Journal of Industrial Information Integration 30 (2022) 100386

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.sciencedirect.com/journal/journal-of-industrial-information-integration

A cloud-based deep learning model in heterogeneous data integration system for lung cancer detection in medical industry 4.0

Chang Gu [a,1], Chenyang Dai [a,1], Xin Shi [b], Zhiqiang Wu [c], Chang Chen [a,*]

[a] Department of Thoracic Surgery, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai, China
[b] Department of Cardiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China
[c] Department of Electrical Engineering, Wright State University, Dayton, OH, USA

ARTICLE INFO

Keywords:
Medical industry 4.0
Lung cancer
Heterogeneous data integration
Cloud-based deep learning
Content-based image retrieval

ABSTRACT

Currently, lung cancer has become one of the most common and deadliest types of cancer. Due to its severity, many countries are now encouraging their at-risk citizens to test and treat lung cancer early. Lung cancer has been worse for poor regions or countries, whose citizens are more susceptible to lung cancer, as the local medical resources and healthcare provider level are inadequate. In recent years, this situation can be significantly improved by leveraging the existing datasets about lung cancer in developed countries. However, due to the poor synchronization of data collection methods, the collected data is heterogeneous, and can't be readily used. Artificial intelligence (AI), big data, cloud computing, and the internet of things accelerate the 4th revolution in the medical industry, and we called it medical industry 4.0. In the medical industry 4.0, lung cancer can be early detected by using a very intelligent approach. In this paper, using AI and cloud platform techniques in the medical industry 4.0, we propose an intelligent detection system including data integration, detection, historical cases comparison, similar cases inquiry, and retrieval for lung cancer. In this system, doctors can integrate the heterogeneous data at hand, and source a large amount of integrated data as a reference when treating the patient. A cloud-based deep learning model is integrated into this system, and then a content-based image retrieval system for similarity comparison is used. Finally, some public datasets are used to train and test this system, and results prove its performance is better than that of some baseline approaches. Then the similar case finding is evaluated with cosine similarity and all similarities reach over 0.93. The heterogeneous data integration system creates a good effect in helping doctors and patients access better diagnosis and treatment for lung cancer.

1. Introduction

With the increase of the aging population, the number of cancer patients continues to increase in the world. As one of the most common cancers, lung cancer has attracted widespread attention from doctors and scholars. At present, the number of deaths from lung cancer ranks first among all cancers in the world. In the United States, the incidence of lung cancer ranks the second in all cancers. The reason that the mortality rate of lung cancer is very high is there are almost no symptoms in the early stage of lung cancer. When the patient shows clinical symptoms, lung cancer has often developed to the middle and late stages and even spread to other parts of the body, and the probability of cure becomes very low. For the patients who survived, the prognosis is also abysmal. Although scholars are vigorously developing drugs for lung cancer treatment, they have little effect on advanced lung cancer. Therefore, detecting lung cancer in the early stage without any symptoms has become the first choice for lung cancer treatment.

Artificial intelligence (AI), big data, cloud computing, and the internet of things (IoT) accelerate the 4th revolution in the medical industry, and we called it medical industry 4.0. In the medical industry 4.0, lung cancer can be early detected by using a very intelligent approach. In the medical industry 4.0, as the medical system integrates with the cloud platform and AI, doctors can better understand their patients. Some cloud platforms can constantly give doctors recommendations on their patients' health condition based on the record of past smoking history, historical physical exam results, etc., as well as

* Corresponding author.
E-mail address: changchenc@tongji.edu.cn (C. Chen).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.jii.2022.100386

Available online 9 August 2022

continuous patient reports with any discomforts or symptoms. If the cloud system suspects a patient might have or be at high risk of lung cancer, the doctor can be alerted to pay extra attention. However, this cloud platform is still limited to more developed countries with a complete cloud medical record system.

While developed countries have taken steps to help their citizens prevent and detect lung cancer in the early stages by integrating cloud platforms and AI tools, poorer countries lack in all aspects. Research showed that the risk of lung cancer increased among individuals living in poor areas. This is mainly due to more intense smoking habits, chronic obstructive pulmonary disease diagnosis, and personal history of cancer [1]. Therefore, the detection technique must first be changed to help improve the overall survival rate in poor areas and countries as doctors are trained through a prolonged academic process and cannot be

