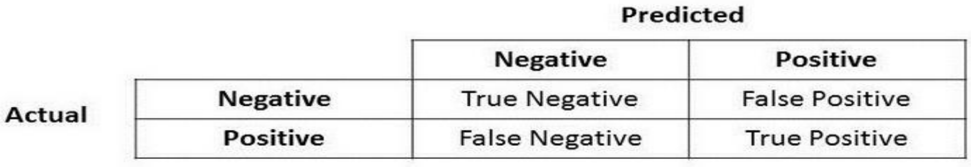
Figure 6.4 Block Diagram of Cancer Classification with Feature Selection

The diagram applied for the classification of colorectal and breast cancer dataset with all the features based on popular machine learning algorithm like SVM, KNN, RF, ANN, and perceptron. The same algorithms are applied after reducing the number of genes by applying RFE, RLR, CFS, mRMR, Relief-F, etc.

5.4 Experimental Results

The gene expression dataset employed in this study both for colorectal and breast cancer is split into training and testing set along with class labels like metastasis and non-metastasis and gene expression features. The dataset split here in several ways, we initially considered 70% of the dataset as training and the 30% of the dataset as testing set. Secondly, 60% of the dataset as training set and the 40% of the dataset as testing set. Finally, 50% of the dataset is considered as training set and the 50% of the dataset as testing set. The confusion matrix for the actual and predicted results with positive and negative ratio with all the parts has shown in Table 6.3.

Table 6.3 Confusion Matrix



		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Accuracy, precision, recall and F-1 score are calculated by using the formula mentioned below based on true positive, false positive ratio in order to evaluate the overall predicted results.

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \tag{6.1}$$

$$Specificity\ (SP) = \frac{TN}{TN+FP} \tag{6.2}$$

$$Sensitivity\ (SE) = \frac{TP}{TP+FN} \tag{6.3}$$

$$Precision\ (PR) = \frac{TP}{TP+FP} \tag{6.4}$$

$$F - score\ (Fs) = 2 \times \frac{SE \times PR}{SE+PR} \tag{6.5}$$

Several classification techniques applied for predicting colorectal and breast cancer from all the 16385 and 24481 features with 120 and 97 samples for detecting metastasis. Machine learning models such as LR, SVM, naive bias, KNN, RF, etc. are applied for both dataset with all features. Moreover, 10-fold cross validation approach is considered in a rotational way to classify cancer. All the classifier results are compared along with prediction rate and several metrics such as F1 score, precision, recall is computed to identify the baseline classifier.

In Table 6.4, the accuracy comparison of the machine learning models is represented with other metrics. From table 5.3, SVM showed accuracy result of 77.9% that is better than other classification models. LR and KNN showed next better accuracy results of 74.9% and 73.4% respectively. The precision, recall, F1-score for all the machine learning models observed in the following table.

Table 6.4 Machine Learning Models with Dimensionality Reduction for Colorectal Cancer

Machine Learning	Avg. Test Accuracy	Precision	Recall	F-1 Score
SVM	77.90	0.77	1.0	0.85
SVM Reduced Features (80)	76.23	0.76	1.0	0.87
LR	74.64	0.75	0.75	0.75
LR Reduced Features (80)	56.50	0.56	0.56	0.56
KNN	73.40	0.73	1.0	0.84
KNN Reduced Features (80)	71.87	0.72	1.0	0.84
Perceptron	76.45	0.74	1.0	0.85
Perceptron Reduced Features (80)	54.10	0.54	0.54	0.54
RF	68.40	0.71	0.71	0.71
RF Reduced Features (80)	75.10	0.8	0.8	0.8
Gaussian Naïve Bias (GNB)	73.59	0.68	0.68	0.68
GNB Reduced Features (80)	72.35	0.65	0.65	0.65

The block diagram mentioned in Figure 6.3 is considered machine learning technique with all the feature and the reduced number of features. For colorectal cancer 16385 is considered to evaluate the predicted results of the classifier applied for comparison. The same machine learning techniques applied after reducing the number of genes to 80 with dimensionality reduction technique and the identified gene sets are the most significant genes for colorectal cancer dataset.

The pictorial diagram of all the machine learning models comparison with accuracy measurement for colorectal cancer is mentioned in Figure 6.5.

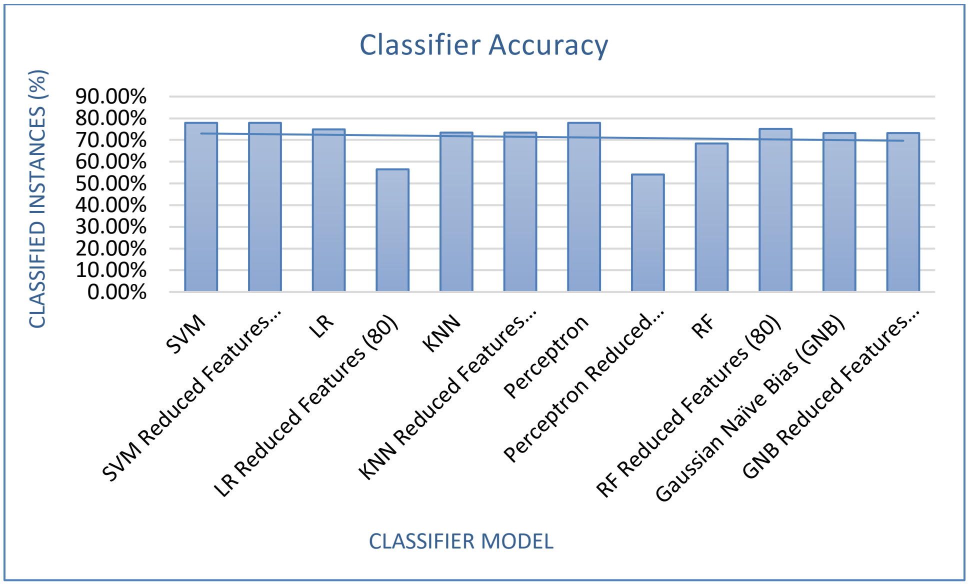


Figure 6.5 Colorectal Cancer Accuracy Comparison

Dimensionality reduction convert the gene set to small number that can improve classification result to identify cancer efficiently and perfectly with less managing time. Classification model's accuracy is cross checked with and without features.

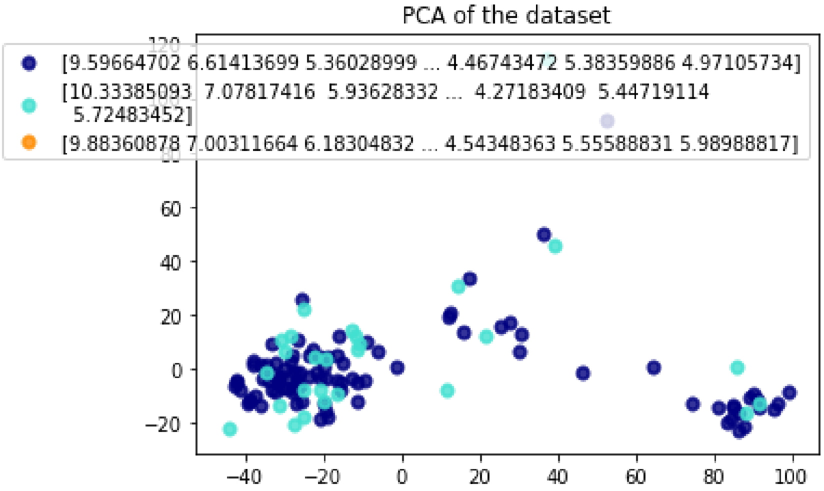


Figure 6.6 PCA to Reduce the Number of Dimensions Colorectal Cancer

In PCA the number of features reduced from 16385 to 80 and the accuracy, precision, recall and F-1 score of all the models with reduced features is closely related to the models with all the features. The PCA plot for colorectal cancer is mentioned in Figure 6.6. In ICA, the number of features reduced to 12 from 16385 depicted in Figure 5.7. Therefore, 12 genes are the significant genes for cancer occurrence. Moreover, t-SNE, factor analysis, and MDS is applied to reduce the number of features depicted in Figure 5.8, 5.9, and 5.10, respectively.

