International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 7 July 2023, pp: 758-770 www.ijaem.net ISSN: 2395-5252

Image-based Histological Features for Lung Cancer Prediction using OptimizedStochastic Learning Model

S.Arif Abdul Rahuman, M. Ashika Nooriya, A. Syed Ali Fathima, F. Aaliya

H.O. D/In Aalim Mu Ham M Ed Salegh Engineering Co Llege Tam I Lnadu, India
2student/ T Aalim Muham Med Saleg H College Tamilnadu, India
Student/!T Aalim Muham Med Salegh College Tamilnadu, India
‘Stu Dent/It Aalim Mu Ham M Ed Salegh College Tam Lnadu, India

Date of Submission: 20-07-2023 Date of Acceptance: 31-07-2023

ABSTRAC T

Acct rding to GHO (Global Health Obseri a tory (GHO) the hich i»re›’alcnce c f a large ›’aricty c f diseases such as Ischaemic heart disease strc ke lund cancer disease ancl lou’cr res| iratcay‘ infectic ns hai’e remained the t i» killers during the past decade. The erou’t1i in the number of mortalities caused by these disease is due to the x re \ delayed sjinptomsdetection. Since in the earlj’ stages the symptoms are insienillficant and similar to those of benign diseases (e.p. the fll4u) we can only defect the ilisease at ad› anced stage. In additic n The high frequency o1“ improper practices that are harmful to health the hereditary factors and the stressful living conditions can increase the death rates,

The sj stain pr i»osed consists of se eral steps including acquiring the ima ge i»reprocessing binarizalion thresholdine and segmentation extraction of features and detection of Optimized Stochastic Learning Model. Detection of the lung C“T image is carried out to extract any significant feature of a classification image and a specific feature extractic n metlic d is implemented.

Scope of the **Project**

- Data Analysis
- Data Preprocessing
- Training the Model
- Testing of Test Data

I. INTRODUCTION

lune cancer (LC) is one o1“ the most common lRalipnant tumors worldwide According to the 2018 Interilatic na1 Agency for Research on Cancer statistics, there will be 2.1 million new cases of LC’ and 1,8 million deaths worldwide . Due to its high morbidit y and mortality. it has become one of the mc sr serious cancers threatening human health.

Clinicians visual analysis of LC histopatliological imaees is one c f the most iinpc rtant methods for evaluating LC subt›'i»cs . Hov’ever, it is complicated and challenging for paiholooists tc review thousands of liistcapathc»logical iina res. and ii is ci’en more difficult for Pc ctors i’ith less experience Therefore. to relieve the i»ressure on doctors and improve tileaccuracy e1’ficiency o1“ dia enosis. it is particularly important to study the com} uter-aided diagnosis mc del of LC. From the }nerspectix e pathology and treatment, LC can be divided into non-small cell lungs carcinoma (NSC LC) and small cell lung carcinoma (SC‘LC), cot hich 80°.fi-85°. 3 are NSCLC’ and rhe rest are SCLC The main histological types c fNSCLC’ are lune adenc carcinoma (ADC’) and lung squamous cell carcinoma (LUSC). The other histological rjpes c f

NSCLC’ are lung adenosquamous carcinoma (ASC’). large-cell carcinoma In particular. ASC is a relatively' rare subt5'pe of NSC’LC that aces unts fear 0.3aC“ñ°fi of all NSC LC s . Due to the diffferent histopatholopical tj i»es of LC, the treatment inetlioils adopted are also diff1“1“erent. 'Julien the lung tissue classification is deteriiiinecl.

the • i»i»ropriate treatment mode can be selected. such as the reasonable application of surgery. chemotherapy, radio thera}ay, mcalecular targeted therapy and immune therapy , In addirion. LC screenings errs rs can be a›’caided, clinicians multifarious wcark |aressure can be sly wed. parients**' suri’iva1 time can be maximized anal the patients quality c f li1’e improved.

Lung Cancer (LC) is the deaclliest and least-funded cancer worldwide . Non-small-cell LC is the major type o1“ LC Luna Adenc carcincoma (LL'AC’) is the most pre› alent liistc logic subtype .Lung nodules manifesting as Grcaund Glass(GG)Subscalid Nodules (SSNs) on Computed Tc mopraphy (C’T) scans ha› e a higher risk of lRalignancy other incidentally detected small solid nodules.

SSNs are often diagnosed as adenocarcinoma and are generally classified into pure GG nodules and part-solid nodules according to their appearance on the lung window settings . A timely and accurate attempt to differentiate the LUACs is of utmost importance to guide a proper treatment plan, as in some cases, a pre-invasive or minimally invasive SSN can be monitored with regular follow-up CT scans, whereas invasive lesions should undergo immediate surgical resection if they are deemed eligible. Most often, the SSNs type is diagnosed based onthe pathological findings performed after surgical resections, which is not desired for prior treatment planning. Currently, radiologists use chest CT scans to assess the invasiveness of the SSNs based on their imaging findings and patterns prior to making decisions regarding the appropriate treatment. Such visual approaches, however, are time consuming, subjective, and error-prone. So far, many studies have used high-resolution and thin- slice (< 1.5mm) CT images for the SSN classification, which require longer analysis times, as well as more reconstruction time . However, lung nodules are mostly identified from CT scans performed for varied clinical purposes acquired using routine standard or low-dose scanning protocols with non-thin slice thicknesses (up to 5mm) . In addition, recent lung cancer screening recommendation, suggests using low-dose CT scans with thicker slice-thicknesses (up to 2.5mm) . Capitalizing on the above discussion, the necessity of developing an automated invasiveness assessment framework that performs well regardless of technical settings has recently arisen among the research community and healthcare professionals.