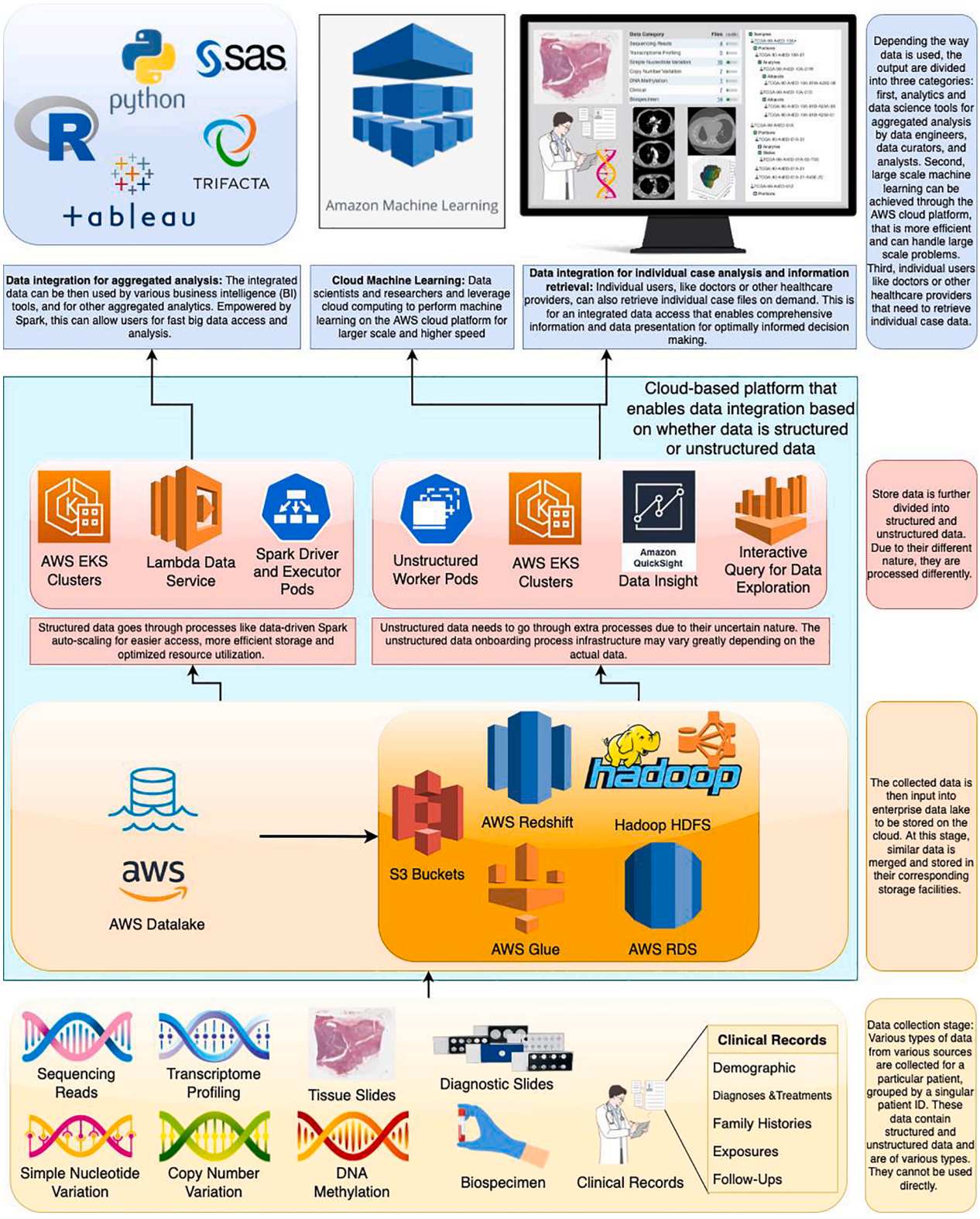

Fig. 1. A framework for medical data integration from the step of data collectin, storage, processing, access and analytics, powered by Amazon web services (AWS).

immediately improved. Then we must turn to the solution of medical detection based on the cloud platform and AI.

Medical detection is a comprehensive understanding of the disease to lay a good foundation for the subsequent treatment or control. The first step in the detection is to provide clues based on the past medical history. Especially cancer patients require some more in-depth tests such as tissue slices, DNA/RNA sequence, etc. All the data are heterogeneous by nature: image, clinical records, sequencing reads, etc. They are not ready to be used easily by intelligence tools or data analytics models, requiring integrated heterogeneous data. Therefore, an indispensable step for the successful medical data analysis is integrating data.

One characteristic of the data collected is its heterogeneity limited by some devices. Data integration is to consolidate heterogeneous data from various sources into a singular and uniform dataset. This allows data connection between source and target systems while simultaneously processing data to reach a previously agreed level of service. A demonstration of data integration is shown in Fig. 1. The process first includes data collection/extraction and transformation for medical data integration: datasets from different sources are collected, standardized, and transformed before being stored in a database. The data collected can be divided into several categories based on data type, such as image, DNA, clinical records, etc. Then depending on the collected data, it's put into various kinds of cloud data storage options, which can be accessed via cloud computing. Once the data is in storage, they need to be treated differently depending on their structural characteristics. For structured data, its processing is more straightforward. It only needs to be systematically cleaned and processed, and then it's ready to be used. For unstructured data, the options depend on the actual use case. After completion of data processing, we create accessing portals for further end-user analysis, which leverages AWS cloud computing capacities [2]. Once all the framework is set up, the end-user can use integrated data by using aggregated tools like Python, R, or business intelligence tools like tableau, or direct access tools that allow doctors to extract comprehensive, integrated medical data for a patient.

Lung cancer is likely to present as lung nodules in the early stage, and chest CT is the primary lung cancer detection method. However, the CT radiation dose is large, it is not suitable for early lung cancer detection. Lungs contain a large amount of air with an extremely low absorption rate for X-rays, so low-dose CT can detect small nodules effectively and has the same detection ability as conventional-dose CT. The traditional detection of lung cancer is to judge whether the patient has lung cancer via observing the results of CT by the doctor. In recent years, big data and computer-aided detection have achieved good results. Medical big data analysis can use pattern recognition and data mining based on massive medical images, pathology, clinical, genetic, and molecular data to obtain in-depth features contained in the data for biomedical research and clinical applications.

Once a lung nodule is cancerous, we want to get as much reference on similar cases to it as possible. Machine learning has enabled similarity comparison between different objects. It's crucial for problems like searching for images/videos similar to a given image/video [3]. The similarity lies in two folds: the first is direct visual similarity, and the second is semantic relation.

As the internet continuously spreads in countries in poverty, the local healthcare providers can connect to the cloud and access records from countries with better healthcare systems. This enables the local doctors to access some of the most extensive medical databases, which can be highly informative. They provide a detailed record for the diagnosis of the cases, treatment plan, medicine prescription, recovery process, and long-term prognosis. Local doctors can search for cases on the cloud platform, refer to these records when diagnosing their patients and make better-informed decisions. However, local doctors are not traditionally trained in the medical system in which the doctors in developed areas were trained. Their ability to refer to historical diagnosis is limited by reasons including but not limited to: language, experience, surgical condition, access to medicine, examination methods, etc. Thus, providing doctors with adequate access to the cloud database to maximize utilization becomes a challenging topic.