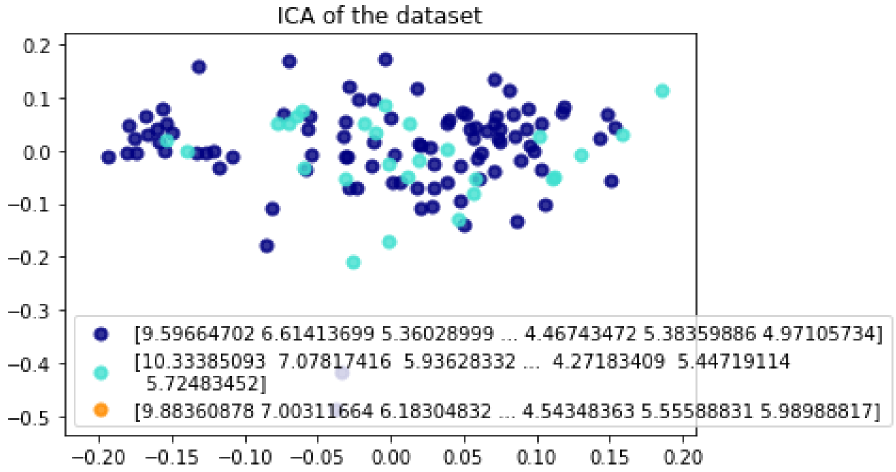


Figure 6.7 ICA to Reduce the Number of Dimensions Colorectal Cancer

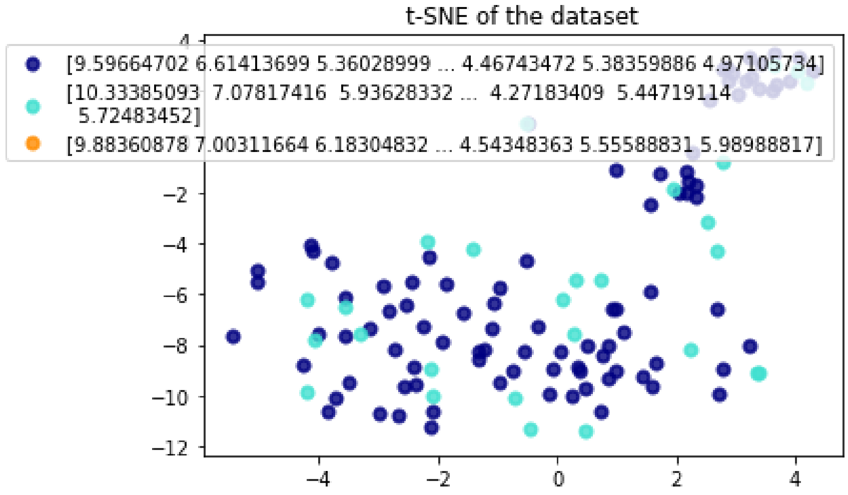


Figure 6.8 t-SNE to Reduce the Number of Dimensions Colorectal Cancer

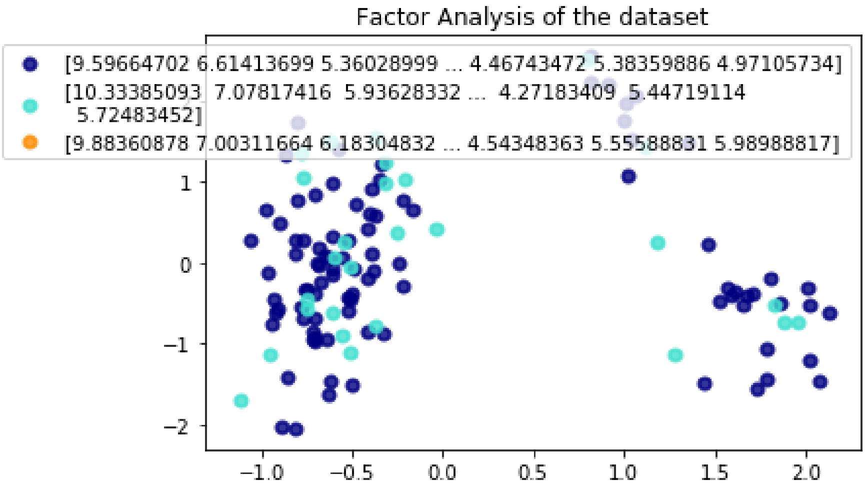


Figure 6.9 Factor Analysis to Reduce the Number of Dimensions Colorectal Cancer

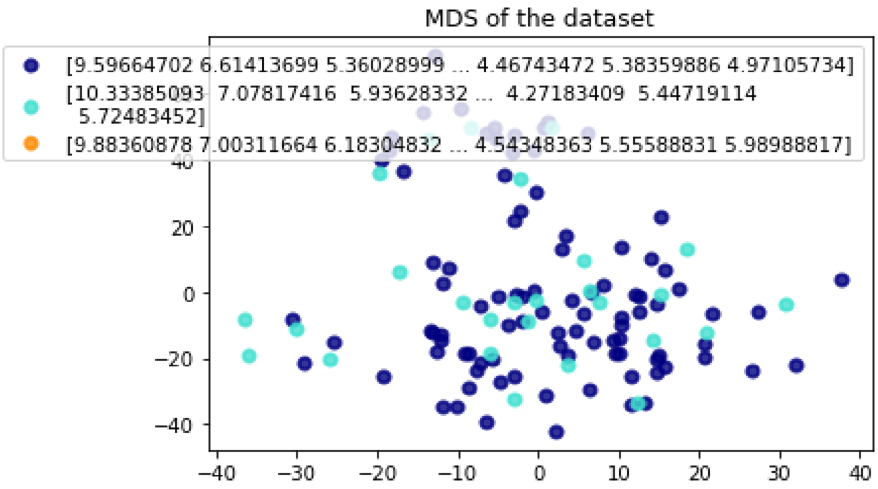


Figure 6.10 MDS to Reduce the Number of Dimensions Colorectal Cancer

In Table 6.5, the accuracy comparison of the machine learning models for breast cancer classification is represented with other metrics. RF showed accuracy result of 72.3% that is better than other classification models [59]. The precision, recall, F1-score for all the machine learning models observed in the following table. The block diagram mentioned in Figure 6.5 is considered machine learning technique with all the feature and the reduced number of features. For breast cancer 24811 is considered to evaluate the predicted results of the classifier applied for comparison.

Table 6.5 Machine Learning Models with Dimensionality Reduction for Breast Cancer

Machine Learning	Avg. Test Accuracy	Precision	Recall	F-1 Score
SVM	63.7	0.62	0.90	0.79
SVM Reduced Features (115)	61.2	0.59	0.85	0.81
LR	64.5	0.68	0.74	0.71
LR Reduced Features (115)	59.1	0.59	0.65	0.61
KNN	58.4	0.63	0.79	0.69
KNN Reduced Features (115)	55.8	0.59	0.77	0.64
Perceptron	67.7	0.67	0.85	0.75
Perceptron Reduced Features (115)	64.1	0.62	0.70	0.62
RF	72.3	0.75	0.95	0.85
RF Reduced Features (115)	69.9	0.80	0.98	0.80
Gaussian Naïve Bias (GNB)	63.8	0.65	0.70	0.65
GNB Reduced Features (115)	61.6	0.63	0.69	0.63

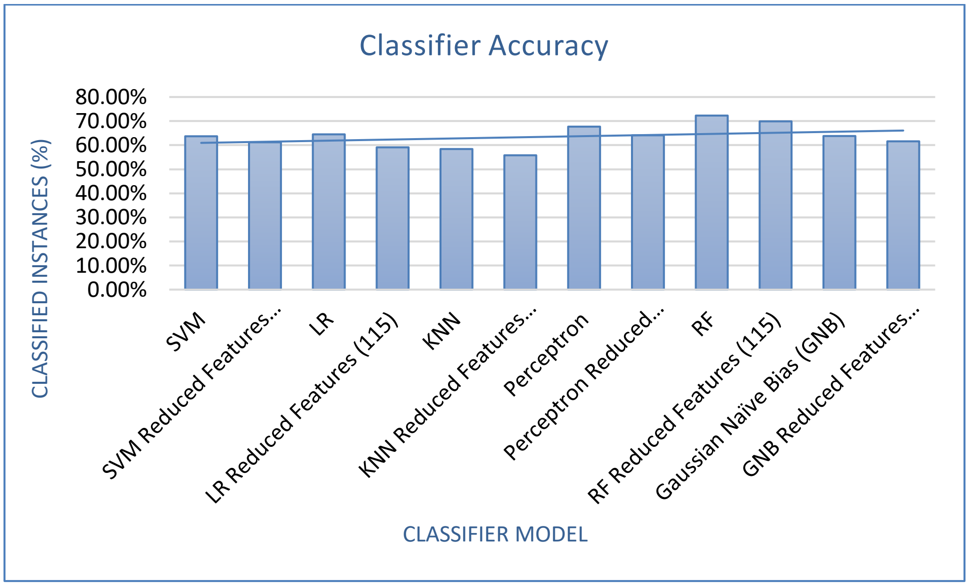


Figure 6.11 Breast Cancer Accuracy Comparison

The same machine learning techniques applied after reducing the number of genes to 120 with dimensionality reduction technique and the identified gene sets are the most significant genes for colorectal cancer dataset. The pictorial diagram of all the machine learning models comparison with accuracy measurement for colorectal cancer is mentioned in Figure 6.11.

In Table 6.6, the feature selection approach like RFE, Relied-F, mRMR, DNN, and CFS is considered both for colorectal and breast cancer. The number of genes for both cancer dataset is reduced using the feature selection approach.

Table 6.6 Number of Selected Genes using Feature Selection

Dataset	Total Genes	RFE	Relief-F	mRMR	DNN	CFS
Colorectal Cancer	16384	95	110	850	700	110
Breast Cancer	24481	120	131	1224	1000	130

Table 6.7 Machine Learning Models with Feature Selection for Colorectal Cancer

Method	SVM	KNN	DT	LR	RF	Ada	GBM
All Features	77.70	73.40	61.52	74.64	68.40	78.29	75.74
RFE	95.23	87.44	79.65	96.17	92.73	90.87	81.48
DNN	94.72	86.31	77.65	93.96	89.48	85.51	77.83
RLR	96.59	83.90	74.71	91.84	88.14	87..07	75.40
mRMR	92.30	81.26	73.35	87.44	87.12	86.46	76.21
HSIC-LASSO	88.08	75.23	69.78	89.43	84.47	89.14	78.05
Relief-F	94.91	79.62	77.32	93.19	92.42	93.59	80.43
CFS	93.43	85.65	74.48	92.95	87.47	88.67	80.78

In Table 6.7, the accuracy comparison of the machine learning models like SVM, KNN, DT, LR, Ada, etc. is represented for colorectal cancer and RLR with SVM showed accuracy result of 96.59% that is better than other classification models. RFE with LR showed next better accuracy results of 96.17%. Figure 6.12 showed the accuracy results of the machine leaning techniques with several feature selection method.

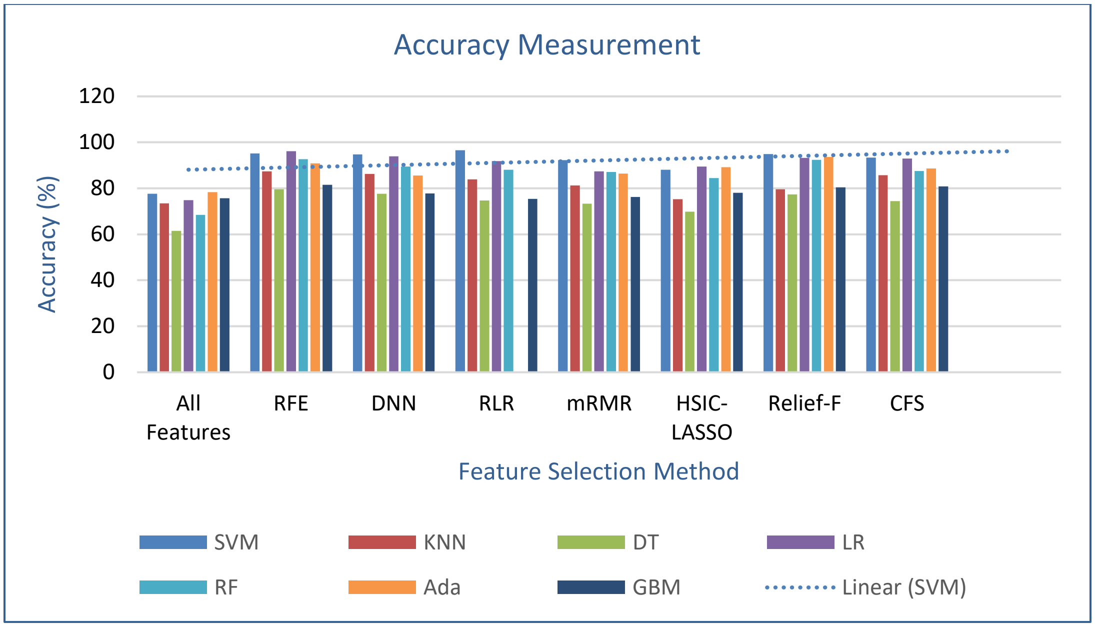


Figure 6.12 Feature Selection Method Accuracy Comparison for Colorectal Cancer

MLP with different number of layers and neurons for colorectal cancer is considered in Table 6.8. It is observed that MLP with 30 layers showed accuracy result of 86.25%

Table 6.8 MLP with Number of Layers for Colorectal Cancer

	15	**30**	**60**	**15-30**	**30-60**
MLP	77.90	86.25	80.48	84.34	83.71

In Table 6.9, the accuracy comparison of the machine learning models like SVM, KNN, DT, LR, Ada, etc. is represented for breast cancer and RFE with SVM showed accuracy result of

90.75% that is better than other classification models. RFE with LR showed next better accuracy results of 89.62%. Figure 6.13 showed the accuracy comparison of the machine leaning techniques with several feature selection method.