Lung cancer is one of the deadliest diseases, with low long- term survival rates. Its treatment is still very heuristic since patients respond differently to the same treatment plans. Therefore, patient-specific models for predicting tumor growth and the treatment response are necessary for clinicians to make informed decisions about the patients therapy and avoid a trial and error based approach. We make a small step in that direction by introducing a model for simulating cancer growth and its treatment inside a 3D lung geometry. In this model, we represent tumor cells by a volume fraction field that varies over space and time. We describe their evolution by a system of partial differential equations, which include patient- and treatment-specific parameters capturing the different responses of patients to the therapies. Our simulation results are compared to clinical data and show that we can quantitatively describe the tumors behavior with some parameter set. This enables us to change therapies and analyze how these changes could have impacted the patients health. Lung cancer is the leading cause of cancer death in the world . In patients with epidermal growth factor receptor (EGFR) mutated tumors, the inhibitors of tyrosine kinase receptor (TKIs) have been shown to significantly improve overall survival . Accurate determination of the EGFR mutation status is highly relevant for the proper treatment of this patients.

The genetic test to determine the mutation of the EGFR involves an invasive procedure in obtaining the tumor tissue sample with risk of complications. In addition, its availability is limited in some regions of the world .Medical images are the most used diagnostic modality in cancer patients, provides information about diagnosis and prognosis. Advances in technology have allowed for improved image resolution, standardization of protocols, and global availability . Medical images are the most used diagnostic modality in cancer patients, provides information about diagnosis and prognosis. Advances in technology have allowed for improved image resolution, standardization of protocols, and global availability .The radiomics features analysis is a non- invasive methodology that converts imaging into high dimensional data, through automatic feature extraction, and has shown a good correlation with histological subtypes of tumors and genetic status in several pathologies, including lung cancer and squamous cell carcinomas of the head and neck. Lung cancer is the second most common cancer in the world, accounting for over 1.6 million deaths annually.1 Approximately 85% of lung cancer is classified as non-small-cell lung cancer (NSCLC).2 The Cancer Genome Atlas (TCGA), a collaborative effort organized by the National Cancer Institute (NCI), published high-quality profiling data on multiple cancer types, including NSCLC.

The resulting rich data provide a major opportunity to uncover the underlying genetic factors of cancers. From the viewpoint of statistical analysis, however, it is a challenging task to identify the markers associated with outcomes and phenotypic variances of NSCLC. With lung cancer being the leading cause of cancer death and a low prevalence of malignancy in pulmonary nodules (1.1% to 12%) , the management of screen-detected pulmonary nodules has become a substantial public health problem. Although lung cancer screening with annual computed tomography is now the standard of care for high-risk patients, the absence of highly accurate and noninvasive diagnostics leads to increased costs, anxiety, morbidity from unnecessary diagnostic procedures of benign lesions, and death from missed malignancy . As such, the use of predictive models to distinguish malignant from benign pulmonary nodules is an area of intense investigation. Deep learning approaches have been moderately successful in estimating the probability of malignancy from medical images. Liao et al.5 utilized a two- step strategy of first identifying pulmonary nodules with a region proposal network and second estimating their malignant potential with a leaky noise-or network. However, this approach in modeling cross-sectional CT scans is different than a clinicians approach to pulmonary nodules as consideration of how nodule features change over time, when possible, is highly predictive of malignancy. In lung cancer specifically, studies have consistently found nodule growth rate to be most closely associated with malignancyout of all studied features.

Motivation

lung cancer is the leading cause of cancer death in the world. Accurate determination of the EGFR (epidermal growth factor receptor) mutation status is highly relevant for the proper treatment of this patients. The survival rate for lung cancer patients is very low compared to other cancer patients due to late diagnostics. Thus, early lung cancer diagnostics is crucial for patients to receive early treatments, increasing the survival rate or even becoming cancer-free. This motivates us to develop this project.

Problem Statement

As a new paradigm, combining ML with physicians is the key route to enhance system efficiency, performance, and reliability. Especially in biomedical applications, if the diagnostic characteristics relied on decipherable diagnostic basis, human expert can support to improve identification and prediction in the performance. Otherwise, consumers may be more reluctant to utilize medical care delivered by AI providers than comparable human providers. However, the current deep ML methods, particularly deep neural networks, suffer from the notorious a anon interpretability issue. A crucial consequence of non-interpretability is the lack of systematic access to the mathematical properties of the ML models, including the representation and generalization abilities, as well as the stability and reliability of the outcomes.

Objectives

To optimize the network parameters so that the lossfunction approaches the global minimum.

To utilize more information than using a patch-basedneural network.

- To randomly shuffle the data points in the training set.
- To extract basic visual patterns like edges, lines andcorners.
- To combine the higher layer to form abstractinterpretations.

Domain Overview

Deep learning has evolved hand-in-hand with the digital era, which has brought about an explosion of data in all forms and from every region of the world. This data, known simply as big data, is drawn from sources Eke social media, internet search engines, e-commerce platforms, and online cinemas, among others. This enormous amount of data is readily accessible and can be shared through fintech applications like cloud computing. Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. achieving results that were not possible before .In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound.

Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers. However, the data, which normally is unstructured, is so vast that it could take decades for humans to comprehend it and extract relevant information. Companies realize the incredible potential that can result from unraveling this wealth of information and are increasingly adapting to AI systems for automated support.

