INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

Extraction and Classification of Lung Cancer Subtypes via Histopathological Imaging

1st P.Saranya,
Assistant Professor
dept. of Computer Science and Engineering
AVS Engineering College
Salem, India

2nd P.R.Dharshinipriya
dept. of Computer Science and Engineering
AVS Engineering College
Salem, India
prdharshinipriya@gmail.com

***Abstract*— Lung cancer (LC) is one of the most serious cancers threatening human health. Histopathological examination is the gold standard for qualitative and clinical staging of lung tumors. However, the process for doctors to examine thousands of histopathological images is very cumbersome, especially for doctors with less experience.**

Therefore, objective pathological diagnosis results can effectively help doctors choose the most appropriate treatment mode, thereby improving the survival rate of patients. For the current problem of incomplete experimental subjects in the computer-aided diagnosis of lung cancer subtypes, this study included relatively rare lung adenosquamous carcinoma (ASC) samples for the first time, and proposed a computer-aided diagnosis method based on histopathological images of ASC, lung squamous cell carcinoma (LUSC) and small cell lung carcinoma (SCLC).

Firstly, the multidimensional features of 121 LC histopathological images were extracted, and then the relevant features (Relief) algorithm was used for feature selection. Finally, through a horizontal comparison with a variety of mainstream classification models, experiments show that the classification effect achieved by the Relief-SVM model is the best. The LUSC-ASC classification accuracy was 73.91%, the LUSC-SCLC classification accuracy was 83.91% and the ASC-SCLC classification accuracy was 73.67%. Our experimental results verify the potential of the auxiliary diagnosis model constructed by machine learning (ML) in the diagnosis of LC.

Keywords—cancer, histopathology, machine learning

I. INTRODUCTION

Lung cancer become one of the most serious cancers threatening human health.
Clinicians' visual analysis of LC histopathological images is one of the most important methods for evaluating LC subtypes.
However, it is complicated and challenging for pathologists to review thousands of histopathological images, and it is even more difficult for doctors with less experience. Therefore, to relieve the pressure on doctors and improve the accuracy and efficiency of diagnosis, it is particularly important to study the computer-aided diagnosis model of LC. From the perspective of pathology and treatment, LC can be divided into non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC), of which 80%-85% are VOLUME 9, 2021. Research on Auxiliary Classification and Diagnosis of LC Subtypes NSCLC and the rest are SCLC. The main histological types of NSCLC are lung adenocarcinoma (ADC) and lung squamous cell carcinoma (LUSC).
The other histological types of NSCLC are lung adenosquamous carcinoma (ASC), large cell carcinoma. In particular, ASC is a relatively rare subtype of NSCLC that accounts for 0.3– 5% of all NSCLCs. Due to the different histopathological types of LC, the treatment methods adopted are also different. When the lung tissue classification is determined, the appropriate treatment mode can be selected, such as the reasonable application of surgery, chemotherapy, radiotherapy, molecular targeted therapy and immune therapy. In addition, LC screening errors can be avoided, clinicians' multifarious work pressure can be slowed, patients' survival time can be maximized and the patient's quality of life improved.
The LC imaging examination methods mainly include: (1) X-ray photography, which is one of the most basic lung imaging examination methods, but the resolution of the photography is low and there are blind spots in the examination . Computed tomography (CT), specifically, chest CT has advantages in detecting early peripheral LC and identifying the location of the lesion. Currently, it is one of the most commonly used imaging methods for preoperative diagnosis and staging of LC. However, one of the limitations of using CT examinations is that for patients undergoing repeated examinations, it is necessary to consider the impact of the radiation dose produced by the operation. Magnetic resonance imaging (MRI) has high sensitivity and specificity for vertebral and bone metastases, but it is not recommended for routine diagnosis of LC Ultrasound is a non-invasive tool, which is usually superior to radiography in the examination of postoperative pulmonary complications (PPCs).

It has developed into a valuable method. Although the above imaging examination methods play an important role in detecting of LC, the results of each examination are only used as a reference for the diagnosis, staging, re-staging, efficacy monitoring and prognosis evaluation of LC, while histopathological examination is the gold standard for tumor qualitative and clinical staging . However, due to the complex texture features of histopathological images, as far as the authors know, there is no computer-aided diagnosis method for ASC, LUSC and SCLC based on histopathological images. .
This paper includes ASC sample data for the first time and introduces a computer-aided model for automatic classification of LC subtypes based on histopathological images of LC. First, seven texture analysis methods are used to extract 265-dimensional features of LC histopathological images, and the relevant features (Relief) algorithm is used for feature selection.