Table 6.9 Machine Learning Models with Feature Selection for Breast Cancer

Method	**SVM**	**KNN**	**DT**	**LR**	**RF**	**Ada**	**GBM**
All Features	63.70	58.40	61.59	64.50	72.30	70.05	63.94
RFE	90.75	81.05	67.70	89.62	87.16	82.55	72.39
DNN	88.67	80.33	72.52	89.35	85.24	79.71	70.53
RLR	89.25	80.19	71.45	87.63	85.27	84..91	70.08
mRMR	87.12	77.93	69.49	84.13	81.56	83.68	69.57
HSIC-LASSO	86.46	72.57	65.25	85.43	80.12	85.29	73.28
Relief-F	89.72	75.45	70.21	88.92	86.14	88.35	74.95
CFS	87.69	79.38	69.94	89.54	85.31	85.26	75.27

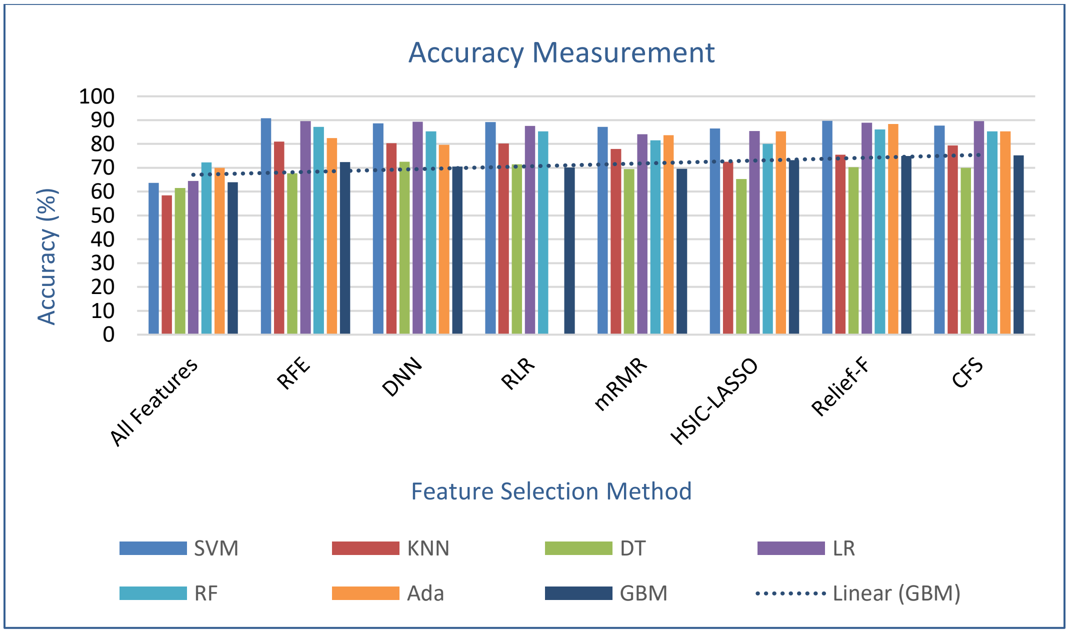


Figure 6.13 Feature Selection Method Accuracy Comparison for Breast Cancer

MLP with different number of layers and neurons for breast cancer is considered in Table 6.10. It is observed that MLP with 15-30 neurons showed accuracy result of 73.98%.

Table 6.10 MLP with Number of Layers for Breast Cancer

	15	**30**	**60**	**15-30**	**30-60**
MLP	67.23	75.35	60.97	73.38	71.62

In Table 6.11, the accuracy comparison of the several machine learning clustering models is represented, and EM showed better results for colorectal cancer and FF showed better results for breast cancer.

Table 6.3 Clustering Methods for all Genes with Accuracy Measurement

Dataset	**K-means**	**EM**	**FF**	**DB**
Colorectal Cancer	81.86	92.78	89.56	83.37
Breast Cancer	72.13	79.52	87.35	76.45

In Table 6.12, ensemble method is considered for both colorectal and breast cancer. It has been identified that AdaBoost when applying with other classification method showed significantly better results. From Table 6.12, AdaBoost-RF showed accuracy result of 98.72% and 97.89% for colorectal and breast cancer, whereas individual AdaBoost got accuracy result of 78.29% and 70.05% for colorectal and breast cancer. Figure 6.14 showed the accuracy comparison of the several ensemble techniques.

Table 6.4 Ensemble Method Accuracy Comparison

Method	**Colorectal**	**Breast**
AdaBoost	78.29	70.05
AdaBoost-RF	98.72	97.89
AdaBoost-SVM	96.67	95.35

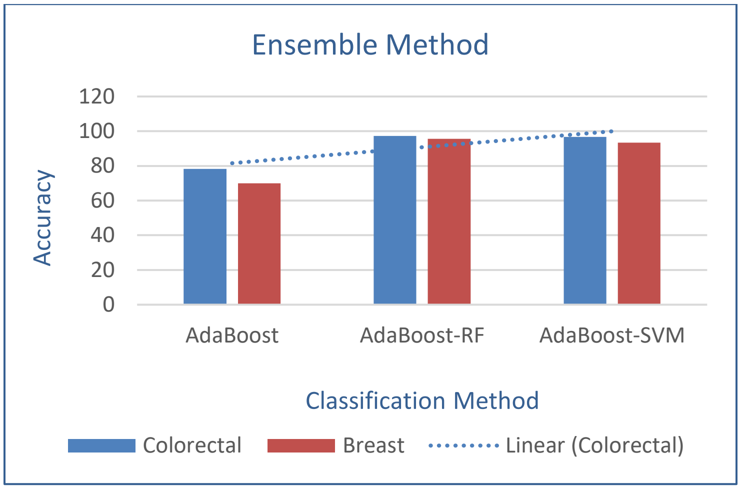


Figure 6.134 Ensemble Method Accuracy Comparison

In Table 6.13, the accuracy comparison of the proposed ensemble model for colorectal and breast cancer dataset with other state-of-the-art methods is represented. The proposed ensemble for colorectal cancer showed accuracy result of 98.72% that is better than other classification models. Moreover, breast cancer ensemble method showed accuracy result of 97.89 that is better than other state-of-the-art classification model.

Table 6.5 Ensemble Method Accuracy Comparison with Other Models

Methods	**Dataset**	**Total Genes**	**Instances**	**Classes**	**Accuracy (%)**
Turgut [60]	Breast	24481	97	2	88.82
Kavitha [61]	Breast	47294	97	2	96.44
Sergey [62]	Colorectal	2000	62	2	93.75
	Breast	24481	97	2	86.09
Abinash [63]	Colorectal	2000	62	2	86.00
	Breast	24481	97	2	97.61
Proposed Ensemble	Colorectal	16384	120	2	98.72
	Breast	24481	97	2	97.89

6.5 Summary

Statistical and machine learning models has fantastic influence in cancer detection from microarray gene expression dataset. The massive features in the dataset is normalized and employed to machine learning models to identify the baseline classifier for both colorectal and breast cancer. The raw dataset with all the gene features is considered and calculate the accuracy, precision, recall, etc. to predict cancer. The same dataset is used for the classification model with reduced number of features by applying feature extraction and feature selection techniques. The reduced features showed significant improvement in cancer identification with better accuracy results and less processing time. Feature extraction techniques like PCA, LDA, t-SNE, MDS, etc. are applied to reduce the number of features. Moreover, feature selection techniques like RFE, RLR, mRMR, Relief-F, etc. are applied for selecting the reduced gene set. Machine learning techniques like SVM, RF, DT, KNN, ANN are applied to identify the baseline classifier. Finally, an ensemble technique is applied to combine a few machine learning techniques and it has been observed that ensemble model showed better accuracy result.

7 INTEGRATIVE ANALYSIS OF HISTOLOGY, IMAGING, AND GENOMICS FOR CANCER BIOMARKER DETECTION

7.1 Introduction

Understanding the biological behavior for cancer prognosis has been a crucial issue over the years. Cancer spreading in an alarming rate and manual identification process is not sufficient for disease diagnosis. Automated identification of cancer lesion has significant impact on the performances of disease prognosis and reduce the processing time of the doctors for cancer detection. Therefore, computer aided system has been performed using machine learning and deep learning models [64-67]. Deep learning model like convolutional neural network with other feature extraction technique has been used for microscopic image classification, radiology image segmentation and classification, and genome expression analysis for retrieving useful information for disease prognosis.

As cancer occurred several parts of the body, an individual analysis of histology, imaging, or genomics is not enough to detect abnormality in the dataset for disease diagnosis. Moreover, individual analysis using machine learning and deep learning models with limited samples might not be beneficiary for identifying appropriate disease outcome as overfitting issue can arise [68]. An integrated analysis of histology, imaging, and genomics for cancer biomarker detection would be a great choice now-a-day. Moreover, histo-genomics analysis, radio-genomic analysis, and molecular-genomic analysis has been a key concern to ensure better diagnosis of the disease [69]. Multiple analysis based on multiple types of dataset can produce significant results for biomarker detection. Correlation analysis among radiology-image, histology-image, and genome sequence can generate better predictive results that might not be possible to achieve from an individual analysis of image, histology, or genome data. Research also shows that combination approaches of any of the above also represent better performances than individual analysis.

In this research, we focused on integrative approach for lung cancer analysis and detection as lung cancer is the prominent cause of cancer-related death. Recent lung cancer study experienced a new type of lung cancer named Non-small cell lung cancer (NSCLC) and around 80% of the lung cancer patients are diagnosed with this type [70][71]. In NSCLC, computed tomography (CT) scan images are enormously used for disease diagnosis. The correlation of CT scan image features with microarray gene expression analysis of NSCLC has significant impact for cancer biomarker detection. Research showed that the combination approach of image modality and gene expression sets achieve better accuracy results rather than considering the individual approach of image or genome analysis [72]. The objective of this research is to evaluate the correlation between CT radiomics features and gene expression sets in NSCLC.

This chapter is organized as follows: Section 7.2 describes the proposed method for lung cancer integrative approach. Section 7.3 describes the dataset. Section 7.4 presents the experimental results. Finally, the conclusion is drawn in Section 7.5.

7.2 Methodology

We consider a radio-genomics association study in this research by incorporating the lung cancer CT image features with microarray gene expression data in NSCLC. The correlation of the two different analysis has been considered to achieve significant results in terms of disease diagnosis. The research carried out in two different steps for image and genome expression analysis. Firstly, correlation map was produced based on the feature extracted from CT scan image and microarray gene expression data in NSCLC. Secondly, evaluation of the model using mathematical and statistical procedure to detect biological findings in the NSCLC dataset. The block diagram of proposed integrated framework for biomarker detection in terms of disease diagnosis has been depicted in Figure 7.1. The figure represented the association study of the radio-genomics and histo-genomics analysis with a view to predict cancer biomarker that could support disease diagnosis and remedial treatment of the patients.

