Multi-Stage Lung Cancer Classification Using Machine Learning

Syed Mehveen[1], D.Murali[2]
[1]PG Student, QUBA College of engineering and technology
[2]Associate Professor, QUBA College of engineering and technology

Abstract: Cancer is the uncontrollable cell division of abnormal cells inside the human body, which can spread to other body organs. It is one of the non-communicable diseases (NCDs) and NCDs account for 71% of total deaths worldwide whereas lung cancer is the second most diagnosed cancer after female breast cancer. The cancer survival rate of lung cancer is only 19%. There are various methods for the diagnosis of lung cancer, such as X-ray, CT scan, PET-CT scan, bronchoscopy, and biopsy. However, to know the subtype of lung cancer based on the tissue type H and E staining is widely used, where the staining is done on the tissue aspirated from a biopsy. Studies have reported that the type of histology is associated with the prognosis and treatment of lung cancer. Therefore, early and accurate detection of lung cancer histology is an urgent need and as its treatment is dependent on the type of histology, molecular profile, and stage of the disease, it is essential to analyze the histopathology images of lung cancer. Hence, to speed up the vital process of diagnosis of lung cancer and reduce the burden on pathologists, Deep learning techniques are used. These techniques have shown improved efficacy in the analysis of histopathology slides of cancer. Several studies reported the importance of convolution neural networks (CNN) in the classification of histopathological pictures of various cancer types such as brain, skin, breast, lung, and colorectal cancer. In this study, the tri-category classification of lung cancer images (normal, adenocarcinoma, and squamous cell carcinoma) are carried out by using ResNet 50, VGG-19, Inception_ResNet_V2, and DenseNet121 for the feature extraction and triplet loss to guide the CNN such that it increases inter-cluster distance and reduces intra-cluster distance.

Index Terms— ResNet 50, CNN, VGG-19, Inception_ResNet_V2 and DenseNet121, Histopathology Images.

1. INTRODUCTION

1.1 OVERVIEW OF THE PROJECT

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", to make predictions or decisions without being explicitly programmed to perform the task.

Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory, and application domains to the field of machine learning. Data mining is a field of study within machine learning and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Reinforcement learning algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment and are used in autonomous vehicles or in learning to play a game against a human opponent. Other specialized algorithms in machine learning include topic modeling, where the computer program is given a set of natural language documents and finds other documents that cover similar topics. Machine learning algorithms can be used to find the unobservable probability density function in density estimation problems.

1.2 SCOPE AND OBJECTIVE SCOPE This project focuses on lung cancer diagnosis and prediction using CNN algorithms. Machine learning algorithms CNN can be trained by feeding it the training data and then the algorithm can predict the data by comparing the given data with the training data. Our goal is to train our algorithm by feeding it with training data. Our goal is to diagnose cancer using different types of input. OBJECTIVE Classification of lung cancer using CNN focuses on the classification of lung cancer. Histopathological images of the lung tissue will be used as input. Once the histopathology image becomes an input test, the image goes through several stages to increase image clarity for better results for classification. Using specific descriptors, it classifies the stage of lung cancer as benign or malignant. It is classified according to the size of the nodules formed in the lungs. The accuracy of the results is increased with CNN. It takes less time for the algorithm to calculate and establish the stage of cancer.

II SYSTEM ANALYSIS

2.1 LITERATURE SURVEY

A literature survey is the most important step in the software development process. Before developing the tool it is necessary to determine the time factor, economy, and company strength.

TITLE1: Lung cancer detection system using lung CT image processing AUTHOR: Amita Dessai

Cancer is the root cause of a large number of deaths worldwide, out of which lung cancer is the cause of the highest mortality rates. Computer tomography scan is employed by radiologists to detect cancer in the body and track its growth. Visual interpretation of databases can lead to cancer detection at later stages, thus leading to late treatment of cancer which only boosts up the cancer death rates.