Therefore, taking the rapid development of biomedical big data and computer-assisted lung cancer detection as an opportunity, we propose a cloud-based deep learning for lung cancer detection and a past similar case finding system. This cloud system uses CT scans and deep learning to detect whether patients have lung cancer. The system will search for similar cases in the vast lung cancer cloud database if the patient has lung cancer. Doctors using this cloud platform can refer to past cases and propose more fact-based diagnoses and treatment plans for patients through past cases, therefore inferring from a large number of historical cases.

This paper makes the following contributions: 1) we propose a cloud-based deep learning model in heterogeneous data integration system for lung cancer detection that takes CT image series as input. 2) For cases deemed to be positive, we use a similarity matching scheme to find the most similar cases in the TCIA dataset, return case IDs and then find them in the TCGA dataset for similar cases. 3) We test our detection and similarity case finding system on the LIDC-IDRI dataset [4] and TCIA/TCGA dataset [5], and it shows a strong performance. 4) We implement the data integration in our proposed system through dataset migration and similar case finding.

The organization of this paper is as follows: In section I, we introduce the background and motivation, and give an overview of the system we are proposing. In section II, we discuss related work in the field. Then, in section III, we perform the architecture of our detection model and similarity comparison algorithm. Section IV discusses the LIDC-IDRI dataset and TCIA/TCGA dataset we are using and the evaluation metrics to test the model. Then in section V, we show experimental results and discuss the performance improvement by parallelizing the training process. In section VI, we conclude the findings and future directions.

2. Related work

2.1. Computer-aided detection

Using computers to aid medical detection on CT images can date back to the early 2000s when Golosio et al. [6] selected the region of interest based on multi-threshold surface triangulation. The model was tested on 84 CT cases, which contained 77 marked nodules, and reached a sensitivity of 79%. Wang et al. [7] used an R-CNN and deep active learning-based method to create a 3D representation of lung nodules based on CT scans and then create spatial segmentations. Xie et al. [8] proposed a knowledge-based collaborative deep learning for lung nodule classification.

2.2. Medical data integration

With the increasing number of medical sensors in recent years, data collection has been more accessible than ever. With a large amount of data that's now available, there now are enormous opportunities and challenges. The deluge of data doesn't only mean the huge volume but also the traffic, uncertainty, diversity, etc. The data facilities designed for traditional use can no longer appropriately handle this data. This data issue has come to light in the healthcare field as well. The integration of diverse medical data has now been a pressing challenge for the medical industry. Coming from varied sources, the medical data is heterogeneous by nature. Thus, integrating medical data can enable data enrichment, enhancement, and more effective use. Some scholars have focused specifically on medical data integration, as the survey by Dhayne et al.[9] has shown. Aceto et al. [10] leveraged industry 4.0 and created a framework that can dynamically integrate healthcare data with the IoT and cloud computing. Tsiknakis et al. [11] proposed a semantic grid infrastructure that enables integrated access and analysis on multilevel medical data.

2.3. AI for the detection in lung cancer

In recent years, there are many different fields where advanced AI-based optimization methods have been applied as solution approaches, such as online learning, scheduling, multi-objective optimization, healthcare, transportation routing, data processing, etc., not just on the topic of lung cancer detection. Some highlights of these AI-based methods can show impressive effectiveness in the aforementioned fields [12–17].

Due to the advantages of CT images in early lung cancer detection, a lot of scholars have implemented some work based on CT images. Computer-aided detection system for lung cancer, which uses computers to assist doctors in detecting lung cancer and analyzing the probability that they are benign or malignant, has been widely recognized.

At present, some scholars have begun to distinguish benign and malignant lung cancer based on CT images. The methods they use are mainly improved based on lung nodule detection methods. Google published the latest achievements in Nature Medicine, using the computer-assisted detection for early lung cancer, outperforming six radiologists with extensive experience [18]. The AI-assisted detection model reduced the number of false-positive rate by 11%. Huang et al. [19] used deep learning in CT readings and other universally available clinical information to help estimate the 3-year lung cancer risk after two CT scans. Guo et al. [20] used a combination of 3D deep learning and radiomics methods to automatically distinguish lung adenocarcinomas, squamous cell carcinomas, and small cell lung cancers. The accurate annotation of lung cancer in CT scans is important, so Lustberg et al. [21] investigated whether using software-generated contouring will save time if used as a starting point for manual organ-at-risk (OAR) contouring for lung cancer patients. Tortora et al. [22] focused on non-small cell lung cancer, and used deep reinforcement learning to auto-train a classifier. Xiao et al. [23] used generative adversarial networks (GANs) to generate data and classify lung cancer cases. Also with data limitations, Ubaldi et al. [24] developed strategies to develop radiomics and machine learning models for lung cancer stage and histology detection.

3. Model architecture

As early stage lung cancer usually expresses as malignant lung nodules, successfully detecting malignant lung nodules and treating them give the patient a much higher chance of recovery. However, nodules are common in lungs, and they are not always malignant, and their detections are prone to false positives and false negatives, both of which can have severe consequences to the patients. Thus, successful malignant nodule detection has become a hot topic in lung cancer detection. In this section, we first discuss the overall framework of the cloud-based deep learning for lung cancer detection, next the lung cancer detection model based on deep learning, and then we will discuss similar case comparison for historical case finding.