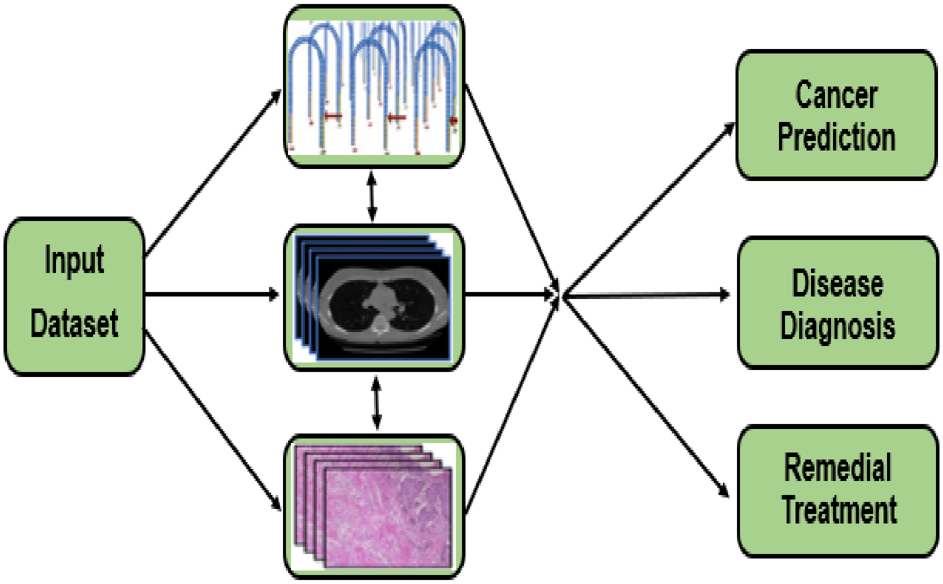


Figure 7.1 Proposed Integrated Framework for Biomarker Detection

The input CT scan images in NSCLC is going through a couple of steps like image augmentation, normalization, segmentation, feature extraction, etc. [73-76]. We deployed the CT images and applied the image pre-processing and analyzing the image in computer vision approach. Deep learning techniques like U-net, V-net, Y-net, W-net has shown significant outcome for cancer lesion detection. Moreover, Convolutional Neural Network (CNN) has great impact in segmentation purpose. We considered U-net architecture for separating the lung from other tissues in CT scan images. End-to-end and pixel-by-pixel processes are deployed to generate the lung patches with a view to separate lung [77]. Tumor area are detected, and nodule detection has been performed for lung cancer detection. Figure 7.2 depicted the lung segmentation task from input image to segmented output in NSCLC.

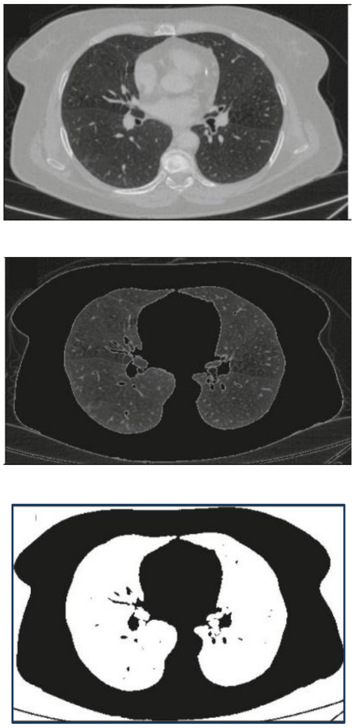


Figure 7.2 Non-Small Cell Lung Cancer Segmentation

Features are extracted from the lung CT scan images. The features are of textural features, HOG (Histogram of Oriented Gradient) features, and wavelet features. Several filtering techniques applied on those features in terms of reducing the redundant features in NSCLC. The microarray gene expression dataset of NSCLC is normalized by considering a range of value before fed into the machine learning model. Dimensionality reduction procedure like PCA, LDA, t-SNE, multidimensional scaling etc. [78] is applied to reduce the number of genes in NSCLC. Numerous clustering technique like k-means, hierarchical clustering technique also applied to determine the number of clusters in gene expression set and it may lead to reduce the number of genes in NSCLC. We considered 180 cluster from the gene expression sets and made relationship among the clusters based on the metagenes. In this research, 130 useful metagenes is generated and we analyzed those genes and compared with performance measurement techniques. Several gene specific tests like p-test and t-test is considered and allocate grades to each gene with a view to identify metagenes with CT scan radiomic features. The association study of CT scan image feature on top of microarray gene expression analysis showed significant outcome for cancer detection in NSCLC.

7.3 Dataset

The dataset used in this research contains CT scan images and microarray gene expression dataset for NSCLC. The publicly available dataset is collected from The Cancer Imaging Archive (TCIA) of 89 samples of NSCLC. More specifically, NSCLC dataset of radiomics and genomics collected from Gene Expression Omnibus (GEO). The publicly available dataset of GEO Series Accession Number is GSE58661. From the dataset, separate analysis has been performed for radiomics analysis and genomics analysis using traditional machine learning and deep learning models. Finally, a correlation of the two processes has been considered for better analysis of the datasets in terms of achieving significant results in disease diagnosis.

7.4 Experimental Results

The proposed model is considered for integrative analysis of CT scan and microarray data for non-small cell lung cancer detection. Technically, Python programming language with Keras and TensorFlow is used for the classification purpose. The hardware and software system specification are mentioned in Table 7.1.

Table 7.1 Hardware and Software Description

Hardware	**Software**
Processor: i7-6000, 2.80 gigahertz	OS: 64-bit Windows 10
Primary Memory: 16 gigabytes RAM	API: Keras [50]
GPU: NVIDIA GeForce GTX 770	Backend: TensorFlow [51]
Storage: Solid State, 250 gigabytes	Language: Python 3.7.3 [52]

We identified 90 textural features, 200 HOG features, and 700 wavelet features from the radiomics CT scan images of NSCLC. The microarray gene expression dataset consists of 60607 number of genes that reduced to 180 most significant genes and finally identified 130 metagenes. The radiomic features are considered with the gene expression sets to form a correlation map based on statistical procedure and identified the pairwise significant feature from the NSCLC dataset. Table 7.2 describes the accuracy comparison of the metagenes based on CT scan feature and the microarray gene expression analysis. It has been identified from the table that metagenes 12 showed significant result by comparing with other metagenes. Figure 7.3 depicted the representation of metagenes in a plot diagram.

Table 7.2 Accuracy Comparison of the Metagenes based on CT Scan Features

Methodology	Accuracy
Metagenes 12	90.12%
Metagenes 18	87.66%
Metagenes 25	89.32%
Metagenes 32	81.68%
Metagenes 53	82.52%
Metagenes 71	78.56%
Metagenes 79	86.95%
Metagenes 84	85.65%
Metagenes 97	77.74%
Metagenes 105	86.81%
Metagenes 119	82.23%

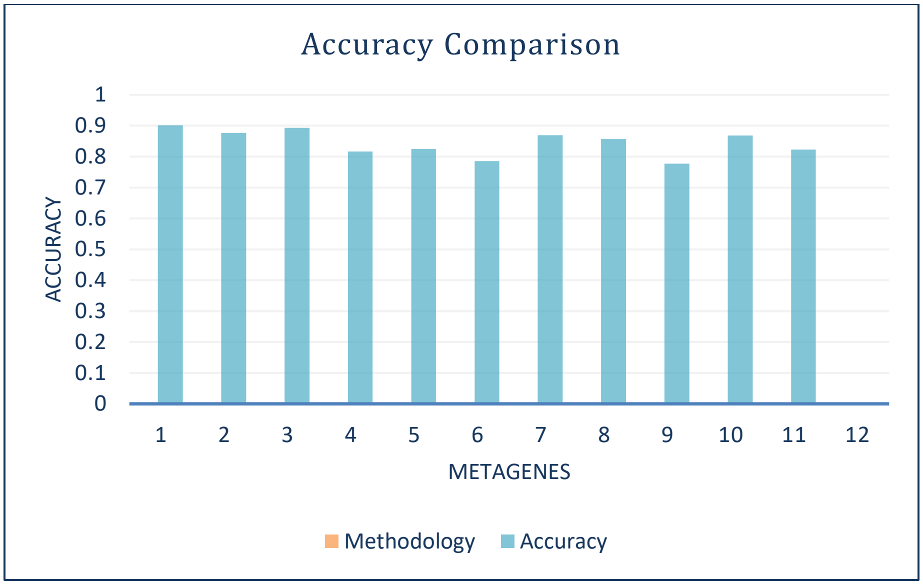


Figure 7.3 Plot Diagram of the Metagenes Performances

7.5 Summary

This research showed an association study of the CT scan images and microarray gene expression dataset of NSCLC. It has been found in many studies that the performance of combination approach in disease diagnosis exceeds the individual study of histology, imaging, or genomics. A radio-genomics analysis of NSCLC has been analyzed here. The image features are extracted from the CT scan images and microarray gene expression dataset is considered for identifying metagenes on top of the radiomic image. This research focused on radio-genomics approach, but histo-genomics, and molecular-genomics approach can be considered in terms of achieving better results in disease diagnosis that may eventually reduce the death rate of the patient.

8 INTERNET OF MEDICAL THINGS (IoMT) BASED BLOCKCHAIN FRAMEWORK FOR ELECTRONIC HEALTH RECORD MANAGEMENT

The health care industry is one of the biggest industries in the market. 11.2% of the total gross domestic product (GDP) was spent on healthcare in Germany, which was third highest in the world. Only countries that spent more were the United States at 16.9%, followed by 12.2% by Switzerland in 2018 [78][79]. 35.24% of the total population in Germany is overweight and 21.29% are obese, the population worldwide is aging, and physician and nursing shortages are anticipated [80]. Hospitals are still spending a sizable amount of resources into processing medical claims and administrative records [81].

Our world is becoming more interconnected and data-driven and so is healthcare. On the other hand, these changes are coming with the cost of high regulation, overhead cost and extra educational requirements for those who participate [81]. Therefore, changes in healthcare are slower. By 2030, the way healthcare is delivered is expected to change drastically due to increased access to data, additive manufacturing AI, and the wearable and implanted devices to monitor our health [82]. The evolution of e-health application and its ability to improve healthcare practices by electronic processes has had a positive influence on the healthcare sector [83].

The range of body functions that can be tracked using wearable in growing. Body functions such as blood pressure, hydration, oxygen level, brain activity (EEG), glucose, respiration, temperature, heart rate, and variability and movement can be tracked through the wearable technology available today [84]. "From assistance for Alzheimer’s patients to understanding complex knee injuries, wearable computing will transform how we understand pharmaceuticals, rehabilitation and preventative care.” [85]. The health-related wearables devices for monitoring and managing the health and well-being of individuals outside the medical institutions are growing to support healthcare [86].

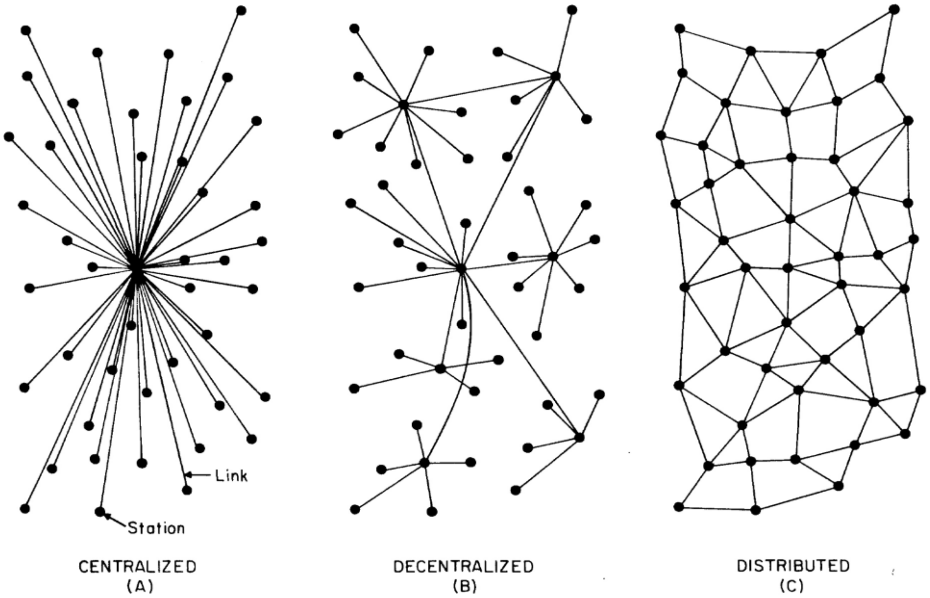


Figure 8.1 Blockchain Information system types [83]

Health-related IoT promises many benefits and is already paving the way for better personalized diagnosis, with healthcare evolving towards a system of predictive, preventive, and precision care [87]. This technology also enables real-time monitoring of patients, fitness and well-being monitoring, medication dispensation and data collection for research in the field of healthcare.