II. Related Work

2.1 Project Challenges

This section discusses the challenges that were overcome to complete this project. Each of the challenges are briefly introduced and detailed in later sections of the report. This section can be useful for those who aim to do a deep learning project with large image datasets.

2.2 Large 3D Dataset Management and Preprocessing

Working with the dataset is the most difficult challenge that had to be overcome for this project. The dataset is very large (around 70 Gigabytes) which makes managing and analyzing the dataset and training the model very computationally heavy. Figure 1.1 shows one instance of a patient's CT Scan which is approximately 60 Megabytes.

A potential solution to this is to migrate the computation workload to a cloud service such as Google Cloud, AWS or Floyd hub which could result in a more efficient work flow for the project. A single CT scan is also 3 Dimensional which can be complex to work with especially during feature selection and data preparation.

2.3 Neural Network Architecture

Design Choices Neural networks are architectures that are designed in of itself and there are many types of architectures out there that exist to solve different problems. A key challenge of the author is to use an architecture that is able to find malignant tumor patterns in the data.

To overcome this, the author has to undertake necessary research to implement the correct model design prior to training. There are many neural network architectures such as multi-layer perceptron, convolutional networks and sequence models. Project Challenge This section discusses the challenges that were overcome to complete this project.

2.4 Doctor's Challenges during Diagnosis

This multidisciplinary team uses a variety of data, CT Scans, X-Rays, Pet Scans and Biopsies to assess whether a patient has lung cancer. These tests help the team to fully diagnose a patient and the approach would be to use all of them to gather data.

In 2016, 134 patients were diagnosed with lung cancer at Beaumont Rapid Access Clinic where 217 patients were diagnosed in total across all hospital services. This means that approximately 1/3 of all referrals at the clinic have lung cancer in Beaumont Hospital. Medical professionals use TNM classification to help characterize lung cancer from basic to advanced forms of malignant tumors. IA being the earliest stage of cancer which is more likely that it was accidentally discovered, the more difficult it will be to perform a biopsy and ultimately the better the prognosis (likelihood to survive).

On the other hand, IV is an advanced stage of cancer which means that it is easier to diagnose, including biopsy, highly likely to cause symptoms and the worse the prognosis.

2.5 Kaggle 3D Unlabeled Dataset:

Data Science Bowl 2017 This dataset was part of the Kaggle competition Data Science Bowl 2017 [4]. The topic of the competition was about lung cancer detection. The dataset was provided by the National Lung Cancer Screening Trial, The Cancer Imaging Archive, Diagnostic Image Analysis Group (Radboud University), Lahey Hospital and Medical Center and Copenhagen University Hospital.

The dataset contains full CT scan images of a patient's lungs. The dimension for this is (512,512, 200) which is (Height, Width, No. of Images). See Figure 2.4a. For this project the dataset has been used to segment different parts of the CT scans as part of feature engineering and visualizations. This dataset was what the author originally wanted to work with as the data was labeled as desired and useful for the project. However, the largest challenge that hindered the author from continuing using this data is the size of the entire dataset. The entire dataset is about 100GB zipped which could not fit on the authors laptop.

III. Literature Survey

3.1 Texture Analysis Based Feature Extraction and Classification of Lung Cancer

Lung cancer is most life-threatening disease, treatment of which must be the primary goal throughout scientific research. The early recognition of cancer can be helpful in curing disease entirely. There are numerous techniques found in literature for detection of lung cancer. Several investigators have contributed their facts for cancer prediction.

These papers largely pact about prevailing lung cancer detection techniques that are obtainable in the literature. A numeral of methodologies has been originated in cancer detection methodologies to progress the efficiency of their detection. Diverse applications like as support vector machines, neural networks, image processing techniques are extensively used in for cancer detection which is elaborated in this work.

The primary goal of pre-processing is to improve the quality of the image so that it can be used to remove or reduce irrelevant parts. To improve .the image's quality, the pre-processing stage is critical. Filters remove the noise and other segments with high recurrence and prepare the datasets for further processing. Lung cancer detection processing methods are depicted in.

3.2 Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier

We used an image processing technique in MATLAB to create an effective lung cancer detection and prediction algorithm. Lung cancer has been detected using multi-stage classification. Using this algorithm, lung cancer predictions have been made.