3.1. Overall framework

In this paper, we propose a cloud system uses deep learning to detect whether patients have lung cancer from their CT scans. If the system detects lung cancer, it will search for similar cases in the vast lung cancer cloud database to retrieve past medical records to help doctors make better-informed treatment plans. Fig. 2 shows the architecture of our intelligent lung cancer detection and similar case finding framework proposed in this paper. The image data (CT scans) are first collected by doctors and uploaded and stored in the cloud database. Then, the data undergoes processing to enable the model to most efficiently utilizes the data. Then, based on the structured/unstructured nature of the data, the data is separated and treated differently to extract their features. Once the features are extracted from the data, the lung cancer detection model is used to determine whether the patient has potential cancer. If the

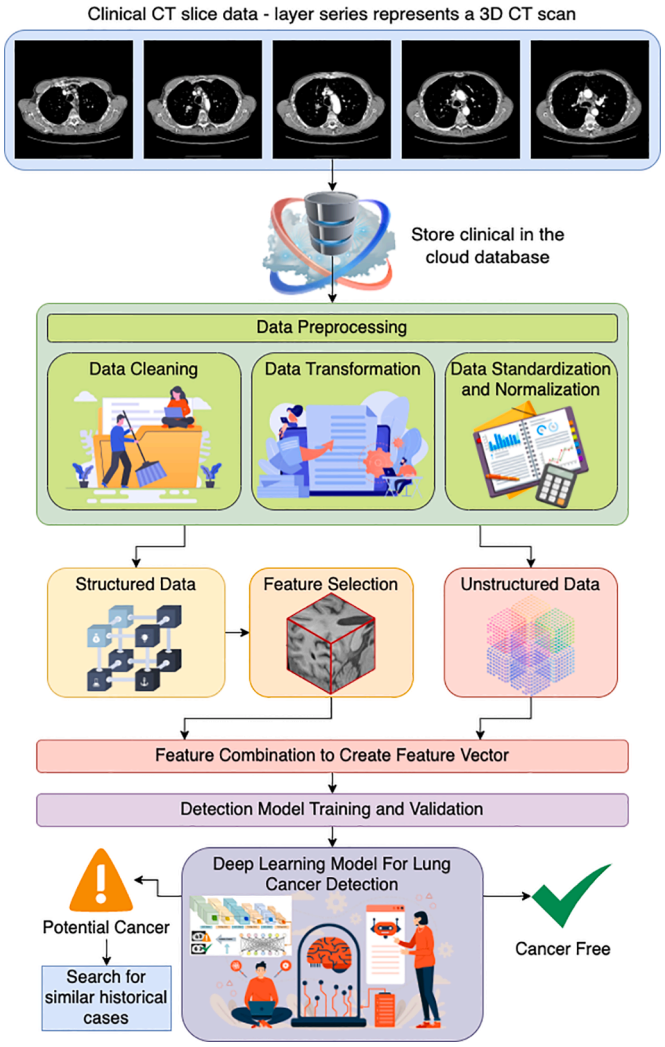

Fig. 2. The architecture of an intelligent lung cancer detection system.

model predicted that the patient might have lung cancer, then the case is ingested by the similar case finding system to search for similar cases. The doctor can define the amount of similar cases he or she wants, and the system will return the results via cloud. In the system we proposed, we are looking at a variety of data types.

While the system can run on individual PC, to further improve the performance of lung cancer detection and similar case finding system proposed in this paper, and to increase accessibility, we integrated the system into the cloud platform. There are several reasons for the integration. First, the dataset volume is very large, and constantly increasing. Training, inferencing and fine-tuning our model on this dataset is challenging for a single workstation. By leveraging cloud computing, this process can be parallelized, and thus significantly reduce the time. The cloud-based system can also access other cloud-based tools easily, such as cloud storage, Spark, AWS, etc. These well-established tools can increase productivity drastically. Secondly, we propose this system for users in less developed areas to access. Their computational resources are limited, and the more portion of the system we move to the cloud, the better they are able to use it. Therefore, we integrated the most computational-intense components of our system to the cloud, and the users can simply access the system via APIs, and only need to upload the necessary CT scan data. The lung cancer detection model makes a decision. If the identification is positive, it searches for similar cases and provides them to the doctor to assist in the detection for the lung cancer. Because the pathology information, inspection

reports, treatment prescriptions and other information of historical cases are stored in cloud storage, doctors can quickly query related information for further data analysis and effective treatment plans.

3.2. Lung cancer detection based on deep learning model

In clinical detection, the pre-operative prediction of the patient based on medical images by the imaging doctor relies on the serialized inter-layer information. In the previous work, the imaging doctor often used 2D deep learning for lung cancer detection, and selected a few slices with important information as the model input. The original 2D deep learning was initially designed for the recognition of two-dimensional shapes. It can directly process two-dimensional images, establish a mapping relationship from low-level signals to high-level semantics, classify and recognize visual images. However, in our case, even if all slices of CT data are sent to the network, the 2D deep learning model structure still ignores the inter-layer information, not considering the correlation of the upper and lower slices. Therefore, in the previous deep learning model design, the serialization characteristics of patient images cannot be fully utilized. 2D deep learning model has certain limitations when analyzing three-dimensional data. The 3D deep learning model is a typical structure to deal with this deficiency. The difference between 2D and 3D deep learning lies in the input of a multi-channel image. 2D deep learning outputs 2D feature maps, while 3D deep learning makes full use of multi-channel information, and its output is 3D feature maps. In 3D deep learning, the dimension of the scanned slice is regarded as the third dimension. A cube is formed by stacking multiple consecutive slice layers. Then, the 3D convolution kernel is used to extract the features of the 3D volume data. In this structure, each feature map in the convolutional layer is connected to multiple adjacent consecutive layers in the previous layer so that the inter-layer information can be fully captured.