The primary goal of the thesis is to establish a decentralized access control system, which provides the opportunity to share digital resources and transfer defined access rights from one organization to another organization without having a central authority across organizational boundaries. In order to be able to answer this broad question, the following issues must be addressed:

- How to create a decentralized access control system?

In a decentralized and collaborative environment, centralized access control systems are not a suitable solution, so developing decentralized access control mechanism is crucial for decentralized systems.

- How can organizations grant or transfer access rights of micro services from one organization to another?

  Traditional access control systems are mostly centralized, so there is a single place to store access permissions. When the access control systems are centralized, granting and transferring access rights becomes a challenge. With decentralized systems in a collaborative environment, access control systems need to grant access rights in a decentralized architecture. Also, transferring the access rights is an essential feature for a decentralized access control system.

- How can consumer organization directly access the provider organization's micro services?

  When the resource provider organization grants access rights for its own microservices, the challenge is how the consumer organization can directly access the microservices without having additional third-party services.

8.1 Background and Motivations

The main objective of the blockchain research is to gain deeper understanding of the blockchain technology and its potential for healthcare systems. To come up with purposeful conclusion, a fundamental research question is formulated:

"What is blockchain technology and how does it fit into current healthcare system?" To answer the main research question, and conduct the study in structured demeanor, the research question is broken down into several sub-questions:

- What is blockchain technology and what are its implications?
- What are the application areas of blockchain?
- What is the current state of healthcare systems?
- What is current state of blockchain regarding healthcare industry?
- Which fields within healthcare sector can make use of blockchain technology? What solution concerning blockchain technology can be implemented in the healthcare sector for secured sharing of medical contents?
- What open issues exists and what are area for future research?

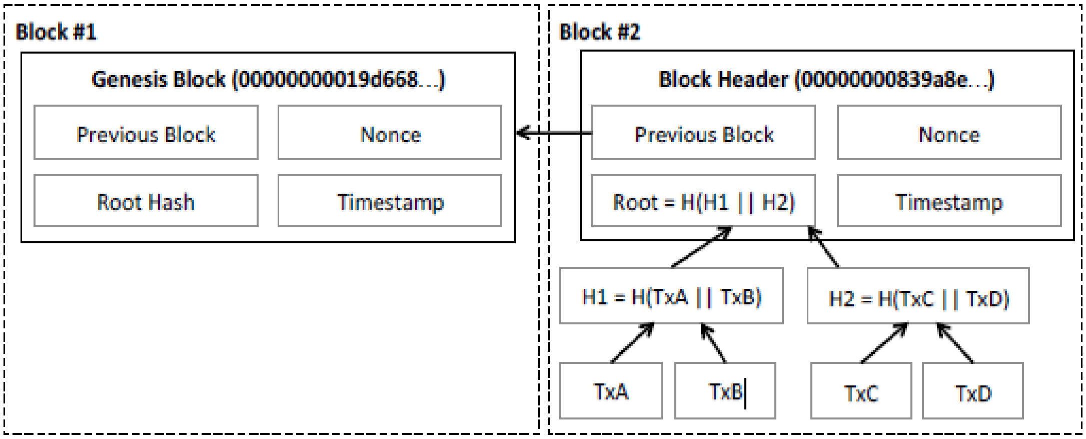


Figure 8.2 Blocks in a Blockchain [85]

This work is focused on the application of blockchain on a medical insurance storage system. In an ideal and basic medical insurance business, there are a group of hospitals, a patient and an insurance company. The insurance company can know a sum of the patient's specified spending records however the company cannot learn the details of the spending records.

The other hospitals can help the insurance company to process a patient's spending records without learning anything about the records. Specifically, we propose a privacy-preserving system which utilizes blockchain technology to store and process medical insurance data, which is characterized by the following features:

- Our system is fully decentralized, as there is no trusted third party to provide authentication.
- The data is verified and securely stored before being included in the blockchain, which provides high credibility to users.
- We adopt the (2, 3) threshold secret sharing protocol among participating hospitals, so that the final data can only be obtained if at least 2 hospitals respond to the insurance company's request.
- Every data stored on the blockchain can be efficiently verified by anyone, and even though the encryption is secure, the verification requires very little computational power. Results are obtained with the use of efficient homomorphic computation.

The remainder of the paper is organized as follows. The investigated problem is formalized in Section 8.2. Section 8.3 introduces challenges of blockchain. In Section 8.4, we investigate the working scenario of inference. The evaluation results are presented in Section 8.5. Section 8.6 concludes the paper.

8.2 Problem Statement

Healthcare is considered as one of the application areas of blockchain technology. But the technology adoption in the healthcare industry is relatively slow, and has been highlighted in the background paper on conceptual issues related to the health system, where the authors state that, "Pragmatic solutions already exist to address many of the greatest global health challenges, yet progress remains frustratingly slow because many health systems are constrained and cannot fully operationalize them.[56] Care coordination between patient and health care provider is increasing in complexity as various chronic conditions in the aging and growing population continue to rise. In many scenarios, the technology available in health care is not sufficient to capture all forms of care being catered. This is mainly due to use of old age technology to transfer information between relevant parties. Health care providers still use legacy systems, and paper-based medical records to retrieve and share medical data. Health care providers are still investing ample amount of resources into processing medical claims and administrative records when most of this can be eradicated using technologies such as Blockchain. Also, when it comes to patient-doctor interaction in Germany, paper-based prescription is still persistent. When someone gets ill and visits the doctor, the prescription for medicine is given in a piece of paper. This paper needs to be taken to a chemist to receive the medicines. In case of the loss of the paper containing prescription, the patient must revisit the doctor.

In this thesis, we try to find the impact blockchain technology can have in the domain of healthcare. On the other hand, we will also investigate the current state of blockchain technology and healthcare industry. We will then try to break healthcare into various subdomains (e.g. Health Information Exchange (HIE), claims adjudication and patient billing management, drug supply chain integrity, pharma clinical trials, etc.) and explore how each section can be improved through blockchain.

8.3 Challenges

Protecting a person’s medical data privacy and ensuring that data integrity is a big challenge for any health care system when they are going to share the data through all over the world with different level of stakeholders for their system purpose. Right now, the standard of data has reached a good standard. Keeping electronic health record (EHR) in different workflows and creating a trusted environment as whole for taking different type of crucial decision for medical fraternity is becoming important. It is important to ensure security of authority who is going to handle the system from up-to-date record of diagnoses, medication and services of individual person that is creating novel research area to improve total system [88]. It is going to be more important to update all the information for all stakeholder's same time with secure protocol.

For improving the total security of the EHR system is most valuable and crucial to give patients a healthy environment. In the country US, the health care system is looking for new secure and user-friendly environment [89]. Due to incentive system and federal laws have given access to health care data. It is also noticed that major hospital cannot share their safely. Patients' needs new type of system, where they can share their information in secure way and hassle freeway for communication with different stockholder. Considering all these in mind a good solution should be required to give confront to user [90].

Thus, building an access control for cross-boundary organization is an important and challenging issue. For example, Organization A grants the access to Organization C to give access to its micro services, which are S1, S2, and S3. Then, if Organization C would like to transfer its access permissions to Organization B, the question is how organization A can grant access to Organization B based on transferred access rights.

Figure 8.3. The primary blockchain structure adapted from [91]. The figure shows the way the ledger is structured in blockchain networks and describes the components that are inside the distributed ledger. The genesis block is assigned automatically when the network is started, with hash default values, and other blocks are inserted in the ledger following the genesis. In block structures, one could include components such as hash from the previous block, nonce, timestamp, block version, and the value known as the root of the Merkle tree. The Merkle tree is used to organize transactions into a blockchain network, and store them with security, as shown in the red

line. The attributes allocated into the block could be modified depending on the consensus protocol used; in summary, this figure presents the characteristics inserted in blockchain structure like Ethereum and Bitcoin implementations.

8.4 Methodology

Electronic health record (EHR) is important to keep a person's information as patient with confidential. IoMT is getting acknowledge from different stakeholders with Blockchain that gives secure path to continue the transaction between different actors. It possible to place the patient in secure way in health care ecosystem through Blockchain technology.

Blockchain is composed of distributed system that keep records of different transactions. It keeps record in digital ledger of collective, unchanging record over time of peer to peer trades created from connected transaction blocks. It trusts on building cryptographic systems to give permission to each participant to share information between them even the peers have no pre-existing trust certification. Creating platform for this new Blockchain based health system needs to establish few backbones before work with this in national wide. It is possible to reduce complexity using this system.

Internet of medical things (IoMT) and blockchain is the key game player in near future. Using these two technologies there could be a huge change and revaluation is possible and what is already begin.

Figure 8.3 represent a model that present a tentative diagram for including IoMT in the medical information system. In the figure, left part is the different devices those are consist of different sensors that collect health information and store in a local database, and it send data to central processing server for further evaluation of the medical information. Patient also have connection with central information processing center.

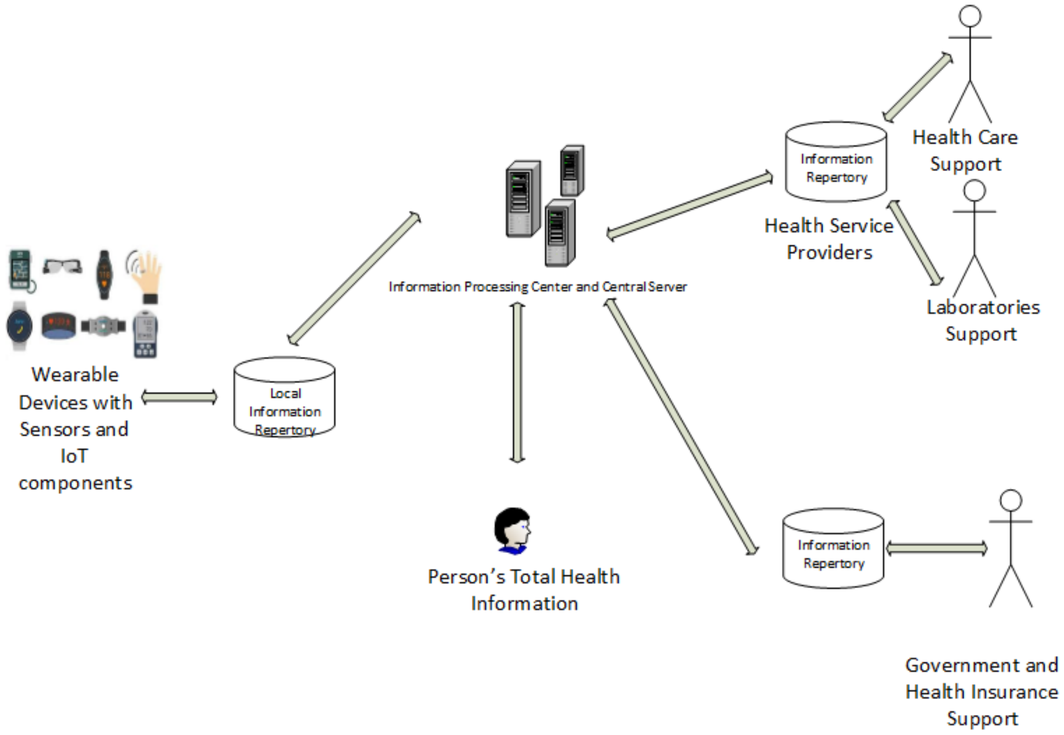

Health Care Support
Information Repertory
Health Service Providers
Laboratories Support
Information Processing Center and Central Server
Wearable Devices with Sensors and IoT components
Local Information Repertory
Person's Total Health Information
Information Repertory
Government and Health Insurance Support

Figure 8.3 Proposed Model

On the right side, the actors named health care support, laboratories support, government and insurance support are all connected with central information center. All these entities could be connected through blockchain system and make a secure way to transfer information to each other and keep track of the updates in same time to all the actors of the system. Blockchain is not a central system, it is a distributed platform to communicate with different actors in a system in secure way.

