ABSTRACT

Past years have experienced increasing mortality rate due to lung cancer and thus it becomes crucial to predict whether the tumor has transformed to cancer or not, if the prediction is made at an early stage then many lives can be saved and accurate prediction also can help the doctors start their treatment. Computed tomography plays a vital role in ensuring the condition of tumor that by checking the size of tumor, location of tumor, etc. In this paper, we have proposed a framework for prediction of cancer at an early stage so that many lives that are in an endangered situation could be revived. Basically, our focus is on two domains of computer science that is Digital Image Processing acronymed DIP and Machine Learning. Digital image processing is well-known for the phase of preprocessing the image. In the further stage, the pre-processed image is exposed to segmentation phase and then the segmented image is passed for feature extraction and finally the extracted features are trained using machine learning classification algorithms like SVM (Support Vector Machines), Random Forest, ANN (Artificial Neural Network). Based on the classification results obtained, prediction is made whether the tumor is benign or malignant. The inevitable parameters such as accuracy, Recall and precision are calculated for determining which algorithm has the highest predictive accuracy.

TABLE OF CONTENTS

CHAPTER NUMBER	TITLE	PAGE NUMBER
	ABSTRACT	**V**
	LIST OF FIGURES	**VIII**
	LIST OF TABLES	
1	**INTRODUCTION**	**1**
	1.1 OVERVIEW	1
	1.2 PURPOSE OF MACHINE LEARNING	2
	1.3 PROBLEM STATEMENT	2
	1.4 OBJECTIVES	3
2	**2.1 LITERATURE SURVEY**	4
	2.2 SYSTEM ANALYSIS	7
	2.2.1 DRAWBACKS OF EXIXTING SYSTEMS	7
	2.2.2 PROPOSED SYSTEM	7
	2.2.3 SYSTEM REQUIRMENTS	8
	2.2.4 HARDWARE REQUIRMENTS	8
	2.2.5 SOFTWARE REQUIRMENTS	8
3	**SYSTEM DESIGN**	**10**
	3.1 PROPOSED WORK	10
	3.2 ARCHITECHTURE DIAGRAMM	12
	3.3 MODILES	12
	3.3.1 PRE PROCESSING LAYER	12
	3.3.2 SEGMENTATION LAYER	13
	3.3.3 FEATURE EXTRACTION LAYER	13
	3.3.4 CLASSIFICATION LAYER	14

4	**SYSTEM IMPLEMENTATION**	**15**
	4.1 UML DIAGRAMS	15
	4.2 DATASET RESEARCH	22
	4.3 PROPOSED ALGORITHMS	23
5	**TESTING METHODS AND RESULTS**	**24**
	5.1 UNIT TESTING	24
	5.2 INTEGRATION TESTING	24
	5.3 ACCEPTANCE TESTING	24
	5.4 ACCURACY TESTING	26
6	**CONCLUSION**	**28**
	6.1 FUTURE SCOPE AND CONCLUSION	28
	6.2 REFERENCES	29
	APPENDIX	**33**
	A. SOURCE CODE	33
	B. OUTPUT SCREENSOTS	36

LIST OG FIGURES

FIGURE NO	FIGURE NAME	PAGE NO
3.1	FLOW DIAGRAM OF PROPOSED WORK	11
3.2	IMPLEMENTATION MODEL	12
4.1	USE CASE DIAGRAMS	16
4.1.1	UPLOAD CT SCANS	17
4.1.2	VIEW DETECTION RESULTS	17
4.1.3	MAKE PREDICTIONS	18
4.1.4	VIEW PREDICTIONS	19
4.1.5	CT SCAN SLICES	20
4.1.6	CANCER MASKS	21
5.4	REGION TABLE FEATURE	27
5.4.1	ACCURACY GRAPH	27

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Machine learning, as a powerful approach to achieve Artificial Intelligence, has been widely used in pattern recognition, a very basic skill for humans but a challenge for machines. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

Since the computer was invented, it has begun to affect our daily life. It improves the quality of our lives; it makes our life more convenient and more efficient. A fascinating idea is to let a computer think and learn as a human. Basically, machine learning is to let a computer develop learning skills by itself with given knowledge. Pattern recognition can be treated like computer being able to recognize different species of objects. Therefore, machine learning has close connection with pattern recognition.