The algorithm then determines the likelihood of lung cancer if the input image does not contain any cells affected by cancer. The algorithm checks the corresponding stage of the cancer, such as the initial, middle, and final stage, if a cancer-affected cell has been identified. Several methods have been used to enhance and segment images prior to each stage of classification. Image enhancement techniques include contrast enhancement, color space transformation, and image scaling.

For segmentation, threshold and marker-controlled watershed based segmentation have been utilized. The proposed system's overall flowchart can be found here. We made gray level cooccurrence framework from picture and from this we computed surface measures from the GLCM. The element was considered for extraction are mean, standard deviation, entropy, RMS, fluctuation, smoothness, kurtosis, IDM, differentiate, relationship, vitality, homogeneity.

3.3 Lung Cancer Disease Diagnosis Using Machine Learning Approach

From the beginning to the present, the most intriguing area of medical research has been the investigation and study of lung diseases. A diagnosis system like this can only help reduce the risk of human life-threatening conditions by detecting malignant growths earlier.

Eventually, a few structures are proposed, but a great number of them are still just ideas. In the following philosophy, the performance of a neural network model is examined to address the common problem in therapeutic imaging applications of recognizing cancerous cells in image data.

A lung cancer identification framework based on AI and deep neural systems is developed in an effort to complete this task. The method relies on supervised learning, which has improved precision, particularly through the use of the deep learning mechanism. A strategy for classifying lung tumors is the CNN classification. Image acquisition, pre-preparation, enhancement, segmentation, feature extraction, and neural framework identification are among the various methods included in the framework. To put it succinctly, the machine learning approach has the potential to provide a once-in-a-lifetime opportunity to enhance low-cost decision support for lung cancer treatment.

3.4 CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images

The regions would assist the pathologist and significantly speed up the entire procedure. A completely automated method for detecting lung cancer in whole slide images of lung tissue samples is presented in this paper. Using a convolutional neural network (CNN), classification is carried out at the image patch level.

The performance of two CNN architectures—VGG and ResNet—is trained and compared. The obtained results indicate that a CNN-based approach may assist pathologists in diagnosing lung cancer.

Index Terms: digital pathology, lung cancer, and convolutional neural networks I. Introduction Lung cancer is the most common cancer-related death cause. Patients with this disease have a poor prognosis, with a 5-year survival rate of less than 20%.

Due to a diagnosis made at an advanced stage of the disease, the majority of patients have a poor prognosis. Early-stage patients have a significantly higher 5-year survival rate of over 70%. In [2], it is demonstrated that screening with low dose computed tomography (LDCT) reduces high-risk population mortality by 20%.

The significance of early detection and diagnosis, which has a significant impact on the outcome of treatment, is emphasized in these findings. Histopathological examination of tissues obtained through bronchoscopy is a standard procedure that is required for early diagnosis following the acquisition of tumor-suspected CT images. Pathologists perform biopsy tissue assessment, which is a time-consuming and error-prone task with a diagnostic accuracy of less than 80% [3]. It is essential for treatment selection to correctly classify the patient into the major histological subtypes of squamous carcinoma, adenocarcinoma, small cell carcinoma, and undifferentiated carcinoma. Since digital pathology scanners now produce high-resolution whole-slide images (WSIs) (up to 160 nm per pixel), it is now possible to use computer vision to automatically detect cancer in WSIs.

Convolutional neural networks (CNNs) are the method of choice right now because of their improved accuracy in a variety of computer vision tasks, including medical imaging [4]. A method for automatically detecting cancer cells in lung tissue WSIs is presented in this paper. In order to lessen the amount of computation required, the first step is to extract the tissue-rich WSI region, or ROI. CNN-based classification of image patches into tumor and normal classes follows next. In the context of the most recent Automatic Cancer Detection and Classification in Whole-slide Lung Histopathology (ACDC@LUNGHP), this task was proposed, and the preliminary findings are outlined in [5]. Other than this, there are no other papers that deal with CNN-based evaluation of lung cancer images. The structure of the paper is as follows: The method is described in Section 3, which follows a brief overview of the related work

in Section 2. Section 4 presents the results, and Section 5 provides a conclusion.

3.5 Deep Learning Methods for Lung Cancer Segmentation in Whole-slide Histopathology Images - the ACDC@LungHP Challenge

A pathologist's evaluation of biopsy tissue is the gold standard for diagnosing lung cancer. The diagnostic accuracy, on the other hand, is less than 80% [4]. Squamous carcinoma, adenocarcinoma, small cell carcinoma, and undifferentiated carcinoma are the most common histological subtypes of malignant lung disease. In order to make the best treatment decisions, it is essential to correctly evaluate these subtypes on biopsy.