TITLE 2: Sex and Smoking Status Effects on the Early Detection of Early Lung Cancer in High-Risk Smokers Using an Electronic Nose AUTHOR: Calum E. MacAulay

Volatile organic compounds (VOCs) in exhaled breath as measured by an electronic nose (e-nose) have utility as biomarkers to detect subjects at risk of having lung cancer in a screening setting.

III SYSTEM DESIGN

3.1 INTRODUCTION

Design is a multi-step that focuses on data structure software architecture, procedural details, algorithms, etc… and an interface between modules. The design process also translates the requirements into the presentation of software that can be accessed for quality before coding begins. Computer software design change continuously as new methods; better analysis and border understanding evolved. Software design is at a relatively early stage in its revolution.

Therefore, software design methodology lacks the depth, flexibility, and quantitative nature that are normally associated with more classical engineering disciplines. However, techniques for software designs do exist, criteria for design qualities are available and design notation can be applied.

3.2 EXISTING SYSTEM

- The Existing system is focused on developing CAD systems for lung cancer detection.
- In hospitals, to detect lung cancer, patients generally undergo a lung examination using the lung surface microscopy technique commonly known as dermoscopy.
- Imaging tests, an X-ray image of your lungs may reveal an abnormal mass or nodule.
- Sputum cytology, if anyone has a cough and has sputum examined under the microscope can sometimes reveal the presence of lung cancer cells.

3.3 EXISTING SYSTEM DISADVANTAGES

- Using a CAD system, the nodule size results in less accuracy.
- Using SVM, the accuracy of results is less the rate is about 55%.

PROPOSED SYSTEM

- To improve the accuracy of feature extraction, CNN algorithms were used.
- Feature extraction is used to determine the size of the nodule.
- It is also used to classify the stage of lung cancer.
- The system takes histopathological images as input and predicts the output.

- It uses algorithms like ResNet 50, VGG-19, Inception_ResNet_V2, and DenseNet121 and filters to enhance the image quality for better results.

PROPOSED SYSTEM ADVANTAGES

- Using a CNN algorithm with feature extraction used to determine the state and stage of lung cancer.
- The accuracy of the system is about 80% to detect lung cancer provided with the dataset. network.

3.3 HISTOPATHOLOGICAL IMAGES

Histopathological images are images of tissue samples that have been stained and prepared for microscopic examination. These images are widely used in the field of pathology for diagnosing diseases, evaluating the severity of diseases, and understanding the microscopic characteristics of various tissues and organs.

FIG 3.1 Histopathological image

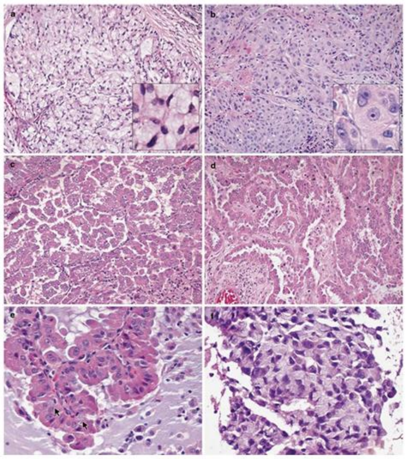


FIG 3.2 Different types of histopathological images of cancer of lung tissues

IV.IMPLEMENTATION AND ANALYSIS

4.1 MACHINE LEARNING

Machine Learning is a subfield of computer science, but is often also referred to as predictive analytics or predictive modeling. Its goal and usage are to build new or leverage existing algorithms to learn from data. To build generalized models that give accurate predictions or to find patterns, particularly with new and unseen similar data.

4.1.1 MACHINE LEARNING CLASSIFICATIONS

Although supervised and unsupervised learning are two of the most widely accepted machine learning methods by businesses today, there are various other machine learning techniques. Following is an overview of some of the most accepted ML methods.