Internet of Medical things is a platform with embedded system consist of different types of sensors that collect health information form a patient. Figure below is a block diagram of the IoMT with different types of sensors.

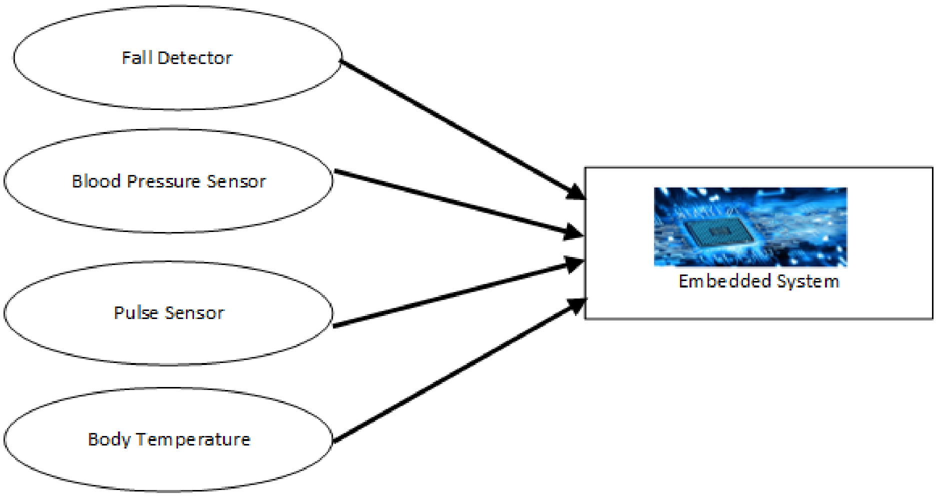


Figure 8.4 IoMT with embedded system with different sensors

This figure 8.4, include fall detection sensor, blood pressure sensor, pulse sensor, body temperature sensor and all sensors are in cooperate with an embedded plat for that can communicate with other part to send information about a patient body condition. IoMT is new game changer in the field of agile medical health sector.

The sequence diagram presents a full scenario for grant access to secure health care system is represented in figure 8.5. There are 6 actors in the sequence diagram. Administrator has access to patient's information. Patient's has IoMT that has connection with web service application. Web service backend application has connection with blockchain.

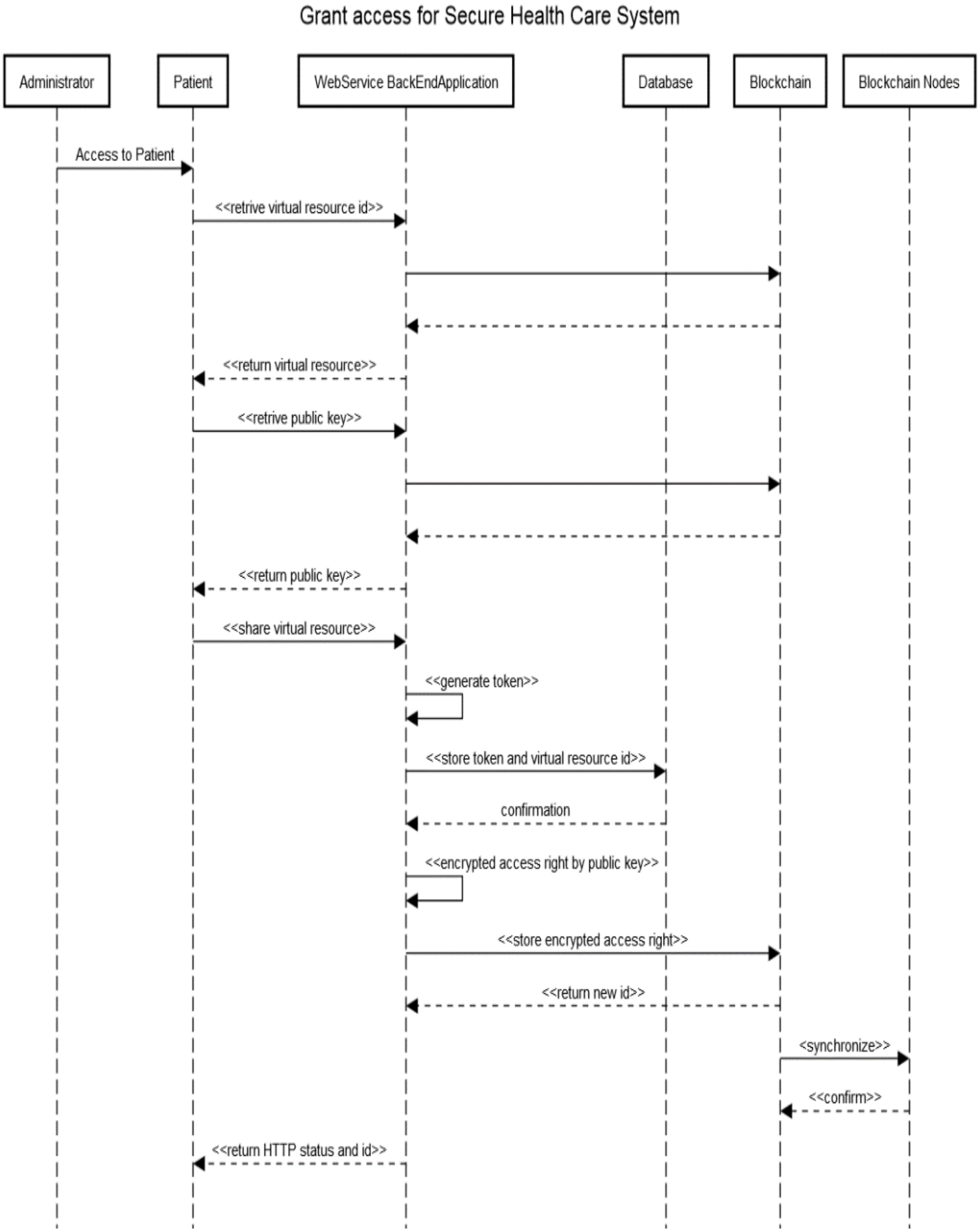

Grant access for Secure Health Care System
Administrator
Patient
WebService BackEndApplication
Database
Blockchain
Blockchain Nodes
Access to Patient
<<retrive virtual resource id>>
<<return virtual resource>>
<<retrive public key>>
<<return public key>>
<<share virtual resource>>
<<generate token>>
<<store token and virtual resource id>>
confirmation
<<encrypted access right by public key>>
<<store encrypted access right>>
<<return new id>>
<synchronize>>
<<confirm>>
<<return HTTP status and id>>

Figure 8.5 Sequential diagram

It sends information to blockchain and get notification from blockchain in different verification stage to secure the communication and update the total information. The system shares public key to secure the shared information. Blockchain Nodes synchronize the information about the update of all actor's information sharing and updating. In this way blockchain keep track of all the transaction between different stockholders of the system and keep the system secure through public key distribution.

8.5 Summary

We explored the current situation of the healthcare sector and explored how blockchain technology can make the healthcare system more efficient and secure. We realized that any scenario that includes data exchange where multiple stake holders are involved can be a good use case for blockchain. Sharing health records, medical records and research data among multiple parties is a common thing in healthcare. Hence, we believe healthcare can be a good application area for blockchain.

Firstly, we investigated the blockchain technology from a general perspective to understand how the technology works and investigated the different layers within blockchain. After having the basic understanding of blockchain technology and why it is gaining popularity, we investigated the different sectors where blockchain technology has had an impact or is starting or have an impact. For this, we discussed how different sectors can leverage blockchain and what solution already exists concerning blockchain.

In addition, the healthcare industry was divided into small sub-sectors to study how each individual sector can leverage blockchain technology. For this, we briefly described how each sector can be made more productive through blockchain and provided some existing solutions. One of the most important and major parts of the thesis was to do a literature review where we investigated published scientific journal papers that discuss how blockchain technology can be implemented and incorporated with the healthcare sector.

9 CONCLUSION AND FUTURE WORK

The majority of available classification and segmentation techniques for detecting cancer biomarkers, focus on cultivating, enhancing and conducting pragmatic investigations on diverse datasets which may not be suitable for advancing the state of the art. Instead, this dissertation examines classification and segmentation techniques from a distinct viewpoint. The main goal is to deliver a prognostication model that uses various biomedical images and microarray gene expression datasets to identify cancer biomarkers for disease diagnosis and provides predictions within reasonable accuracy. The model presented in this research showed significant contributions compared with state-of-the-art methods. Moreover, for the protection of biomedical datasets, blockchain technology is chosen for secured sharing of biomedical contents and genome expression data. This chapter provides an outline of our main approaches.

9.1 Synopsis of the Research

The principal purpose of our research is to scrutinize and investigate classification and segmentation methods for cancer biomarker detection from multimodal images of different organs of the body such as lung, breast, and colorectal, etc. to obtain insight to develop models that can improve performance in comparison to the state-of-the-art methods. The necessity of deeper knowledge of the classification, segmentation, and feature representation methods and identifying their characteristics is essential to improve the overall performance. This aided in building a strong classification and segmentation technique for better disease that can be used for more accurate diagnosis. Moreover, feature extraction and feature selection techniques are applied for microarray gene expression analysis for metastasis prediction. Dimensionality reduction techniques like PCA, ICA, LDA, and MDS are considered essential to reduce the number of genes in microarray datasets to get improve the overall accuracy with the selected significant genes. The literature review conducted at the beginning of this research dissertation introduced background survey of biomedical image classification, segmentation, and genome expression analysis for disease diagnosis. Moreover, we considered a blockchain technique in a distributed environment for secured sharing of medical contents among patients, doctor, and caregivers. The presented classification and segmentation techniques observed significant outcomes providing the efficacy of these methods.