However, there aren't enough qualified pathologists to meet the huge clinical needs, especially in countries like China, where there are a lot of lung cancer patients. Lung cancer screening with low-dose Computed Tomography was recently implemented in the United States thanks to the National Lung Screening Trial (NLST), the largest randomized control trial. Additionally, the Dutch-Belgian lung cancer screening trial (NELSON), the second-largest randomized control trial, demonstrates the advantages of screening for lung cancer. Whole-slide histopathology images, biopsies, and resected tumors are likely to result from the implementation of lung cancer screening in the United States and Europe. At the same time, there is a severe shortage of pathologists and a heavy workload. By automatically evaluating lung biopsies, an artificial intelligence (AI) system may effectively resolve the aforementioned issues.

IV. Software Description

4.1 Python

Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.It is used for:

- web development (server-side),
- software development,
- mathematics,
- System scripting.

Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc.).Python has a simple syntax similar to the English language. Python has syntax that allows developers to write programs with fewer lines than some other programming languages.

Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that

prototyping can be very quick. Python can be treated in a procedural way, an object-oriented way or a functional way. Flexible and powerful, Python was originally developed in the late 1980s at the National Research Institute for Mathematics and Computer Science by Guido van Rossum as a successor to the ABC language.
Since its introduction, Python has grown in popularity thanks to what is seen as a clear and expressive syntax developed with a focus on ensuring that code is readable. Python is a high-level language .
This means that Python code is written in largely recognizable English, providing the Pi with commands in a manner that is quick to learn and easy to follow. This is in marked contrast to low-level languages, like assembler, which are closer to how the computer
"thinks" but almost impossible for a human to follow without experience.
The high-level nature and clear syntax of Python make it a valuable tool for anyone who wants to learn to program. It is also the language that is recommended by the Raspberry Pi Foundation for those looking to progress from the simple Scratch.
Python is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options.

4.2 Anaconda Navigator
Anaconda is an open-source distribution of the Python and R programming languages for data science that aims to simplify package management and deployment. Package versions in Anaconda are managed by the package management system, Conda, which analyzes the current environment before executing an installation to avoid disrupting other frameworks and packages.
The Anaconda distribution comes with over 250 packages automatically installed. Over 7500 additional open-source packages can be installed from PyPI as well as the Conda package and virtual environment manager. It also includes a GUI (graphical user interface), Anaconda Navigator, as a graphical alternative to the command line interface. Anaconda Navigator is included in the Anaconda distribution, and allows users to launch applications and manage Conda packages, environments and channels without using command-line commands. Navigator can search for packages, install them in an environment, run the packages and update them.

4.3 NumPy
NumPy is a library in Python that allows for efficient numerical computing in Python. This library is highly optimized to do mathematical tasks. In the project workflow NumPy is heavily used in data pre-processing and preparation One of the main features about NumPy is its highly efficient n-dimensional array (Nd array). Compared to a list in Python a NumPy array can be dimensions and has more features associated with the Nd array. NumPy can also perform more efficient mathematical operations compared to the math library in Python.

4.4 Pandas
Pandas is also a library in Python, like NumPy is also used for data pre-processing and preparation. One of the main features about pandas is the Data Frame and Series data structure. These data structures are optimized and contain fancy indexing that allow a variety of features such as reshaping, slicing, merging, joining and etc. to be available. Pandas and NumPy are extremely powerful when used together for manipulating data.

4.5 Matplotlib
Matplotlib is a Python plotting library that allows programmers to create a wide variety of graphs and visualizations with ease of use. The great feature about Matplotlib is that it integrates very well with Jupyter Notebook and creating visualizations is simplified. Matplotlib also works very well with pandas and NumPy.

4.6 TensorFlow
TensorFlow is an open-source deep learning library by Google. It was originally developed by Google's engineers who were working on Google Brain and has been used for research on machine learning and deep learning. TensorFlow at its core is about computations of multidimensional arrays called tensors but what makes TensorFlow great is its ability to be flexible to deploy computations on different devices such as CPU's and GPU's .

4.7 Keras
Keras is also a Deep Learning Framework that abstracts much of the code in the other Frameworks like TensorFlow and Theano. Compared to the other frameworks Keras is more minimalist.