Machine Learning is a scientific research of statistical procedures and methods which they are used by computer systems designed to perform such functions without specific instructions, rather than trusting in the models and conclusions. This is believed to be part of an artificial intelligence. Machine Learning algorithms sets up a mathematical model based on data examples called "training data" to make predictions without the completion of a task being explicitly programmed.

1.2 PURPOSE OF THE MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

1.3 PROBLEM STATEMENT

With the rapid increase in population rate, the rate of diseases like cancer, chikungunya, cholera etc., are also increasing. Among all of them, cancer is becoming a common cause of death. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow older or become damaged, they die, and new cells take their place. When cancer cells develop, however, this orderly process breaks down. As cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumor. This tumor starts spreading to different of body. Tumors are of two types benign and malignant where benign (non-cancerous) is the mass of cell which lack in ability to spread to other part of the body and malignant (cancerous) is the growth of cell which has ability to spread in other part

of body this spreading of infection is called metastasis. There is various type of cancer like Lung cancer, leukemia, and colon cancer etc. The incidence of lung cancer has significantly increased from the early 19th century. There is various cause of lung cancer like smoking, exposure to radon gas, secondhand smoking, and exposure to asbestos etc.

1.4 OBJECTIVE

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

.

CHAPTER2

2.1 LITERATURE SURVEY

In the 21st century, cancer is still considered a serious disease as the mortality rates are high. Among all cancer types, lung cancer ranks first regarding morbidity and mortality [1, 2]. There are two main categories of lung cancer: non-small-cell lung cancer (NSCLC) and small cell lung cancer (SCLC). For non-small-cell lung cancer, a subcategorization into lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) is further used. These types of cancers account for approximately 85% of lung cancer cases [3]. Compared with the diagnosis of benign and malignant, further fine-grained classification of lung cancers such as LUSC, LUAD, and SCLC is of great significance for the prognosis of lung cancer. Accurately determining the category of lung cancer in the early diagnosis directly influences the effect of the treatment and thus the patients' survival rate [1, 4]. Positron emission tomography (PET) and computed tomography (CT) are both widely used noninvasive diagnostic imaging techniques for clinical diagnosis in general and for the diagnosis of lung cancer in particular [4]. Immunohistochemical evaluation is considered the gold standard for lung cancer classification. However, this procedure requires a tissue biopsy, an invasive procedure with the inherent risk of a delayed diagnosis and thus exacerbation of the patient's pain.

Advances in artificial intelligence research enabled numerous studies on the automatic diagnosis of lung cancer. The use of data in lung cancer-type classification is roughly divided into three categories: CT and PET image data as well as pathological images [5]. The well-known data science community Kaggle provides high-quality CT images for participants with the task to distinguish malignant or benign nodules from pulmonary nodules. Kaggle competitions repeatedly produce excellent deep learning approaches for these tasks [6, 7]. With the progresses in the research of automatic lung cancer diagnosis, studies are no longer limited to the classification of benign and malignant nodules and data sets are no longer limited to CT images [8–12]. Wu et al. [9] use quantitative imaging characteristics such as statistical, histogram-related, morphological, and textural features from PET images to predict the distance metastasis of NSCLC, which shows that quantitative features based on PET images can effectively characterize intratumor heterogeneity and complexity. Two recent publications propose the application of deep learning to pathological images to classify NSCLC and SCLC [10] and to classify transcriptome subtypes of LUAD [11]. The complexity of the clinical diagnosis of lung cancer is also characterized by the wide range of imaging modality, which is employed in the diagnosis [13, 14].