First, we offered a lung nodule detection and classification method for early diagnosis of lung cancer from publicly available lung images from LUNA, Kaggle, and LDIC-IDRI datasets. U-net with z-score normalization technique applied for pre-processing of the images with a view that focuses on white pixels for nodule identification. Moreover, an autoencoder with local binary pattern and flattening technique applied for lung nodule detection. Additionally, a 3D DenseNet classifier is applied for better detection of the nodule in lung CT images. The proposed model presented significant improvements for detecting nodule in lungs. We can apply the same approach for the 3D DenseNet model for multi organ disease classification to show the efficacy of the method.

Second, we established an ensemble of pre-trained Convolutional Neural Network with transfer learning for microscopic image classification for detecting cancer types. Breast cancer histopathological image and cervical cancer microscopic images is considered with ensemble CNN model. Data augmentation and pre-processing techniques applied to augment the dataset before fed into the neural network. We have compared the model with other state-of-the-art methods and achieved significant performance for breast cancer detection and cervical cancer analysis.

Third, we proposed an approach for colorectal cancer detection from GEO microarray gene expression data. Machine learning techniques have been applied to gene expression data and compared the accuracy, precision, recall, etc. Finally, we reduced the number of genes by applying a several dimensionality reduction techniques and applied the same machine learning algorithms in order to compare the performances before and after the reduction of the genes. We were able to identify the most significant genes that are responsible for the colorectal cancer which represents a significant improvement.

Fourth, we offered a distributed environment secured with blockchain technique for the protection of sharing the medical contents among patients, doctors, insurance company, third party, and caregivers. A patient centric approach as well as electronic health record management technique applied to maintain the security while sharing, modifying the medical contents. Blockchain also applied for remote patient monitoring with internet of medical things for data collection and sharing through blockchain. A Hyperledger and an Ethereum technique is considered for building the blockchain architecture.

To conclude, our results have been compared to the state-of-the-art methods and showed considerable improvement compared to well-known image classification and segmentation techniques used for disease diagnosis. Metastasis prediction also showed significant results in comparison with feature extraction and feature selection approaches. Further improvement can be possible by focusing on robust ensemble deep learning model for disease diagnosis.

9.2 Future Work

In this dissertation, it has been observed that the proposed model for image and genome classification and segmentation showed significant accuracy improvement for detecting cancer biomarker. However, more accuracy improvement can be considered by focusing on robust deep learning models for cancer classification. In terms of microarray gene expression analysis, more feature extraction and selection technique can be considered for the benchmark dataset. Blockchain has been used extensively in healthcare to secured sharing of medical content. But robust cryptographic algorithm can be considered for smooth and secure sharing of the data.

There are several major concerns that should be addressed to contribute more in this research are stated below.

9.2.1 Analysis of the Proposed Classification and Segmentation Model for Disease Diagnosis

In this analysis, following points are the key considerations for ensemble classification and segmentation method-

1. More real-life datasets should be introduced to evaluate the model to overcome the overfitting issue.
2. The ensemble classification method can be compared with other state-of-the-art ensemble method with multimodal multi-organ images. The running time of 3D DenseNet for volumetric image classification can be reduced by incorporating robust models.
3. Data generation and augmentation using ensemble GAN can improve the accuracy of the model.
4. Cost effective deep Bayesian Active Learning could be a good choice for biomedical image classification. More noble deep learning architecture can be proposed for cancer classification.

9.2.2 Association of Radio-Genomics, Histo-Genomics, and Biomolecular-genome analysis

For performance evaluation of classification model using image, genome, and histology association study has the following key points-

1. In terms of disease diagnosis, imaging or genomics or microscopic study is not enough for finding better biomarker. Moreover, association study could be a good choice to ensure better accuracy results.
2. Functional correlations among different association study of image-genome, histo-genome or biomolecular-genome can significantly improve the performances. Multiple way analysis of disease diagnosis would be reliable for disease diagnosis.
3. Association study needs the dataset for the same sample size, and it is hard to manage the confidential biomedical dataset.

9.2.3 Blockchain Protocol in a Distributed System

The following points are the main considerations for secured sharing of the biomedical contents based on blockchain-

1. Analysis of the different protocol to generate the blockchain index and compare the performances of the protocol could be good idea
2. Ensemble graph algorithm can be considered for determining the architecture of the blockchain in a distributed system. Greedy graph algorithm can be applied for blockchain architecture.

REFERENCES

[1] E. Tasci and A. Ugur, “Shape and texture based novel features for automated juxta pleural nodule detection in lung CTs,” Journal of medical systems, vol. 39, no. 5, p. 46, 2015.

[2] A. M. Santos, A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, “Automatic detection of small lung nodules in 3d CT data using gaussian mixture models, Tsallis entropy and SVM,” Engineering applications of artificial intelligence, vol. 36, pp. 27–39, 2014.

[3] Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio, “Pylearn2: a machine learning research library,” arXiv preprint arXiv:1308.4214, 2013.

[4] H. Lee, J. Lee, and S. Cho, “View-interpolation of sparsely sampled sinogram using convolutional neural network,” in Medical Imaging 2017: Image Processing, vol. 10133. International Society for Optics and Photonics, 2017, p. 1013328.

[5] H. Lee, J. Lee, H. Kim, B. Cho, and S. Cho, “Deep-neural-network based sinogram synthesis for sparse-view CT image reconstruction,” arXiv preprint arXiv:1803.00694, 2018.

[6] J. Kim, J. Kwon Lee, and K. Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 2, pp. 295–307, 2016.

[8] M S. Sajjadi, B. Scholkopf, and M. Hirsch, “Enhancenet: Single image super-resolution through automated texture synthesis,” arXiv preprintarXiv:1612.07919, 2016.

[9] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, “Exploiting image-trained cnn architectures for unconstrained video classification,” arXiv preprint arXiv:1503.04144, 2015.

[10] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian. “Multi-scale convolutional neural networks for lung nodule classification,” In IPMI, 2015.

[11] X. Yan, J. Pang, H. Qi, Y. Zhu, C. Bai, X. Geng, M. Liu,D. Terzopoulos, and X. Ding. “Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies,” In ACCV, 2016.

[12] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, “Community-enhanced deanonymization of online social networks," in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, ser. CCS '14. New York, NY, USA: ACM, 2014, pp. 537-548.

[13] S. E. Park, X. Han, T. L. Berg, and A. C. Berg, “Combining multiple sources of knowledge in deep CNN’s for action recognition,” in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE,2016, pp. 1–8.

[14] Zhang, Y.; Wu, L.; Wang, S., “Magnetic resonance brain image classification by an improve artificial bee colony algorithm. Progress Electromagnetic Resolution,” 116, 65–79, 2011.

[15] Chaplot, S.; Patnaik, L.M.; Jagannathan, N.R., “Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network,” Biomed. Signal Process Control, 1, 86–92.

[16] Safaa E.Amin, M.A. Mageed, “Brain Tumor Diagnosis Systems Based on Artificial Neural Networks and Segmentation Using MRI,” IEEE International Conference on Informatics and Systems, INFOS 2012.

[17] Heba Mohsen, EL-Sayed A. EL-Dahshan, EL-Sayed M. EL-Horbaty, Abdel-Badeeh M. Salem, “Classification using deep learning neural networks for brain tumors,” IEEE International Conference on Informatics and Systems, INFOS 2017.

[18] Suchita Goswami, Lalit Kumar P. Bhaiya, “Brain Tumor Detection Using Unsupervised Learning based Neural Network”, IEEE International Conference on Communication Systems and Network Technologies, 2013.

[19] J.Seetha and S. Selvakumar Raja., “Brain tumor classification using Convolutional neural network,” Biomedical and Pharmacology Journal vol.11(3), 1457-1461.

[20] L.D. Nguyen, D. Lin, Z. Lin, J. Cao, “Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation,” Circuits and Systems (ISCAS), 2018 IEEE International Symposium on, IEEE, 2018, pp. 1–5.

[21] R. Awan, N.A. Koohbanani, M. Shaban, A. Lisowska, N. Rajpoot, “Context-aware learning using transferable features for classification of breast cancer histology images,” International Conference Image Analysis and Recognition, Springer, 2018, pp. 788–795.

[22] A. Kensert, P.J. Harrison, O. Spjuth, “Transfer learning with deep convolutional neural network for classifying cellular morphological changes,” bioRxiv (2018) 345728.

[23] S. Vesal, N. Ravikumar, A. Davari, S. Ellmann, A. Maier, “Classification of breast cancer histology images using transfer learning,” International Conference Image Analysis and Recognition, Springer, 2018, pp. 812–819.

[24] Naik, J.; Prof. Patel, Sagar, “Tumor Detection and Classification using Decision Tree in Brain MRI,” IJEDR, ISSN:2321-9939, 2013.

[25] H. Zhang, J. E. Fritts, S. A. Goldman, “Image Segmentation Evaluation: A Survey of Unsupervised Methods”, Computer Vision and Image Understanding, Pp. 260-280, 2008.

[26] Cui W, Wang Y, Fan Y, Feng Y, Lei T, “Localized FCM clustering with spatial information for medical image segmentation and bias field estimation,” Int J Biomed Imaging 2013, Article ID 930301,8 pages [PMC free article] [PubMed].

[27] Liu T, Fang Y, Zhang H, Deng M, Gao B, Niu N, Yu J, Lee S, Kim J, Qin B, et al, “HEATR1 negatively regulates Akt to help sensitize pancreatic cancer cells to chemotherapy,” Cancer Res 76: 572-581, 2016.

[28] Tsukamoto Y, Fumoto S, Noguchi T, Yanagihara K, Hirashita Y, Nakada C, Hijiya N, Uchida T, Matsuura K, Hamanaka R, et al, “Expression of DDX27 contributes to colony-forming ability of gastric cancer cells and correlates with poor prognosis in gastric cancer,” Am J Cancer Res 5: 2998-3014, 2015.

[29] Liu WB, Jia WD, Ma JL, Xu GL, Zhou HC, Peng Y and Wang W, “Knockdown of GTPBP4 inhibits cell growth and survival in human hepatocellular carcinoma and its prognostic significance,” Onco target 8: 93984-93997, 2017.

[30] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2. IEEE, 2006, pp. 2169–2178.

[31] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[32] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Volume 1 - Volume 01, ser. CVPR ’05, 2005, pp.886–893.

[33] O. Deniz, G. Bueno, J. Salido, and F. De la Torre, “Face recognition using histograms of oriented gradients,” Pattern Recogn. Lett., vol. 32, no. 12, pp. 1598–1603, Sep. 2011.

[34] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, “Fast human detection using a cascade of histograms of oriented gradients,” in Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, ser. CVPR ’06, 2006, pp. 1491–1498.

[35] Cheng, Erkang, et al., “Discriminative vessel segmentation in retinal images by fusing context-aware hybrid features,” Machine vision and applications 25.7 (2014): 1779-1792.

[36] Wang Xiancheng, Li Wei, Miao Bingyi, Jing He, Zhangwei Jiang, Wen Xu, Zhenyan Ji, Gu Hong, Shen Zhaomeng, “Retina Blood Vessel Segmentation Using A U-Net Based Convolutional Neural Network,” International Conference on Data Science (ICDS2018)

[37] Liskowski, Paweł, and Krzysztof Krawiec, “Segmenting Retinal Blood Vessels with Deep Neural Networks,” IEEE transactions on medical imaging 35.11 (2016): 2369-2380.

[38] R. Hu and J. Collomosse, “A performance evaluation of gradient field hog descriptor for sketch-based image retrieval,” Computer Vis. Image Understand., vol. 117, no. 7, pp. 790–806, Jul. 2013.