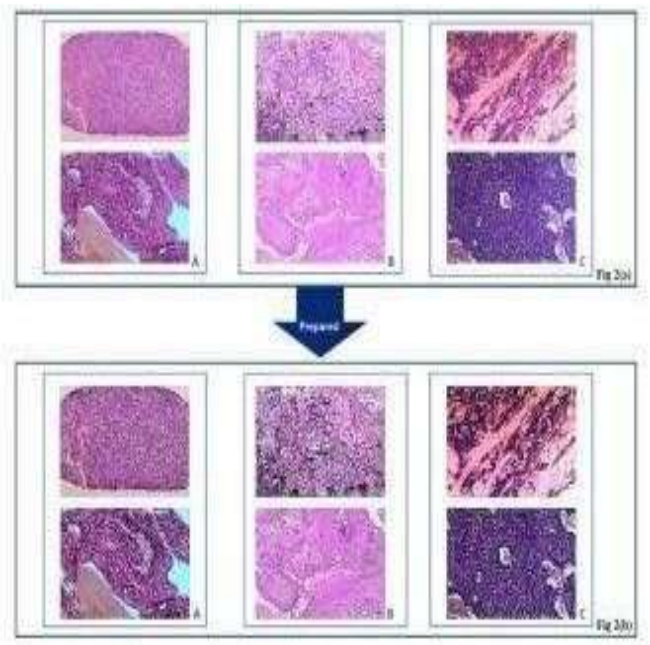
V. MODULES

5.1 Data Collection
Research on Auxiliary Classification and Diagnosis of LC Subtypes FIGURE 2. 2(a). The unprocessed histological images of LUSC, ASC and SCLC. 2(b). The processed histological images of LUSC, ASC and SCLC. To better display the image details, two pathological tissue images of each category were selected for display in this study, of which A was LUSC, B was ASC, and C was SCLC. 22 had LUSC, 27 had ASC and 45 had SCLC. All subjects had signed informed consent in this retrospective study.
For improving the generalization performance of the model, this study took an average of 12 pathological tissue sections from each patient as the research object, and finally selected 121 histopathological images [19], [20]. According to the classification standards provided by the World Health Organization, LC patients were collected from the hospital pathology database, and the relevant pathological diagnosis results were confirmed by pathology.

All tumor tissues included in the study were made into histopathological sections by hematoxylins and eosin (H&E) staining, which were collected through the microscope of the hospital pathology department and stored as JPG image files with four resolutions: 744 × 554, 2048 × 1536, 1024 × 768, 640 × 480 [6].

Histopathological Image Preprocessing Histopathological examination is the gold standard for qualitative and clinical tumor staging . Histopathological images have been widely used by doctors for diagnosis and treatment, and are an important basis for predicting patient survivals. The

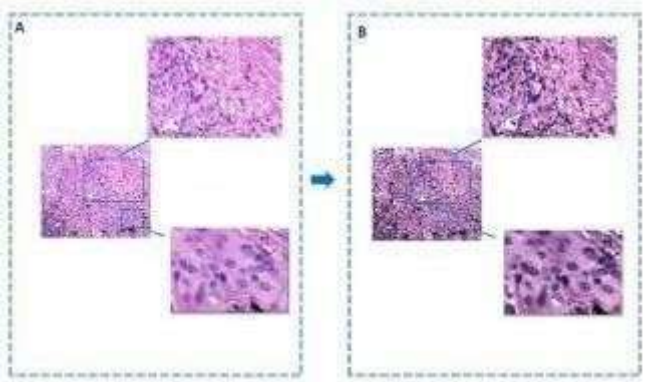
histopathological images of LUSC, ASC and SCLC. According to reports, the following problems exist in histopathological images: The histopathological images are faced with a large number of rich geometric .



5.1.1 Feature Extraction Of Histopathological Images Of Lung Cancer

Because the rich features presented in histopathological images are an important basis for clinicians to carry out diagnosis, the effective extraction of image features is the key to improving the accuracy Because the rich features presented in histopathological images are an important basis for clinicians to carry out diagnosis, the effective extraction of image features is the key to improving the accuracy

Extraction methods on LC histopathological image classification. 1) Handcrafted Texture Extraction Methods We extracted 265-dimensional features using seven handcrafted texture extraction algorithms, including the Hu invariant moments, GLGCM, wavelet transform, GLCM, LBP, GLDS and Markov random field, as shown in Table .



5.2 Machine Learning Algorithm

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

IBM has a rich history with machine learning. One of its own, Arthur Samuel, is credited for coining the term, “machine learning” with his research (PDF, 492 KB) (link resides outside IBM) around the game of checkers. Robert Nealey, the self-proclaimed checkers master, played the game on an IBM 7094 computer in 1962, and he lost to the computer. Compared to what can be done today, this feat almost seems trivial, but it’s considered a major milestone within the field of artificial intelligence.