About Parulas

Pandas is a popular Python package for data science, and with good reason: it offers powerful, expressive and flexible data structures that make data manipulation and analysis easy, among many other things. The Data Frame is one of these Pandas a high-level data manipulation tool developed by Wcs McKinney. It is built on the Numpy package and its key data structure is called the Data Frame. Data Frames allow you to store and manipulate tabular data in rows of observations and columns of variables. Pandas is built on top of the NumPy package, meaning a lcit of the structure of NumPy is used or repbcated in Pandas. Data in pandas is often used to feed statistical analysis in SciPy, plotting functions from Matplotlib, and machine learning algorithms in Scikit-learn. Jupiter Notebooks offer a good environment for using pandas to do data exploration and modeling, but pandas can also be used in text editors just as easily. Jupyter Notebooks give us the ability to execute code in a particular cell as opposed to running the entire file. this saves a lot of time when working with large and transformations. Notebooks also provide

an easy way to visualize pandas Data Frames and plots. As a matter of fact, this article was created entirely in a Jupyter Notebook.

There are two types of data structures in pandas: Series and Data Frames.

1. Series: a pandas Series is a one dimensional data structure (one dimensional and array) that can store âC" and for every value it holds a unique index, too.
2. Data Frame: a pandas Data Frame is a two (or more) dimensional data structure âC" basically a table with rows and columns. The columns have names and the rows have indexes. Those who are familiar with R know the data frame as a way to store data in rectangular grids that can easily be overviewed. Each row of these grids corresponds to measurements or values of an

   while each column is a vector containing data for a specific variable. This means that a data frames rows do not need to contain, but can contain, the same type of values. they can be numeric, character, logical, etc. Now, Data Frames in Python are very similar: they come with the Pandas library, and they are defined as two-dimensional labeled datawith of potentially different types.

In general, you could say that the Pandas DataFrame consists of three main components: the data, the index, and the columns.
Firstly, the DataFrame can contain at that is:
A Pandas Data Frame : A one-dimensional labeled array capable of holding any data type with axis labels or index

- An example of a Series object is one column from a DataFrame.
- A NumPy and **array,** which can be a record or structureda two-dimensional and **array**
- Dictionaries of one-dimensional and arrays, lists,dictionaries or Series.
- Some of the key features of Python Pandas are as follows:
- It provides Data Frame objects with default and customized indexing which is very fast and efficient.

There are tools available for loading data of different file formats into in-memory data objects.

- It is easy to perform data alignment and integratedhandling of missing data in Python Pandas.
- It is very simple to perform pivoting and reshaping ofdata sets in Pandas.
- It also provides indexing, label-based slicing, and sub-setting of large data sets.

We can easily insert and delete columns from a datastructure.
At a aggregation and transformations can be done usinggroup by.
Data aggregation and transformations can be done usinggroup by.
High-performance merging and joining of data can bedone using Pandas.

- It also provides time series functionality. Inserting and deleting columns in data structures.Merging and joining data sets.

Reshaping and pivoting data sets.

Aligning data and dealing with missing data. Manipulating data using integrated indexing for DataFrame objects.

Performing split-apply-combine on data sets using thegroup by engine.

Manipulating high-dimensional data in a data structure with a lower dimension using hierarchical axis indexing. Subnetting, fancy indexing, and label-based slicing data sets that are large in size.

Generating data range, converting frequency, date shifting, lagging, and other time-series functionality.

Reading from files with CSV, XLSX, TXT, among other formats.

Arranging data in an order ascending or descending.Filtering data around a condition.

Analyzing time series. Iterating over a data set

Numpy

Numpy is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. If you are already familiar with MATLAB, you might find this tutorial useful to get started with Numpy. A Numpy array is a grid of values, all of the same type, and is indexed by a tuple of non- negative integers.

The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension. NumPy is, just like SciPy, Scikit-Learn, Pandas, etc. one of the packages that you just can at miss when you are learning data science, mainly because this library provides you with an array datastructure that holds some benefits over Python lists, such as: being more compact, faster access in reading and writing items, being more convenient and more efficient.

NumPy is a Python library that is the core library for scientific computing in Python. It contains a collection of tools and techniques that can be used to solve on a computer mathematical models of problems in Science and Engineering. One of these tools is a high-performance multidimensional array object that is a powerful data structure for efficient computation of arrays

and matrices. To work with these arrays, there as a vast amount of high-level mathematical functions operate onthese matrices and arrays.

An array is basically nothing but pointers. It as a combination of a memory address, a data type, a shape, and strides:

The data pointer indicates the memory address of the fJirst byte in the array,

- The data type or dtype pointer describes the kind of elements that are contained within the array,
- The shape indicates the shape of the array, and
- The strides are the number of bytes that should be skipped in memory to go to the next element. If your strides are (10,1),
- you need to proceed one byte to get to the next column and 10 bytes to locate the next row

Or, in other words, an array contains information about the raw data, how to locate an element and how to interpret an element. With NumPy, we work with multidimensional arrays. We all dive into all of the possible types of multidimensional arrays later on, but for now, we all focus on 2-dimensional arrays. A 2-dimensional array is also known as a matrix, and issomething you should be familiar with. In fact, it as just adifferent way of thinking about a list of lists. A matrix hasrows and columns. By specifying a row number and acolumn number, we are able to extract an element from a matrix. We can create a NumPy array using the Numpy array function. If we pass in a list of lists, it will automatically create a NumPy array with the same number of rows and columns. Because we want all of the elementsin the array to be float elements for easy computation, weall leave of the header row, which contains strings. One of the limitations of NumPy is that all the elements in an array have to be of the same type, so if we include the header row, all the elements in the array will be read in as strings. Because we want to be able to do computations like find the average quality of the wines, we need the elements to all be floats. NumPy has several advantages over using core Python mathematical functions, a few of which are outlined here:

1. NumPy is extremely fast when compared to core Python thanks to its heavy use of C extensions.
2. Many advanced Python libraries, such as Scikit-Learn, Scipy, and Keras, make extensive use of the NumPy library. Therefore, if you plan to pursue a career in data science or machine learning, NumPy is a very good tool to master. NumPy comes with a variety of

built-in functionalities, which in core Python would take a fair bit of custom code.