[39] W.-J. Choi and T.-S. Choi, “Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach,” Entropy, vol. 15, no. 2, pp. 507–523, 2013.

[40] A. Chon, N. Balachandar, and P. Lu, “Deep convolutional neural networks for lung cancer detection,” tech. rep., Stanford University, 2017.

[41] Y. LeCun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision.,” Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), pp. 253–256, IEEE, 2010.

[42] K. Alex, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems 25 (NIPS 2012), pp. 1097–1105, 2012.

[43] H. Suk, S. Lee, and D. Shen, “Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis,” Neuro Image, vol. 101, pp. 569–582, 2014.

[44] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, “Unsupervised deep feature learning for deformable registration of MR brain images.,” Medical Image Computing and Computer-Assisted Intervention, vol. 16, no. Pt 2, pp. 649–656, 2013.

[45] J. Glaister, A. Wong, and D. A. Clausi, “Segmentation of Skin Lesions from Digital Images Using Joint Statistical Texture Distinctiveness,” IEEE Transactions on Biomedical Engineering, vol. 61, pp. 1220-1230, 2014.

[46] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, Oct. 2015, pp. 234–241

[47] M.H. Jafari, N. Karimi, and K. Najarian, “Skin Lesion Segmentation in Clinical Images Using Deep Learing,” IEEE International Conference on Pattern Recognition, 2016.

[48] Brahim AIT SKOURT, Abdelhamid EL HASSANI, Aicha MAJDA, “Lung CT Image Segmentation Using Deep Neural Networks,” The First International Conference on Intelligent Computing in Data Sciences, Procedia Computer Science 127 (2018) 109–113.

[49] Hieu Trung Huynh*and Vo Nguyen Nhat Anh, “A deep learning method for lung segmentation onlarge size chest X-ray image,” IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), 2019.

[50] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I. Chang, “Deep learning of feature representation with multiple instance learning for medical image analysis,” IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 1626–1630, 2014.

[51] D. Kumar, A.Wong, and D. A. Clausi, “Lung nodule classification using deep features in ct images,” 12th Conference on Computer and Robot Vision, pp. 133–138, June 2015.

[52] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, “Chest pathology detection using deep learning with non-medical training,” Proceedings - International Symposium on Biomedical Imaging, vol. July, pp. 294–297, 2015.

[53] W. Sun, B. Zheng, and W. Qian, “Computer aided lung cancer diagnosis with deep learning algorithms,” in SPIE Medical Imaging, vol. 9785, pp. 97850Z–97850Z, International Society for Optics and Photonics, 2016.

[54] J. Tan, Y. Huo, Z. Liang, and L. Li, “A comparison study on the effect of false positive reduction in deep learning based detection for Juxtapleural lung nodules: CNN vs DNN,”

in Proceedings of the Symposium on Modeling and Simulation in Medicine, MSM ’17, (San Diego, CA, USA), pp. 8:1–8:8, Society for Computer Simulation International, 2017.

[55] R. Golan, C. Jacob, and J. Denzinger, “Lung nodule detection in CT images using deep convolutional neural networks,” in 2016 International Joint Conference on Neural Networks (IJCNN), pp. 243–250, July 2016.

[56] Kaggle, “Data science bowl 2017.” https://www.kaggle.com/c/datascience- bowl-2017/data, 2017.

[57] LUNA16, “Lung nodule analysis 2016.” https://luna16.grandchallenge.org/, 2017.

[58] M. Firmino, A. Morais, R. Mendoa, M. Dantas, H. Hekis, and R. Valentim, “Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects,” BioMedical Engineering Online, vol. 13, p. 41, 2014.

[59] S. Hawkins, H. Wang, Y. Liu, A. Garcia, O. Stringfield, H. Krewer, Q. Li, D. Cherezov, R. A. Gatenby, Y. Balagurunathan, D. Goldgof, M. B. Schabath, L. Hall, and R. J. Gillies, “Predicting malignant nodules from screening CT scans,” Journal of Thoracic Oncology, vol. 11, no. 12, pp. 2120–2128, 2016.

[60] Spanhol FA, Oliveira LS, Cavalin PR, Petitjean C, Heutte L., “Deep features for breast cancer histopathological image classification,” Systems, Man, Q7 and Cybernetics (SMC), IEEE International Conference On. Banff: IEEE; 2017. p. 1868–73.

[61] Spanhol FA, Oliveira LS, Petitjean C, Heutte L., “Breast cancer histopathological image classification using convolutional neural networks,” Neural Networks (IJCNN), 2016 International Joint Conference On. Vancouver: IEEE; 2016. p. 2560–7.

[62] Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, Campilho A., "Classification of breast cancer histology images using convolutional neural networks," PloS one. 2017;12(6):0177544.

[63] Ojala T, Pietikainen M, Maenpaa T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans Patt Anal Mach Intell. 2002;24(7):971–87.

[64] Ojansivu V, Heikkilä J., "Blur insensitive texture classification using local phase quantization," In: Int Confer Image Signal Proc. Berlin: Springer; 2008. p. 236–43.

[65] Setio, A. A. A.; Traverso, A.; de Bel, T.; Berens, M. S.; van den Bogaard, C.; Cerello, P.; Chen, H.; Dou, Q.; Fantacci, M. E.; Geurts, B.; et al, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge.," Medical Image Analysis 42:1–13, 2017.

[66] Shen, W.; Zhou, M.; Yang, F.; Yang, C.; and Tian, J., "Multiscale convolutional neural networks for lung nodule classification," IPMI, 588–599. Springer, 2015.

[67] Shen, W.; Zhou, M.; Yang, F.; Dong, D.; Yang, C.; Zang, Y.; and Tian, J., "Learning from experts: Developing transferable deep features for patient-level lung cancer prediction," MICCAI, 124–131. Springer, 2016.

[68] Suzuki, K.; Li, F.; Sone, S.; and Doi, K., "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network," IEEE Transactions on Medical Imaging 24(9):1138–1150, 2005.

[69] Veta M, Pluim JP, Van Diest PJ, Viergever MA, "Breast cancer histopathology image analysis: A review," IEEE Trans Biomed Eng. 2014; 61(5):1400–11.

[70] Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S., “Breast cancer multi-classification from histopathological images with structured deep learning model,” Sci Rep. 2017;7(1):4172.

[71] Kowal M, Filipczuk P, Obuchowicz A, Korbicz J, Monczak R., “Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images,” Computer Biol Med. 2013;43(10):1563–72.

[72] Filipczuk P, Fevens T, Krzyzak A, Monczak R., “Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies,” IEEE Trans Med Imaging. 2013;32(12):2169–78.

[73] Spanhol FA, Oliveira LS, Petitjean C, Heutte L., “A dataset for breast cancer histopathological image classification,” IEEE Trans Biomed Eng. 2016;63(7): 1455–62.

[74] Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, Hulsbergen-Van De Kaa C, Bult P, Van Ginneken B, Van Der Laak J., “Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis,” Sci Rep. 2016; 6:26286.

[75] Cire¸san DC, Giusti A, Gambardella LM, Schmidhuber J., “Mitosis detection in breast cancer histology images with deep neural networks,” International Conference on Medical Image Computing and Computer-assisted Intervention. Berlin: Springer; 2013. p. 411–8.

[76] Bayramoglu N, Kannala J, Heikkilä J., “Deep learning for magnification independent breast cancer histopathology image classification,” Pattern Recognition (ICPR), 2016 23rd International Conference On. Cancun: IEEE; 2016. p. 2440–5.

[77] He K, Zhang X, Ren S, Sun J., “Deep residual learning for image recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE; 2016. p. 770–8.

[78] Wu J, Leng C, Wang Y, Hu Q, Cheng J., "Quantized convolutional neural networks for mobile devices," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE; 2016. p. 4820–8.

[79] Courbariaux M, Hubara I, Soudry D, El-Yaniv R, Bengio Y., "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," arXiv preprint arXiv:1602.02830. 2016.

[80] Srinivas S, Babu RV., "Data-free parameter pruning for deep neural networks," arXiv preprint arXiv:1507.06149. 2015.

[81] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, J. Tomaszewski, F. A. Gonz´alez, and A. Madabhushi., "Accurate and reproducible invasive breast cancer detection in wholeslide images: A deep learning approach for quantifying tumor extent," Scientific Reports, 7:46450, 2017.

[82] H. Daume III and D. Marcu., "Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research," 26:101–126, 2006.

[83] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell., "Decaf: A deep convolutional activation feature for generic visual recognition," International conference on machine learning, pages 647–655, 2014.

[84] A. Dosovitskiy, J. T. Springenberg, and T. Brox., "Unsupervised feature learning by augmenting single images," CoRR, abs/1312.5242, 2013.

[85] D. Eddelbuettel, M. Stokely, and J. Ooms. Rprotobuf, "Efficient cross language data serialization in r," arXiv preprint:1401.7372, 2014.

[86] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 542(7639):115–118, 2017.

[87] C. Fitzmaurice, C. Allen, R. M. Barber, L. Barregard, Z. A. Bhutta, H. Brenner, D. J. Dicker, O. Chimed-Orchir, R. Dandona, L. Dandona, et al., "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study," JAMA oncology, 3(4):524–548, 2017.

[88] Nishimura Y, Takiguchi S, Ito S, and Itoh K, "Evidence that depletion of the sorting nexin 1 by siRNA promotes HGF-induced MET endocytosis and MET phosphorylation in a gefitinib-resistant human lung cancer cell line," Int J Oncol 44: 412-426, 2014.

[89] Naushad SM, Reddy CA, Kumaraswami K, Divyya S, Kotamraju S, Gottumukkala SR, Digumarti RR and Kutala VK, "Impact of hyperhomocysteinemia on breast cancer initiation and progression: Epigenetic perspective," Cell Biochem Biophys68: 397-406, 2014.

[90] Han L, Wu Z and Zhao Q., "Revealing the molecular mechanism of colorectal cancer by establishing LGALS3-related protein-protein interaction network and identifying signaling pathways," Int J Mol Med 33: 581-588, 2014.

[91] Takahashi Y, Sawada G, Kurashige J, Uchi R, Matsumura T, Ueo H, Takano Y, Eguchi H, Sudo T, Sugimachi K, et al., "Amplification of PVT-1 is involved in poor prognosis via apoptosis inhibition in colorectal cancers," Br J Cancer110: 16 4-171, 2014.

[92] Gantt GA, Chen Y, Dejulius K, Mace AG, Barnholtz-Sloan J and Kalady MF, "Gene expression profile is associated with chemo-radiation resistance in rectal cancer," Colorectal Dis 16: 57-66, 2014.

[93] Di Franco S, Turdo A, Todaro M and Stassi G, “Role of type I and II interferons in colorectal cancer and melanoma,” Front Immunol 8: 878, 2017.

[94] Song D, Huang R, Tang Q, et al, “Identification of EZH2-related key pathways and genes in colorectal cancer using bioinformatics analysis,” Chin J Colorectal Dis 5: 475-479, 2016.

[95] Liang B, Li C and Zhao J, “Identification of key pathways and genes in colorectal cancer using bioinformatics analysis,” Med Oncol 33: 111, 2016.

[96] Guo Y, Bao Y, Ma M and Yang W, “Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis,” Int J Mol Sci 18: pii: E722, 2017.