Over the next couple of decades, the technological developments around storage and processing power will enable some innovative products that we know and love today, such as Netflix’s recommendation engine or self-driving cars. Machine learning is an important component of the growing field of data science.

Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

Machine learning algorithm into three main parts.

1. A Decision Process: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabeled, your algorithm will produce an estimate about a pattern in the data.

2. An Error Function: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.

3. An Model Optimization Process: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

5.4 Real-world machine learning use cases

Speech Recognition: It is also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, and it is a capability which uses natural language processing (NLP) to process human speech into a written format. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or provide more accessibility around texting.

Customer Service: Online chatbots are replacing human agents along the customer journey. They answer frequently asked questions (FAQs) around topics, like shipping, or provide personalized advice, cross-selling products or suggesting sizes for users, changing the way we think about customer engagement across websites and social media platforms. Examples include messaging bots on e-commerce sites with virtual agents, messaging apps, such as Slack and Facebook Messenger, and tasks usually done by virtual assistants and voice assistants.

Computer Vision: This AI technology enables computers and systems to derive meaningful information from digital images, videos and other visual inputs, and based on those inputs, it can take action. This ability to provide recommendations distinguishes it from image recognition tasks. Powered by convolutional neural networks, computer vision has applications within photo tagging in social media, radiology imaging in healthcare, and self-driving cars within the automotive industry.
Recommendation Engines: Using past consumption behavior data, AI algorithms can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.
Automated stock trading: Designed to optimize stock portfolios, AI-driven high frequency trading platforms make thousands or even millions of trades per day without human intervention.

VI. TESTING

After the source code has been completed, documented as related data structures. Completed the project has to undergo testing and validation where there is subtitle and definite attempt to get error. The project developer treads lightly, designing and execution test that will demonstrates that the program works rather than uncovering errors, unfortunately errors will be present and if the project developer doesn't find them, the user will find out. The project developer is always responsible for testing the individual units i.e., modules of the program.
In many cases developer also conducts integration testing i.e., the testing step that leads to the construction of the complete program structure. This project has undergone the following testing procedures to ensure its correctness./

6.1 Unit Testing

Unit testing is a method by which individual units of source code, sets of one or more computer program modules together with associated control data, usage procedures, and operating procedures are tested to determine if they are fit for use. Intuitively, one can view a unit as the smallest testable part of an application. In procedural programming, a unit could be an entire module, but is more commonly an individual function or procedure.

6.2 Integration Testing

Integration testing is a systematic technique for constructing the program structure. While at the same time conducting tests to uncover errors associated with interfacing. The objective is to take unit tested modules and build a program structure that has been dedicated by design.
In this integration testing its done using the main module and based on the type of integration testing the subordinate tables and other criteria along with their path, is replaced one at a time with actual modules.

6.3 Validation Testing

It is said that validation is successful when the software functions in a systematic manner that can be reasonably accepted by the customers. This type of testing is very important because it is the only way to check whether the requirements given by user have been completely fulfilled.
The input given to various forms are validated effectively. The validated input is given for all modules. Each module is tested independently.

6.4 White Box Testing

White-box testing (also known as clear box testing, glass box testing, transparent box testing, and structural testing) is a method of testing software that tests internal structures or workings of an application, as opposed to its functionality (i.e. black-box testing). In white-box testing an internal perspective of the system, as well as programming skills, are used to design test cases. The tester chooses inputs to exercise paths through the code and determine the appropriate outputs. This is analogous to testing nodes in a circuit, e.g. in circuit testing (ICT).
While white-box testing can be applied at the unit, integration and system levels of the software testing process, it is usually done at the unit level. It can test paths within a unit, paths between units during integration, and between subsystems during a system– level test. Though this method of test design can uncover many errors or problems, it might not detect unimplemented parts of the specification or missing requirements.

6.5 Black Box Testing

Black-box testing is a method of software testing that examines the functionality of an application (e.g. what the software does) without peering into its internal structures or workings (see white-box testing). This method of test can be applied to virtually every level of software testing: unit, integration, system and acceptance. It typically comprises most if not all higher level testing, but can also dominate unit testing as well.
Test procedures
Specific knowledge of the application's code/internal structure and programming knowledge in general is not required. The tester is aware of what the software is supposed to do but is not aware of how it does it. For instance, the tester is aware that a particular input returns a certain, invariable output but is not aware of how the software produces the output in the first place.
Test cases :
Test cases are built around specifications and requirements, i.e., what the application is supposed to do. Test cases are generally derived from external descriptions of the software, including specifications, requirements and design parameters. Although the tests used are primarily functional in nature, non-functional tests may also be used. The test designer selects both valid and invalid inputs and determines the correct output without any knowledge of the test object's internal structure.