In our case, our data can be considered 3D. Simply applying 2D deep learning to the data layer by layer, which is essentially dimension reduction, will lose the spatial information of the data between layers. Thus, we propose a 3D deep learning model to process lung cancer image data. This method can simultaneously extract 3D features from different frames through a 3D convolution operation to better analyze 3D objects. The feature extracted by the 3D deep learning model is:

$$f^{i}_{(x,\ y,\ z)} = \sigma \times \left(b_i + \sum_{p=0}^{p_i} \sum_{q=0}^{q_i} \sum_{r=0}^{r_i} w_{pqr} f^{i-1}_{(x+p,\ y+q,\ z+r)} \right) \tag{1}$$

where $f^{i}_{(x,\ y,\ z)}$ is the feature value of ith layer, at (x, y, z). b_i is the bias value at ith layer, and $(p_i,\ q_i, r_i)$ is the kernel size of the ith layer. σ is the sigmoid function.

The 3D deep learning model used in this paper is designed based on the model in [25], it expands the original 2D structure to adapt to 3D data, and achieve the task of lung cancer classification. The basic framework of 3D deep learning model is shown in Fig. 3. A series of slices are used to construct 3D data in the unit of the nodule, which is used as the input of 3D deep learning model. Then we improved the existing model and transferred the model parameters pre-trained on the multi-modal medical image dataset to this task for 3D feature extraction,

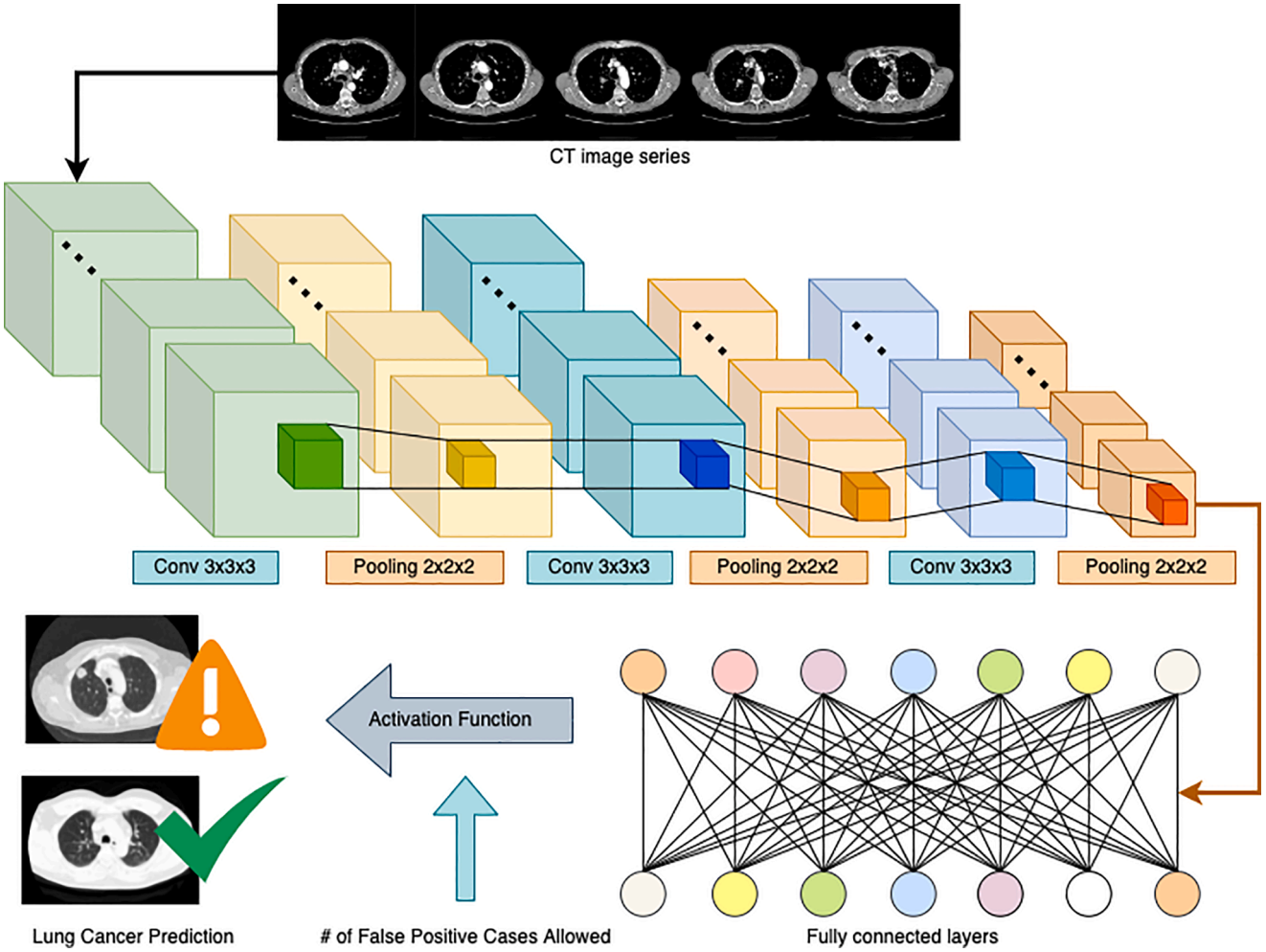

Fig. 3. The architecture of the deep learning model with the number of false-positive cases allowed.

making full use of the sequence level features of the medical image. Compared with traditional methods, the extracted features are more robust and satisfy the task of medical image classification. We add a global average pooling layer to down-sample the extracted features and normalize the output elements to maintain the consistency of feature expression. Then, we use PCA to reduce the dimensionality of the data and input the post-processed training set with the new integrated feature expression into the classifier. The classifier consists of two convolutional layers and a pooling layer. The output of the fully connected layer is two nodes representing two categories.

3.3. Similar case search and comparison

In a typical image search, the user inputs a keyword, and then the search engine returns an image or images associated with the keyword. This procedure is similar to when we are using an image as the keyword. When we give the search system an image as the query, it should return images that are "similar" to the query image. The definition of "similar" depends on our use case. This procedure is commonly referred to as Content-Based Image Retrieval (CBIR) systems, which is shown in Fig. 4. CBIR system usually is divided into the following steps [26].