Supervised Learning

These algorithms are trained using labeled examples, in different scenarios, as input where the desired outcome is already known. Equipment, for instance, could have data points such as "F" and "R" where "F" represents "failed" and "R" represents "runs".

Unsupervised Learning

This method of ML finds its application in areas where data has no historical labels. Here, the system will not be provided with the "right answer" and the algorithm should identify what is being shown. The main aim here is to analyze the data and identify a pattern and structure within the available data set. Transactional data serves as a good source of data sets for unsupervised learning.

Semi-supervised Learning

This kind of learning is used and applied to the same kind of scenarios where supervised learning is applicable. However, one must note that this technique uses both unlabelled and labeled data for training. Ideally, a small set of labeled data, along with a large volume of unlabelled data is used, as it takes less time, money, and effort to acquire unlabelled data. This type of machine learning is often used with methods, such as regression, classification, and prediction. Companies that usually find it challenging to meet the high costs associated with labeled training processes opt for semi-supervised learning.

Reinforcement Learning

This is mainly used in navigation, robotics, and gaming. Actions that yield the best rewards are identified by algorithms that use trial and error methods. There are three major components in reinforcement learning, namely, the agent, the actions, and the environment. The agent in this case is the decision maker, the actions are what an agent does, and the environment is anything that an agent interacts with. The main aim of this kind of learning is to select the actions that maximize the reward, within a specified time. By following a good policy, the agent can achieve the goal faster.

4.2 IMAGE

An image is an array or a matrix of square pixels arranged in columns and rows. An image is an artifact, for example, a two-dimensional picture, which has a similar appearance to some subject usually a physical object or a person.

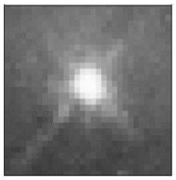


FIG 4.1 - An Image - An Array or a Matrix of Pixels Arranged in Columns and Rows.

4.2.1 PIXEL

Image processing is a subset of the electronic domain where the image is converted to an array of small integers, called pixels, representing a physical quantity such as scene radiance, stored in a digital memory and processed by a computer or other digital hardware.

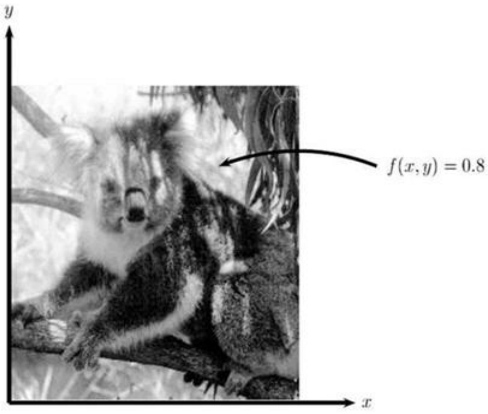
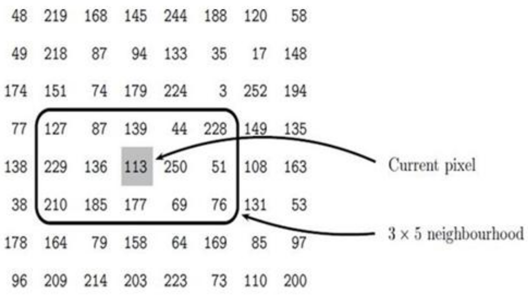


FIG 4.2 - A Grey Scale Image



48	219	168	145	244	188	120	58
49	218	87	94	133	35	17	148
174	151	74	179	224	3	252	194
77	127	87	139	44	228	149	135
138	229	136	113	250	51	108	163
38	210	185	177	69	76	131	53
178	164	79	158	64	169	85	97
96	209	214	203	223	73	110	200

FIG 4.3 - Pixel representation

Supervised classification

Supervised classification uses the spectral signatures obtained from training samples to classify an image. With the assistance of the Image Classification toolbar, you can easily create training samples to represent the classes you want to extract. You can also easily create a signature file from the training samples, which is then used by the multivariate classification tools to classify the image.