6.6 Unit testing

In computer programming, unit testing is a method by which individual units of source code, sets of one or more computer program modules together with associated control data, usage procedures, and operating procedures are tested to determine if they are fit for use.
Intuitively, one can view a unit as the smallest testable part of an application. In procedural programming, a unit could be an entire module, but is more commonly an individual function or procedure. In object-oriented programming, a unit is often an entire interface, such as a class, but could be an individual method. Unit tests are created by programmers or

occasionally by white box testers during the development process.

Ideally, each test case is independent from the others. Substitutes such as method stubs, mock objects, fakes, and test harnesses can be used to assist testing a module in isolation. Unit tests are typically written and run by software developers to ensure that code meets its design and behaves as intended. Its implementation can vary from being very manual (pencil and paper)to being formalized as part of build automation.

Testing will not catch every error in the program, since it cannot evaluate every execution path in any but the most trivial programs. The same is true for unit testing. Additionally, unit testing by definition only tests the functionality of the units themselves. Therefore, it will not catch integration errors or broader system-level errors (such as functions performed across multiple units, or non-functional test areas such as performance).

Unit testing should be done in conjunction with other software testing activities, as they can only show the presence or absence of particular errors; they cannot prove a complete absence of errors. In order to guarantee correct behavior for every execution path and every possible input, and ensure the absence of errors, other techniques are required, namely the application of formal methods to proving that a software component has no unexpected behavior.

Software testing is a combinatorial problem. For example, every Boolean decision statement requires at least two tests: one with an outcome of "true" and one with an outcome of "false". As a result, for every line of code written, programmers often need 3 to 5 lines of test code.

This obviously takes time and its investment may not be worth the effort. There are also many problems that cannot easily be tested at all – for example those that are nondeterministic or involve multiple threads. In addition, code for a unit test is likely to be at least as buggy as the code it is testing. Fred Brooks in The Mythical Man-Month quotes: never take two chronometers to sea. Another challenge related to writing the unit tests is the difficulty of setting up realistic and useful tests. It is necessary to create relevant initial conditions so the part of the application being tested behaves like part of the complete system. If these initial conditions are not set correctly, the test will not be exercising the code in a realistic context, which diminishes the value and accuracy of unit test results.

To obtain the intended benefits from unit testing, rigorous discipline is needed throughout the software development process. It is essential to keep careful records not only of the tests that have been performed, but also of all changes that have been made to the source code of this or any other unit in the software. Use of a version control system is essential. If a later version of the unit fails a particular test that it had previously passed, the version-control software can provide a list of the source code changes (if any) that have been applied to the unit since that time.

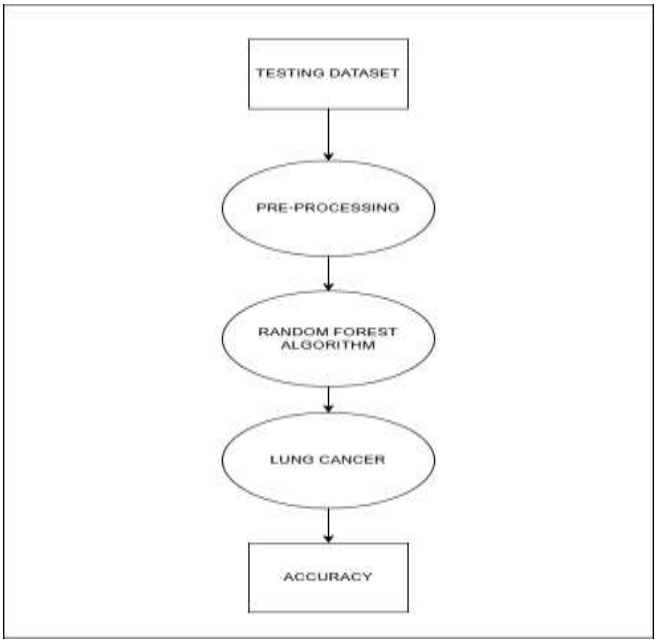
It is also essential to implement a sustainable process for ensuring that test case failures are reviewed daily and addressed immediately if such a process is not implemented and ingrained into the team's workflow, the application will evolve out of sync with the unit test suite, increasing false positives and reducing the effectiveness of the test suite.