Mathplotlib

Plotting of data can be extensively made possible in aninteractive way by Matplotlib, which is a plotting library that can be demonstrated in Python scripts. Plotting of graphs is apart of data visualization, and this property can be achieved by making use of Matplotlib. Matplotlib makes use of many general-purpose GUI toolkits, such as wx1 ython, Tkinter,QT, etc., in order to provide object- oriented APIs for embedding plots into applications. John D. Hunter was theperson who originally wrote Matplotlib, and its lead developer was Michael Droettboom. One of the free andopen-source Python library which is basically used for technical and scientific computing is Python SciPy.Matplotlib is widely used in SciPy as most scientific calculations require plotting of graphs and diagrams. Matplotlib is a plotting library like GNU plot. The main advantage towards GNU plot is the facl that Matplotlib is a Python module. Due to the growing interest in python the popularity of Matplotlib is continually rising as well. Another reason for the attractiveness of Matplotlib lies in the fact that it is widely considered to be a perfect alternative to MATLAB, if it is used in combination with Numpy and Scipy. Whereas MATLAB is expensive and closed source, Matplotlib is free and open source code. It is also object-oriented and can be used in an object oriented way. Furthermore it can be used with general-purpose GUItoolkits like wxPython, Qt, and GTK+. There is also a procedural "pylab", which designed to closely resemble thatof MATLAB. This can make it extremely easy for MATLABusers to migrate to Matplotlib. Matplotlib can be used to create publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Another characteristic of Matplotlib is its steep learning curve, which means that users usually make rapid progress after having started. The official website has to say thefollowing about this: "Matplotlib tries to make easy things easy and hard things possible. You can generate plots,histograms, power spectra, bar charts, error charts,scatterplots, etc, with just a few lines of code."

Deep Learning

Deep learning is a computer software that mimics the network of neurons in a brain. It is a subset of machine learning and is called deep learning because it makes use of deep neural networks. Deep learning algorithms are constructed with connected layers:

- The first layer is called the Input Layer
- The last layer is called the Output Layer
- All layers in between are called Hidden Layers. The word deep means the network join neurons in more than two layers.

A deep neural network provides state-of-the-art accuracy in many tasks, from object detection to speech recognition. They can earn automatically, without predefined knowledge explicitly coded by the programmers. To grasp the idea of deep learning, imagine a family, with an infant and parents. The toddler points objects with his little finger and always says the word 'cat.' The toddler points objects with his little fJinger and always says the word ’cat.' As its parents are concerned about his education, they keep telling him 'Yes, that is a cat' or 'No, that is not a cat.' The infant persists in pointing objects but becomes more accurate with 'cats.’ The little kid, deep down, does not know why he can say it is a cat or not. He has just learned how to hierarchies complex features coming up with a cat by looking at the pet overall and continue to focus on details such as the tails or the nose before to make up his mind. A neural network works quite the same. Each layer represents a deeper level of knowledge, i.e., the hierarchy ofknowledge. A neural network with four layers will learn more complex feature than with that with two layers.

The learning occurs in two phases.

1. The fifirst phase consists of applying a nonlinear transformation of the input and create a statistical model as output.
2. The second phase aims at improving the model with a mathematical method known as derivative.

The neural network repeats these two phases hundreds to thousands of time until it has reached a tolerable level of accuracy. The repeat of this two-phase is called an iteration Classifification of Neural Networks

1. Shallow neural network: The Shallow neural network has only one hidden layer between the input and output.
2. Deep neural network: Deep neural networks have more than one layer. For instance, Google LeNet model for image recognition counts 22 layers. The computational models in Deep Learning are loosely inspired by the human brain. The multiple layers of training are called

Artificial Neural Networks (ANN)

Neuron

Artifificial Neural Networks contain layers of neurons. A neuron is a computational unit that calculates a piece of information based on weighted input parameters. Inputs accepted by the neuron are separately weighted. Inputs are summed and passed through a non-linear function to produce output. Each later of neurons detects some additional information, such as edges of things in a picture or tumors in a human body. Multiple layers of neurons can be used to detect additional information about input parameters.

Nodes

Artificial Neural Network is an interconnected group of nodes akin to the vast network of layers of neurons in a brain

Each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another.

Inputs

Inputs are passed into the first layer. Individual neurons receive the inputs, with each of them receiving a specific value. After this, an output is produced based on these values.

Outputs

The outputs from the first layer are then passed into the second layer to be processed. This continues until the final output is produced. The assumption is that the correct output is predeii»«i. Each time data is passed through the network, the end result is compared with the correct one, and tweaks are made to their values until the network creates the correct final output each time.

Some of the commonly used neural networks are as follows:

1. Artificial Neural Network (ANN)
2. Convolutional Neural Network (CNN)
3. Recurrent Neural Network (RNN)
4. Deep Neural Network (DNN)
5. Deep Belief Network (DBN)

Artificial neural networks are one of the main tools used in machine learning. As the aceneurala part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output later can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize. Generally, the working of a human brain by making the right connections is the idea behind ANNs. That was limited to use of silicon and wires as living neurons and dendrites. Here, neurons, part of human brain. That was composed of 86 billion nerve cells. Also, connected to other thousands of cells by Axons. Although, there are various inputs from sensory organs. That was accepted by dendrites. As a result, it creates electric impulses. That is used to travel through the Artificial neural network. Thus, to handle the different issues, neuron send a message to another neuron. As a result, we can say that ANNs are composed of multiple nodes. That imitate biological neurons of the human brain. Although, we connect these neurons by links. Also, they interact with each other. Although, nodes are used to take input data. Further, perform simple operations on the data. As a result, these operations are passed to other neurons. Also, output at each node is called its activation or node value.