Unsupervised classification

Unsupervised classification finds spectral classes (or clusters) in a multiband image without the analyst's intervention. The Image Classification toolbar aids in unsupervised classification by providing access to the tools to create the clusters, the capability to analyze the quality of the clusters, and access to classification tools.

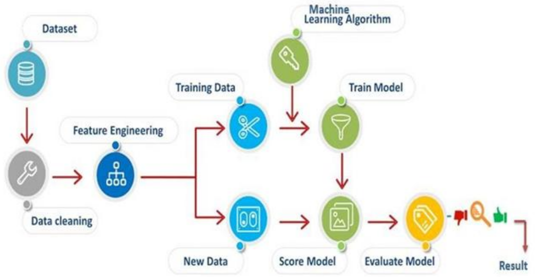


FIG 4.13 – Architecture of Unsupervised Classification

4.4 PYTHON TECHNOLOGY

Python is an interpreted, object-oriented programming language similar to PERL, that has gained popularity because of its clear syntax and readability. Python is said to be relatively easy to learn and portable, meaning its statements can be interpreted in several operating systems including UNIX-based systems, Mac- OS, MS-DOS, OS/2, and various versions of Microsoft Windows 98. Python was created by Guido

Van Rossum, a former resident of the Netherlands, whose favorite comedy group at the time was Monty Python's flying circus. The source code is freely available and open for modification and reuse.

Python has a significant number of users. A notable feature of Python is its intention of source statements to make the code easier to read. Python offers Dynamic datatype, ready-made classes, and interfaces to many systems calls and libraries. It can be extended using the C or C++ language.

V. CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Machine learning systems and its application are extending gigantically across the world in every sector like the Medical industry to suggest doctors, to automate the pharmaceuticals industry. In the field of Computer Science used to automate websites and provide suggestions based on search history and past clicks. Machine learning obtains its knowledge from experts and stores it in the feed. From experts, it develops its database and shares its knowledge with other machines or it may have a central repository. Now a day most of the medical diagnosis is done by using machine i.e. computer to obtain faster results and reduce human error in peculiar cases and conditions.

5.2 FUTURE ENHANCEMENT

Currently, the project uses only the command line interface (CLI) on all operating systems. In the future, deeper learning models with greater capacity and new design methods may be developed to better capture complex patterns in histopathological images accurately and increase the reliability of various stages of lung cancer. Future developments may focus on improving AI descriptive methods to provide clear and meaningful explanations for the prediction of lung cancer classification models. This helps build trust and confidence in doctors' and patients' standards and facilitates their adoption in the hospital.

REFERENCE

[1] Recent Advances in Classification and Diagnosis of Lung Cancer" by Jin et al. (2021)

[2] "Machine Learning and Deep Learning Approaches for Lung Cancer Classification: A Survey" by Zhang et al (2021)

[3] "Lung Cancer Classification by Machine Learning Algorithms: A Systematic Review" by Shahriyari et al (2020)

[4] Agarwal, N., Balasubramanian, V. N., & Jawahar, C.V. (2018). Improving multiclass classification by deep networks using DAGSVM and Triplet Loss. Pattern Recognition Letters, 112, 184–190.

[5] N. Savage, "How AI is improving cancer diagnostics," Nature News, 25-Mar 2020. [Online]. Available: https://www.nature.com/articles/d41586-020-00847-2. [Accessed: 13-Jun 2020].

[6] Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000).[Dataset]. Available: https://www.kaggle.com/andrewmvd/lung-and-colon-cancer histopathological-images [Accessed: 18 May 2020].

[7] Histology of the Lung. YouTube, 2016.

[8] https://ieeexplore.ieee.org/document/9793061

[9] https://tlcr.amegroups.com/article/view/21998/html

[10] https://www.sciencedirect.com/science/article/pii/S2405959520304732

[11] https://beei.org/index.php/EEI/article/view/4579

[12] https://www.nature.com/articles/s41598-023-29656-z