Unit testing embedded system software presents a unique challenge: Since the software is being developed on a different platform than the one it will eventually run on, you cannot readily run a test program in the actual deployment environment, as is possible with desktop programs.

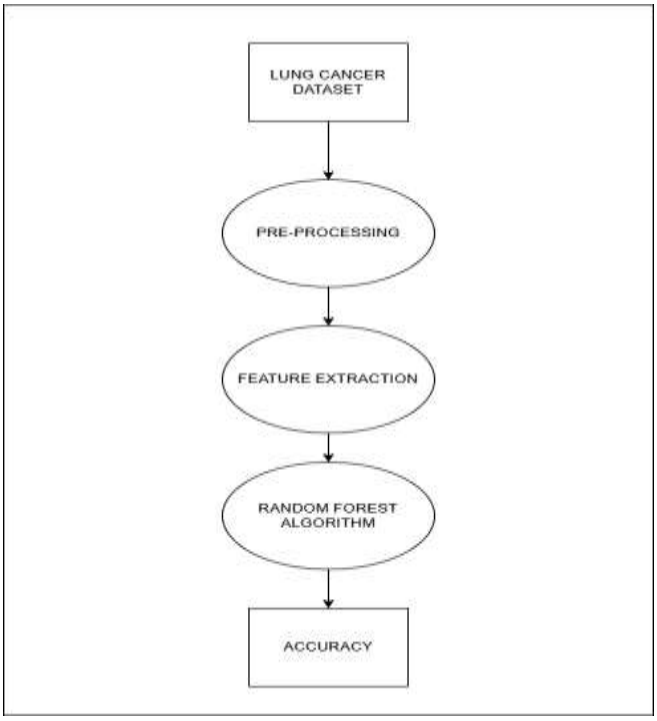
VII. Data Flow Diagram

Data flow diagrams illustrate how data is processed by a system in terms of inputs and outputs. Data flow diagrams can be used to provide a clear representation of any business function. The technique starts with an overall picture of the business and continues by analyzing each of the functional areas of interest. The technique exploits a method called top-down expansion to conduct the analysis in a targeted way.

DFD 0:



DFD 1:



7.3 Architecture Diagram

VIII. CONCLUSION

Currently, ML Algorithms play a significant role in early LC prediction, and with the help of these techniques available data can be used to make predictions or decisions. Study work provided a proposed system followed by MLa in predicting early LC which provides the researcher with better knowledge in ML Technique for early prediction of LC. Moreover, to make ML approaches easier in the field of LC, different types of the dataset used, various data preprocessing methods implemented, essential features that are been selected and extracted are been explained in detail. Also, the performance of different MLa is evaluated. The parameters used in constructing an efficient and accurate ML model for early prediction of Lung cancer is a piece of additional information. This study will help the researchers to identify ML techniques that produces more accuracy and efficiency .

8.1 Future Work

To construct an efficient and accurate ML model for early LC, the model can be developed with the following parameters Data should be collected from large and highly qualified authorized canter's. "e.g." www.cancerimagingarchive.net Data collected should be preprocessed by a powerful technique such that no important data is lost. Highly correlated Features with the output should be identified for best results. Using the Hybrid ML model, early prediction of LC can produce accurate results. Several ML tools and various platforms can be made available for researchers to provide good results. There are also many data analytical tool that can provide useful information for future data analysis.

REFERENCES

[1] Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre L A, Jemal A, "Global Cancer Statistics 2018", doi: 10.3322/caac.21492

[2] V Noronha , R Dikshit, N Raut, A Joshi, C S Pramesh, K George, J P Agarwal, Munshi A, Kumar P, " Epidemiology of lung cancer in India: Focus on the differences between non-smokers and smokers", 2012. volume:49, Page: 74-81.

[3] Cheung LC, Katki H A, Chaturvedi A K, Jemal A, Berg C D, "Preventing Lung Cancer Mortality by Computed Tomography Screening: The Effect of Risk-Based Versus U.S. Preventive Services Task Force Eligibility Criteria", doi:10.7326/M17-2067.

[4] Aberle D R, Adams A M, Berg C D, et al , " Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening", doi: 10.1056/NEJMoa1102873.

[5] Koning H J D, Meza R, Plevritis S K, Haaf K T, Munshi V N, Jeon J, et.al, "Benefits and Harms of Computed Tomography Lung Cancer Screening Strategies: A Comparative Modeling Study for the U.S. Preventive Services Task Force", 2014; 160(5):311-20.