Previous research already proved that deep learning approaches can not only use the feature distribution patterns from different pulmonary imaging modalities but even merging different features to achieve the computer-aided diagnosis. Liang et al. [15] employ multichannel techniques to predict the IDH genotype from PET/CT data using a convolutional neural network (CNN), while other approaches use a parallel CNN architecture to extract several features of different imaging modalities [16, 17].

Compared with the classification of the benign and malignant, the classification of the three types of lung cancer from medical images are more suitable to constitute a fine-grained image recognition problem as diverse distributions of features and potential pathological features need to be considered. Because the fine-grained features which need to extract in images, and meanwhile the lesion region is a small part of the whole image, the deep learning framework is susceptible to feature noise. At present, most methods based on various deep learning frameworks have proved to have certain bottleneck in fine-grained problems. In order to solve this problem, the previous research mainly implements the attention mechanism from the two dimensions (channel and spatial) of the feature representation. The channel attention mechanism models the relationship between feature channels [18], while the spatial attention mechanism ensures that noise is suppressed by weighting feature representation spatially [19–21]. So far, spatial attention mechanism has been used in medical image processing to enhance extracted features [20, 21]. The channel attention mechanism has been used in the detection and classification of pulmonary disease [22, 23]. The presentation of these attention mechanisms illustrates the source of characteristic noise from different perspectives. There are few related studies on how to use the attention mechanism more effectively on images with different imaging modalities, so the deep learning model based on the multimodality dataset still has problems in fine-grained problems.

Many works has already been proposed for prediction of cancer by various researchers among then Palani et al., [5] has proposed IoT based predictive modeling by using fuzzy C mean clustering for segmentation and incremental classification algorithm using association rule mining and decision tree for classification for classifying the tumor sets and based on the output generated by incremental classification model convolutional neural network has been applied with other features for predicting benign or malignant.

Lynch et al., [6] Various machine learning algorithm are implemented for predicting the survivability rate of person, performance is measured based on root mean square error. Each model is trained using 10-fold cross validation, as the parameters are preprocessed by assigning default value so cross

validation is used for avoiding over fitting.

FENWA et al., [3] proposed a model whether feature like contrast, brightness from the image dataset is extracted using texture based feature extraction and on that two type of machine learning algorithm are applied one is artificial neural network another one is support vector machine and then performance has been evaluated on both the algorithm to compare which algorithm is giving more accuracy.

Öztürk et al., [7] proposed a model where a five type of feature extraction techniques were used in individual classification algorithm to predict at which features extraction technique which machine learning algorithm is giving more accuracy.

Jin et al., [8] proposed a model where the original image is first converted into binary image the erosion and dilution has been operated on that image after that image has been segmented on the segmented image region of interest extraction is applied to identify volume or size of the tumor and after extraction convolutional neural network is applied with softmax classification layer to recognize the tumor is cancerous or not.

Sumathipala et al., [9] proposed a model where the image data are taken from LIDC-IDRI, after collecting the image data image filtration has been implemented, filtration is done based on the patient who went through biopsy and module level is equal to 30 and then images whose module level is equal to 30 is segmented and then Logistic regression and random forest has been applied for prediction.

2.2 SYSTEM ANALYSIS

2.2.1 DRAWBACKS OF EXISTING SYSTEMS

In some cases, the application still does not have accurate results. Further optimization is needed.

Priority information is needed for segmentation. Database extension is required for greater accuracy.

Only a few diseases are covered. Therefore, the work must be expanded to cover more diseases.

Possible causes that can cause misclassification can be: Symptoms of the disease vary from cigarettes, optimizing the characteristics needed, more training patterns are needed to cover and predict more cases - the actual disease.

2.2.2 PROPOSED SYSTEM

The main theme of this project, is to detect the lung cancer and to take precautions to avoid or clear that diseases. It overcomes the drawbacks of existing system.

The implementation phase begins with smoking as input and do the following steps:

- Pre-Processing layer.
- Segmentation layer.
- Feature Extraction of layer.
- Machine learning classifier.

After performing all the above steps, we can detect the disease of lung cancer.