First, we need to define a feature extractor. By defining the feature extractor, we tell the computer what features of the images we extract. Then these features will be extracted from the images to compare against those of the images in the to-be-searched cloud-based database.

Second, we need to preprocess the data. After we selected the feature extractor in the previous step, we will proceed to apply this to all the images in the database. In this way, we create a unique descriptor of each image in the database that our similar image searcher can compare against the query image and return the result if similar images are found. The step of translating images into features is a hashing process, and the overall procedure is usually called "indexing." When treating a big dataset, we can leverage the computing capability of cloud computing to parallelize the process to make it more efficient.

Then, once the dataset is prepared, we need to define a comparison metric. These are the criteria that we use to determine whether two images are similar and how similar. After extracting the features from the previous steps, we should get a feature vector for each image. Then, we calculate the distance between the target image feature and the candidate image feature. If the distance falls within a set threshold, we can conclude that the two images are similar. Common similar functions often used in the machine learning and computer vision field are Euclidean, Manhattan, Cosine, and Chi-squared distances. Here in this paper, our distance is defined as:

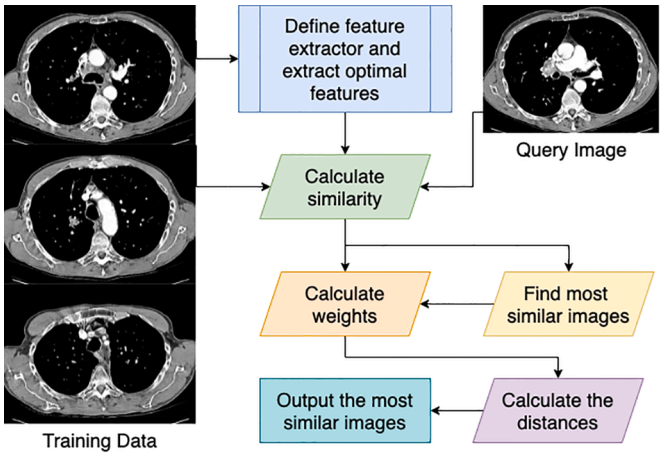

Fig. 4. Content-Based Image Retrieval (CBIR) systems.

$$l(\theta,\ Orig,\ Cand) = \sum_{i=1}^{w_{Orig}} \sum_{j=1}^{h_{Orig}} |f(\theta,\ Orig) - f(\theta,\ Cand)| \tag{2}$$

where *Orig*, *Cand* represent the original and candidate image to be compared. θ is the weights, f is the feature function, and w_{Orig}, h_{Orig} are the width and height of the image.

Finally, we traverse the database and search for images that are similar to our query image. In this process, we can return all the images with similarities higher than the threshold (or, in other words, distance lower than the threshold). Or, we can also sort them based on the similarity before returning the results.

In our use case, we first preprocess all the cases in the lung cancer database. In this step, we process each of the lung cancer cases and extract their image features to be used later. When we identify a nodule as malicious with the detector mentioned in the previous subsection, we extract its image features and create its feature vector. Then we compare against the nodules in the database.

4. Experiments

4.1. Dataset

The dataset in this paper consists of two parts: first, we use the image data from the LIDC-IDRI dataset to train our model. Then, we search similar historical cases in the TCIA/TCGA dataset integrated in the cloud platform. We are selecting the dataset in this way because the LIDC-IDRI dataset is almost ten times larger and can accommodate the training for a more sophisticated lung cancer detection model. Then, when we conduct a similarity test, we can compare the results to the data from TCIA/TCGA dataset.

The LIDC-IDRI dataset contains the CT image data 1010 patients. At the same time, four radiologists marked the boundaries of the lung nodules they found for each case. For each nodule, the degree of malignancy is scored, and the score from 1 to 5 corresponds to the degree of benign and malignant from low to high. The benign and malignant scores of radiologists are judged based on their experience and may be different from the real benign and malignant results of lung nodules. Only 157 of the 1010 patients had clinical data. Among them, 27 patients had no benign or malignant results, 36 patients had benign nodules or other non-malignant diseases, 43 patients had primary lung cancer, and 51 patients had metastatic cancer. The clinical data comes from long-term follow-up results, tissue biopsy results, or surgical results. Data integration in medical applications typically involves several steps: data consolidation, virtualization, and propagation. Here, since we are using a well-organized dataset, we can skip the steps above. But in other use cases where the data is not standardized, these steps would be necessary. In our use case, we emphasize the role of using deep learning in medical data integration. Deep learning can extract high-level information from data, thus can be used to integrate various kinds of data. For instance, in our case, it can be used to combine medical image data (CT scans) with clinical data of patient symptoms, as well as genomic data like DNA sequencing reads and transcriptome profiling.

We will use clinical data and average radiologist scores as two gold standards to train and test the model in this paper. We excluded patients with missing benign and malignant results and patients with metastatic cancer in the clinical data. At the same time, we assume that benign nodules or nodules in patients with other non-malignant diseases are benign, and those lung nodules in patients with primary lung cancer are malignant. The average radiologist score is the average of all radiologists' scores on the benign and malignant degree of the same lung nodule. If the average radiologist score is less than or equal to 3, it is considered benign, and if the average radiologist score is more than 3, it is considered to be malignant. Combining the patient's clinical data and

the boundary markers of lung nodules, CT images of 127 lung nodules have corresponding clinical standards. Combining the benign and malignant degree score and the boundary markers of lung nodules, the CT images of 2391 lung nodules have the corresponding radiologist's gold standard for scoring.