Related Survey

ECiFR Assessment in Lung Cancer CT Images: Analysis of Local and Holistic Regions of Interest Using Deep Unsupervised Transfer Learning Description-

This study proposed an approach based on a pre-trained encoder to work as a feature extractor, followed by an MLP for the all classification of the EGFR mutation status. Different regions of analysis were used in order to study the relevance of information from all the lung structures in this complex classification task.

Classifier Ensemble Based on Computed Tomography Attenuation Patterns for Computer-Aided Detection

This paper proposes a CAD system that uses a classier ensemble based on CT attenuation patterns to detect pulmonary nodules in low-dose 3D CT scans automatically. igned the algorithm to discard nodule candidate detections thal have centroid close to each other in an exam. Also, we designed a classier ensemble for false positive reduction, which fuses the individual classification through average to improve nodules classification sensitivity. Our CAD system achieved a sensitivity of 94.90% and an average of 1.0 FP/Scan on the publicly available LUNA16 dataset.

Lung Cancer Prediction from Text Datasets Using Machine Learning

In this work, SVM is used to predict the development of lung cancer. The fundamental objective of this system is to provide consumers with an early warning, allowing them to save both money and time. The performance evaluation of the proposed method produced positive results, demonstrating that SVM can be used effectively by oncologists to aid in the identification of lung cancer. If the prediction is right, it is possible that the doctor will be able to prepare a better prescription and present the patient with an earlier diagnosis.

Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges

This paper has presented a systematic review of current techniques in diagnosis and cure of several cancers affecting human body badly. The focus of this article is to review, analyze, categorize methodologies of different types of cancer and uncover existing limitations. The review has presented six types of cancers lung cancer, breast cancer, brain tumor, liver cancer, leukemia, and skin cancer. Additionally, this study has presented four significant stages of automated cancer diagnosis such as image pre-processing, tumor segmentation, feature extraction, and classification using benchmark

Existing System

Lung cancer is known to be one of the most dangerousdiseases which are the main reason for disease and death when diagnosed in primitive stages. Since lung cancer can only be detected more broadly after it spread to lung parts and the occurrence of lung cancer in the earlier stage is very difficult to predict. It causes a greater risk as radiologists and specialist doctors assess the existence of lung cancer. For this reason, it is important to build a smart and automatic cancer prediction system that is accurate and at which stage of cancer or to improve the accuracy of the previous cancer prediction that will help determines the type of treatment and treatment depthdepending on the severity of the disease.

In this paper, the Adaptive Hierarchical Heuristic Mathematical Model (AHHMM) has been existing for the deep learning approach. **To analyze** deep learning based on the historical therapy scheme in the development of Non- Small Cell Lung Cancers (NSCLC) automated radiation adaptation protocols that aim at optimizing local tumor regulation at lower rates of grade 2 RP2 radiation pneumonitis. Furthermore, the system existing consists of several steps including acquiring the image, preprocessing, binarization, thresholding, and segmentation, extraction of features and detection of deep neural network (DNN). Segmentation of the lung CT image is carried out to extract any significant feature of a segmented image, and a specific feature extraction method is implemented. The test evaluation showed that the model existing could detect 96.67 % accuracy of the absence or presence of lung cancer. Lung cancer is a dangerous disease, and early-stage detection is therefore necessary. This paper presents deep learning assisted Adaptive Hierarchical Heuristic Mathematical Model (AHHMM) to predict lung cancer on computed tomography images. This paper uses the Modified K-means algorithm to pre-classify pictures into slices of images in the same image, where the DlsfN will concentrate on the image classification ofimages in similar images. The next thing is the convolution layer with ltered edges to scan for lung cancer thoroughly. Then, estimating the weighted mean function which replaces the pixel utilizing the cumulative distribution and likelihood distribution method improved the images quality. The injured portion has been segmented by a pixel-kke value measurement after the image has been improved. Based on the similarity calculation, spectral- related features has been extracted. The existsing AHHMM system predicts computed tomography scanning images of lung cancer successfully. At the end of the system, you can say that the system is satisfying its desires. The endings of the evaluation showed that around 90°/o of the images has correctly identified. Such results show that DNN is useful in cyst diagnosis for classifying lung cancer. Hybridized Heuristic Mathematical Model will be implemented in future for predicting the lung cancer at earlier stage.

Drawbachs of Existing System

It is difficult to specify how neurons should be modeled. Difficult to be used in large-scale parallel computing.

Cannot be implemented real time

- Solutions have been proved in effective
- High level of communication and computation overheads High complexity of installing and maintaining

Prone to Error

Proposed System

The method proposed in this paper is an end-to-end structure which fully takes the end-to-end advantages of deep learning and directly classify lung CT images without lung parenchymal segmentation and lung nodule segmentation that the current lung nodule classification algorithms require.

The proposed structure, Optimized Stochastic Learning Model, adopts many structures of modem neural network, for instance, Dense-Block, batch normalisation (BN) and dropout. Meanwhile, LDNNET is an adaptive architecture based on convnets combining soft max classifier which are utilized to alleviate the problems of training deep convnets. Wc input the original lung CT images into the network and finally get the classification results of lung CT images. The neural network was built according to this requirement, consisting of 27 hidden layers23 are convolution and deconvolution layers and four are max pooling layers. In the last layer of the neural network, the output tensor shape is 512 512 7 and is normalized using the sigmoid function, representing the neural networks final output. Each hidden layer is batch normalized, except for the convolution (blue arrows) and pooling (red arrows) layers, and leaky ReLU is used as activation function. All convolution and pooling layers represented by yellow and green arrows add the noise described above. Tàe obvious characteristic of this kind of neural network is its skip and non-skip connections. In training, the networks output needs input cross-entropy cost function to calculate the loss value, and is trained using Adam optimiser to minimize the loss value. The information each neuron contains cornes from the larger area of the neuron in the input image. That is to say, the information contained by the neuron in the middle hidden layer is a high-level summary of a large number of neurons in the underlying layer, which contains advanced features. In the pooling process, the neural network will lose part of the information, which will easily lead to the loss of information on small nodules and ultimately result in errors within the networks judgjnent results. In the process of recognizing small nodules, only small areas in the original image need to be analyzed; therefore, the neural networks recognition of small nodules must rely on the underlying features

Advantages of Proposed System Excellent empirical performance

Fast and efficient, but also as accurate as the state-of-the-art algorithms

- Boost the Performance Quick and Efficient to use Simple to use and interpret Computational Complexity is significantly reduced
- Capability to model high-dimensional features. and animproved accuracy rate.

Hardware Reiärements

Processor	1.4 GHz 64-bit processor
Disk Space	100GB Free Space
RAM	Minimum 8GB
Graphies Device	Super VGA (1024 x 768) or higher-resolution

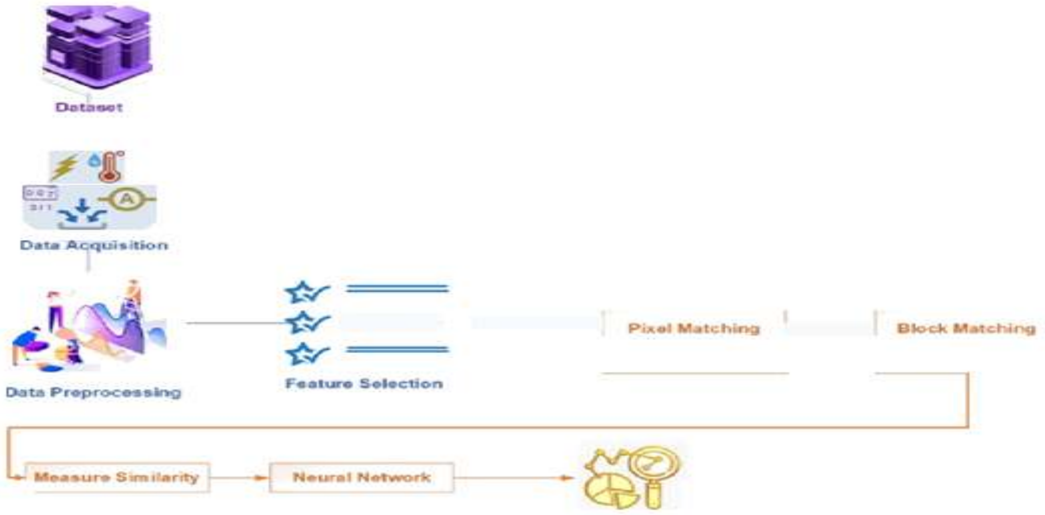
Software Requirements

Python Anaconda

Jupiter Notebook

- TensorPlow Package
- Plotly Package
- Matplotlib Package

Architecture



Algorithm Implemented
Optimized Stochastic Learning Model Algorithm

Advantages of IinplementedAlgorithm

It can simplify and speed up learning and inference of the networks and make the learning problem much easier. Yields output maps for inputs of any size, the output dimensions are typically reduced by subsampling. It can take input of arbitrary size of data to generate correspondingly-sized output with efficient inference and learning.

System Implementation

Module 1 : **Image Preprocessing**

Thresholding techniques attempt to binarize a grayscale image based on pixel density. Thresholding provides a simple way to achieve segmentation operation over the foreground and background regions of an image. A parameter called intensity threshold determines the output produced. Depending on whether the pixel intensity was greater that on lesser than the threshold value, it would be replaced by either a white or black pixel. In Morphology, the image is examined using a small template that is known as a structuring element. A structuring element such as kernel is used as a reference to compare with the corresponding pixels at all potential locations in the image. These operations are suited for used on binary images. A kernel must be specified for performing morphological operations and it influences the nature of the results. Four morphological approaches tested in this paper are erosion, dilation, opening, and closing. Blurring removes high-frequency contents in the image such as noise. Three blurring methods tested in this paper are average fdter, Gaussian filter, median filter.

In average Fdter, the image is convolved using a normalized box fdter that takes the average of entire pixels under a specified kernel area and changes the central element. Gaussian kernel is used in the Gaussian fdter to perform blurring. It is able to remove the Gaussian noise.

Module 2 : **Dataset Splitting**

Splitting your dataset is essential for an unbiased evaluation of prediction performance. In most cases, its enough to split your dataset randomly into three subsets:

1. The training set is applied to train, or fit, your model. For example, you use the training set to find the optimal weights, or coefficients, for various algorithms.
2. The validation set is used for unbiased model evaluation during hyperparameter tuning. For example, when you want to find the optimal number of neurons in a neural network or the best kernel for a support vector machine, you experiment with different values. For each considered setting of hyperparameters, you fit the model with the training set and assess its performance with the validation set.
3. The test set is needed for an unbiased evaluation of the final model. You shou1dn't use it for fitting or validation. Underfitting is usually the consequence of a model being unable to encapsulate the relations among data.

For example, this can happen when trying to represent nonlinear relations with a Enear model. Underfitted models will likely have poor performance with both training and test sets. Overfitting usually takes place when a model has an excessively complex structure and learns both the existing relations among data and noise. Such models often have bad generalization capabilities. Although they work well with training data, they usually yield poor performance with unseen (test) data.

Module 3 : **Model Training**

Convolutioa layer

The Convo layer is occasionally known as the feature extraction layer since the data features are extracted within this layer. First, a part of the data is associated with the Convo layer to make a convolution operation and calculate the dot product between the approachable field and rate. The outcome of the process is a single number of output capacities. The Convo layer also holds the ReLU activation function to build all negative valuesto zero

Pooling layer

The pooling layer is used to decrease the spatial capacity of the input data after convolution. The layer can use two layers of convolution. If we put a fully connected layer after the Convo layer without first including a pooling or max pooling layer, then it will be computationally expensive, which we do not want. Therefore, max pooling must be used to reduce the spatial volume of the input data.