Because the data in the LIDC-IDRI dataset is collected from different institutions, and different CT machines are used for collection, there are differences in lateral resolution, layer thickness, and filtering methods in the original CT images, which brings difficulties to feature extraction. However, such data can realistically simulate the complex and changeable CT results in reality, which is a challenge. While there are differences on data due to its source, the quality of the data is maintained at a reasonable standard by the dataset manager. Thus, the data does not need extensive processing before it can be utilized. In this paper, a rather simple preprocessing is performed on the image, and the purpose is to normalize the size of the original CT images. The original image is converted into an isotropic image by linear interpolation so that the resolution of the image in the three directions is 1 mm. We are performing a sample of the LIDC-IDRI dataset in Fig. 5. Fig. 5. (a) shows some of the annotations marked by the radiologists, such as texture, malignancy of the nodule, etc. Fig. 5. (b) shows an example of the CT scan with the nodule of interest circled in red. Please note that one patient may have one or more nodules of interest. Then in Fig. 5. (c), we demonstrate the 3D layer view of the nodule and its dimensions, constructed by using multiple layers of the CT scans. Finally, in Fig. 5. (d), we show the annotations of the four radiologists and perform consensus consolidation and pad slices for visualization, which we used as the input of our model.

TCGA project was jointly launched by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) in the United States. TCGA mainly collects clinical data, genome mutations, mRNA expression, miRNA expression, methylation, and other data on various human cancers (including subtypes) and is an important data source for cancer researchers [27]. A screenshot from the TCGA dataset patient case file is shown in Fig. 6. As shown in the figure, for each record, it contains various types of data, including DNA related sequencing data, clinical records, biospecimen, etc., and in this paper we use the biospecimen data. Currently, TCGA has studied 36 cancer types in total. TCGA uses large-scale genome analysis technology to understand the molecular mechanism of cancer through extensive cooperation. TCGA aims to deepen our understanding for cancer through molecular characterizations, establish a rich genomics data resource for the broad research community, help advance health and science technologies, and change how cancer patients are treated in the clinical. TCGA hosts a massive clinical, genetic, and pathological data. In contrast, the corresponding radiological image data, such as the CT scans, are stored on TCIA. The TCIA dataset contains detailed image data of the cancer patients recorded in TCGA, as well as their pathology reports. A sample tissue specimen is shown in Fig. 7 with its pathology report. The

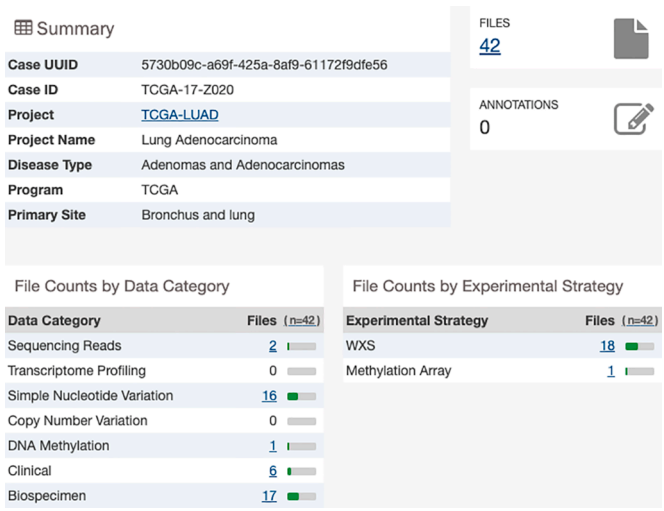

Fig. 6. A sample data file of the TCGA dataset.

LIDC-IDRI dataset is a subset of TCIA/TCGA dataset that consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. We do want to show a sample of the type of data available from the TCGA dataset once we find a match using our similarity algorithm.

In our paper, we first use the data from LIDC-IDRI to help train and test the detection model, then use the data from TCIA as a diagnosis history archive and compare the detected positive lung cancer cases in the detection model with the instances in the TCIA. Then, we select the most similar cases in TCIA and find the corresponding clinical, genetic, and pathological data from TCGA to recommend them to users as references for detection, treatment, and prognosis.

4.2. Evaluation metrics

To evaluate the overall detection effect of the proposed system, we can analyze the system performance to understand the characteristics of results. Since the detection results are only a possibility of a region of the lung CT image to have cancer, we cannot simply use a dichotomous result to verify the effectiveness of the system, and a better solution is needed. Some scholars use the technique of comparing the distance of the center point, that is, the distance between the center of the detected region in the image and the center of the true nodule is compared less than the same threshold. If it is less than this threshold, it is considered that the true nodule has been detected by the detection system. Some scholars set this threshold as the radius of the nodule to be detected, while others set this threshold as a constant, such as 5 mm. The method adopted in this paper is: if the detected nodule overlaps the true nodule,

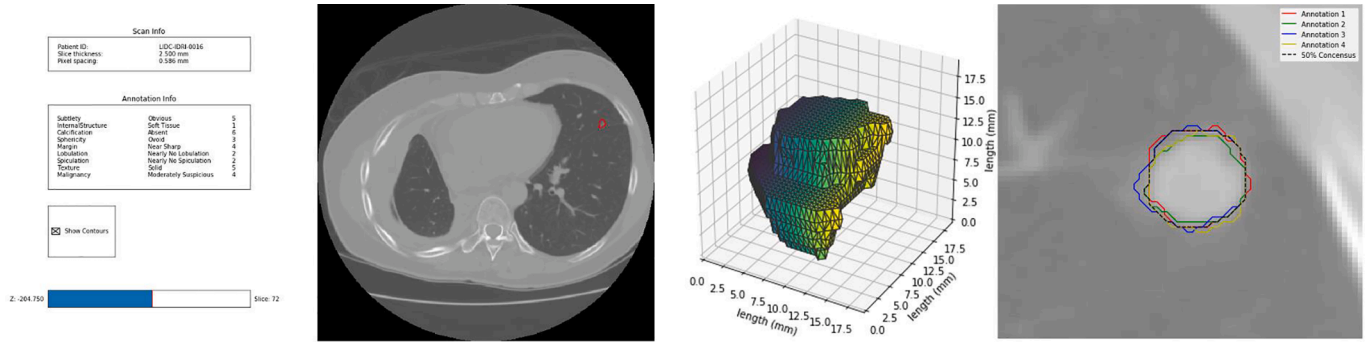

Fig. 5. (a) annotations marked by the radiologists. (b) an example of the CT scan with the nodule of interest circled in red. (c) the 3D layer view of the nodule and its dimensions, constructed by using multiple layers of the CT scans. (d) annotations of the four radiologists, and their average.