II. CONCLUSION

Lung cancer is a dangerous disease and early-stage detection is therefore necessary. This paper presents deep learning assisted Optimized Stochastic Learning Model to predict lung cancer on computed tomography images. This paper uses the Modified K- means algorithm to pre-classify pictures into slices of images in the same image where the neural network will concentrate on the image classification of images in similar images. The next thing is the convolution layer with filtered edges to scan for lung cancer thoroughly. Then estimating the weighted mean function which replaces the pixel utilizing the cumulative distribution and likeEhood distribution method improved the images quality

Future Work

The proposed syslem result is useful in cyst diagnosis for classifying lung cancer. Optimized Stochastic Learning Model will be implemented in future for predicting the lung cancer at earlier stage.

REFERENCES

[1]. Francisco Silva , Tania Pereira , Joana Morgado , Julieta Frade , Jos Mendes , Cludia Preitas , Eduardo Negro , Beatrix Flor De Lima , Miguel Correia Da Silva , Antnio J. Madureira , Isabel Ramos , Venceslau Hespanhol , 5os Lus Costa , Antnio Cunha and Hlder

[2]. P. Oliveira Published Year 2021 C. Ani1 Kumar,1 S. Harish,1 Prabha Ravi,2 Murthy SVN,3 B. P. Pradeep Kumar,4 V. Mohanave1,5,6 Nouf

[3]. Alyami,7 S. Shanmuga Priya,8 and Amare Kebede Asfaw S ,p, S. G. Armato et at., The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database Jan. , , , L. Alzubaidi, M. A. Fadhel, O. Al- Shamma, J. Zhang, J. Santamara, Y. Duan, and S. R. Oleiwi, Towards a better understanding of transfer p. ,Jun. , , J. D. Horbar, E. M. Edwards, L. T. Greenberg, K. A. Morrow, R. F. Soil, M. E. Buus-Frank, and J. S. Buzas,Variation in performance of neonatal Mar. , Art. no. e, ., K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, , . MA,USA: MIT Press, , ., American Cancer Lung Cancer. Accessed: (,). cancer/about/key- statistics.html Society. , D. Ost, A. M. Pein,and S. H. Feinsilver, The solitary pulmonary nodule, /NEJMcp, , Z. Wu, R. Ge, G. Shi, L. Zhang, Y. Chen, L. Luo, Y. Cao, and H. Yu, convolutional neura network for false-positive reduction in pulmonary nodule detection, doi: , ., /, -, /aba, c. , Q. Don, H. Chen, L. Yu, S. Qin, and P.-A. Heng, Multilevel contextual , -D CNNs for false positivereduction in pulmonary nodule detection, , ., /TBME., ., , F. Milletari, N. Navab, and S.-A. Ahmadi, V-Net: Fully convolu- , ., /, DV., ., . , H. Zhu, G. Han, Y. Peng, W.Zhang, C. Lin, and H. Zhao, Functional- realistic CT image super-resolution for earlystage pulmonary nodule doi:, ., /j.future., ., ., ., S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real- time object detection with region proposal networks, in Proc. , th Int. Conf. , J. Mei, M. M. Cheng, G. Xu, L.

[4]. R. Wan, and H. Zhang, SANet: A slice-aware network for pulmonary nodule detection, Trans. Pattern Anal. Mach. Intell., early access, Mar. , , doi: , .,

[5]. /TPAMI., ., .F. Pereira, D. Menotti, and L. F. de Oliveira, A , D lung nodule candidate detection by grouping DCNN , D candidates, in Proc. , th Int. Joint Conf. , ., /, . , M. N. Cribbs and D. J. C. Mackay, Variational Gaussian process clas- Nov. , . , L. Cong, L. Zhou, H. Liu, and J. Wang, Outcomes of high-ow nasal cannula versus non-invasive positive pressure ventilation for patients with acute exacerbations of chronic obstructive pulmonary disease, Int. J. Clin. , L. Zhou, L. Guan, W. Wu, X. Li, X. Chen, B. Guo, Y. Huo, S. Xu, Y. Yang, and R. Chen, High-pressure versus low pressure home non- invasive positive pressure ventilation with built-in software in patients with Art. no. , . ,R. Liu, H. Wang, and X. Yu, Shared-nearest-neighbor-based clustering Jun. , . P. Shen, J. Chao, and J. Zhao, forecasting exchange rate using deep belief , G. Wang, J. Qiao, J. Bi, W. Li, andM. Zhou, TL-GDBN: Growing deep , G. Lu, D. Li, and L. Zhang, Clinical investigation of depression in elderly patients with chronic obstructive pulmonary disease, China Med., J. M. Marin, S. J. Carrizo, C. Casanova, P. Martinez-Camblor, J. B.

Soriano, A. G. N. Agusti, and B. R. Celli, Prediction of risk of COPD exacerbations Mar. , , X. Guo, M. Zhou, S. Liu, and L. Qi, Lexicographic multi objective scatter search for the optimization of sequence dependent selective disassembly. , R. Veevers and S. Hayward, Morphing and docking visualization of biomolecular structures using multidimensional scaling, J. Mol. Graph. , S. Cihose, J. Mitra, S. Khanna, and J. Dowling, An improved patient-specific mortality risk prediction in ICU in a random forest c1assi- Aug. , , , B. Sluban and N. Lavra, Relating ensemble diversity and performance: Sul. , , , R. Matteo and V. Giorgio, Ensemble Methods: A Review. London, U.K.: Chapman &amp; Hall, , . , P. D. Gopalan and S. Pershad, Decision-making in ICUA systematic review of factors considered important by ICU clinician decision makers Apr. , . , W.-H. Hsieh, D.-H. Shih, P.-Y. Shih, and S.-B. Lin, An ensemble classier with a case-based reasoning system for identifying Internet addiction, Int., S. L. Javan, M. M. Sepehri, M. Layeghian Javan, and T. Khatibi, An intelligent warning model for early prediction of cardiac arrest in Sep. , ., R. Dybowski, V. Gant, P. Weller, and R. Chang, Prediction of outcome in critically ill patients using artificial neural network synthesis by genetic , K. Lin, Y. Hu, and G. Kong, Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model, Int. J. , G. Carioli, P. Bertuccio, P. Boffetta, F. Levi, C. La Vecchia, E. Negri, and M. Malvezzi, European cancer mortality predictions for the year , May , . , O. H. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, A genetic algorithm-based feature selection, Int.

[6]. J. Electron. Commun. Comput. , A. E. W. 5o hnson, N. Dunkley, L. Mayaud, A. Tsanas, A. A. Kramer, and G. D. Clifford, Patient specific predictions in the intensive care unit using , A. Awad, M. Bader-El-Den, J. McNicholas,and 5. Briggs, Early hospital mortality prediction of intensive care unit patients using an ensemble Dec. , , M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, Multiple imputation by chained equations: What is it and how does it work, Int. J. Methods mpr., . , N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic minority over- sampling technique SMOTE, : Synthetic minority Jan. , , G. Molenberghs and U. Hasselt, Models for discrete longitudinal data, in Models for Discrete Longitudinal Data. New York, NY, USA: Springer-Verlag, Oct. , ., B. 5. Weled, Book review: The little ICU book of facts and formulas. Paul 1 Marino MD PhD, with contributions from Kenneth M Sutin MD. Philadelphia: Wolters Kluwer/Lippincott Williams &amp; Wilkins. , , F. Seccareccia, F. Pannozzo, P. Dima, A. Minoprio, A. Menditto, C. Lo Noce, and S. Giampaoli, Heart rate as a predictor of mortality: The Aug. , . , N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, and X. Zhao, A novel coronavirus from patients with pneumonia in China, , New England J. Med., to be published.

[7]. S. Al Youha, S. A. Dot, M. H. Jamal, S. Almazeedi,M. Al Haddad, M. AlSeaidan, A. Y. Al-Muhaini, F. Al-Ghimlas, and S. K. Al-Sabah, Val-idation ofthe Kuwait progression indicator score for predicting progression of severity in COVID, medRxiv, to bepublished. , P. Ramachandran, M. Gajendran, A. Perisetti, K. O. Elklioly, A. Chakraborti, G. Lippi, and

[8]. H. Goyal, Red blood cell distribution width (RDW) in hospitalized COVID-, patients, medRxiv, to be published. , B. H. Poy, J. C. T. Carlson, E. Reinertsen, R.P. Valls, E. Palanques-Tost, C. Mow, M. B. Westover, A. D. Aguirre, and J. M. Higgins, Elevated RDW is associated with increased mortality risk in COVID-, medRxiv, to be published. , L. Yan, H. T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, , W. Liang, J. Yao, and 5. He, Early triage of critically ill COVID-, , . , Y.-P. Liu, G.-M. Li, J. He, Y. Liu, M. Li, R. Zhang, Y.-L. Li, Y.-Z. Wu, and B. Diao, Combined use of the neutrophil-tolymphocyte ratio and CRP to predict , -day disease severity in , hospitalized

patients with COVID-, no. , p. , May , .

[9]. , Y. Shang, T. Liu, Y. Wei, J. Li, L. Shao, M. Liu, Y.

[10]. Zhang, Z. Zhao, H. Xu, Z. Peng, X. Wang, and F. Zhou, Scoring systems for predicting mor- Jul. , Art. no. , . , L. Zhang, X. Yan, Q. Fan, H. Liu, X. Liu, Z. Liu, and Z. Zhang, D- dimer levels on admission to predict in- hospital mortality in patients with , . , Q. Gu, Z. Li, and J. Han, Linear discriminant dimensionality reduction, in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases, , , .

[11]. , X. Shi, Q. Li, Y. Qi, T. Huang, and J. Li, An accident prediction approach based on XCiBoost, in Proc. , th Int. Conf. Intell. Syst. Knowl. Eng. , S. K. Pal and S. Mitra, Multilayer perceptron, fuzzy sets, and classi- , A.Zlotnik and V. Abraira, A general-purpose nomogram generator for predictive logistic regression models, Stata 5.: Promoting Commun. , J. Wang, X.Wu, Y. **Tian,** X. Li, X. Zhao, and M. Zhang, Dynamic changes and diagnostic and prognostic signicance of serum PCT, Hs-CRP and S-, protein in central nervous system infection, Exp. Therapeutic , B. Yildiz, H. Poyraz, N. Cetin, N. Kural, and O. Colak, High sensitive C-reactive protein: A new marker for urinary tract infection, VUR and , . , E. W. Steyerberg et al., Internal validation of predictive models: Efciency of some procedures for logistic regression analysis, 5. Clin. , A. Bemheim, X. Mei, M. **Huang,** Y. Yang, Z.

[12]. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li,S. Li, H. Shan, A. Jacobi, and M. Chung, Chest CT ridings in corona irus disease- ,(COVlD-,):Relationship to duration , , /radiol., . , T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, Automated detection of COVID-, cases using Jun. , Art. no. , doi: ,/j.compbiomed., ., . ,s-Fernando Roberto Pereira , loo Mario Clementin De Andrade , Dante Luiz Escuissato and Lucas Ferrari De Oliveria