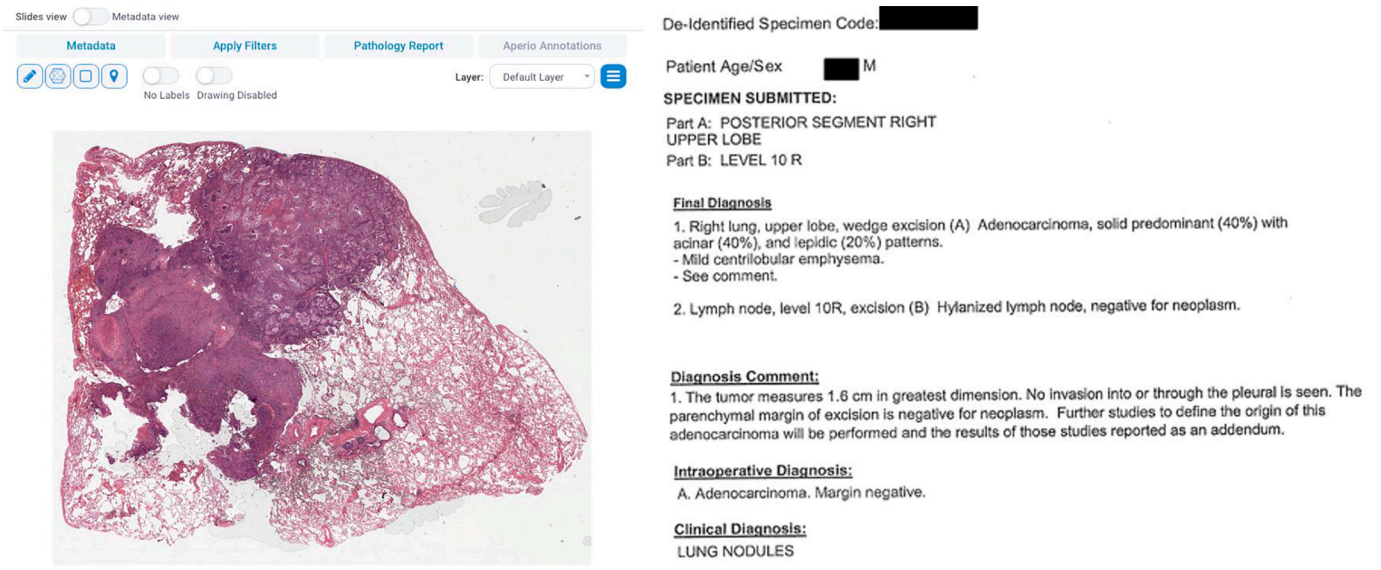

Fig. 7. A sample tissue specimen with its pathology report from the TCIA/TCGA dataset.

the detected nodule is considered as an actual positive test. Otherwise, this nodule is regarded as a false positive.

To measure whether the lung nodules in the patient's body are all detected, scholars generally use the accuracy to measure the detection accuracy. Accuracy is defined by the following:

True Positive (TP)=the number of true positive nodules that are judged as positive nodules by the detection system.
True Negative (TN)=the number of true negative nodules that are judged as negative nodules by the detection system.
False Positive (FP)=the number of true negative nodules that are judged as positive nodules by the detection system.
False Negative (FN)=the number of true positive nodules that are judged as negative nodules by the detection system.

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

We assume the output result of the detection system is a probability rather than a binary decision of zero or one. In that case, the ROC curve can be used to measure the performance of the detection system. The x-axis of the ROC curve is false positive, and the y-axis is true positive. The curve tests different thresholds to balance true and false positive rates, and how they change as the threshold changes. Test samples over the threshold are judged as true nodules, and those below the threshold are considered not a nodule. Every time a threshold is taken, a set of true and false positives are obtained, and the ROC curve is obtained by connecting these points. The quantitative indicator for the ROC curve is Area Under the Curve (AUC). The larger the AUC value, the better the detection effect. It is generally believed that an AUC below 0.7 indicates poor detection performance, a value between 0.7 and 0.85 is moderate, and a value over 0.85 is excellent. During the detection of lung nodules, a lot of false positives may be generated. The detection system should have low false positives while obtaining high true positives.

5. Results and analysis

5.1. Detection and similarity comparison results

To test the effectiveness of the cloud-based deep learning model for lung nodule detection proposed in this paper, we will now select patient data in the test set for verification. We used ROC to perform statistical analysis on the test by calculating AUC to show its performance. In order to prove the proposed model in this paper, we also compared our model to other baseline models like faster RCNN and Multi View CNN to detect the same dataset.

Table 1 is the accuracy comparison that shows the results of the three deep learning models in detecting lung nodules on the LIDC-IDRI dataset. From Table 1, it can be seen that the model proposed in this paper performs well in the task of lung nodule detection, and its overall accuracy is better than that of baseline methods. As the number of false positives tolerance increases, so does the accuracy. When the average number of false positives in each group of CT images is 10, the accuracy reaches 95.6%. At the same time, looking at Fig. 8, we can see that our proposed method has the highest AUC, which is 0.87, higher than that of baseline methods. Therefore, the proposed method outperforms baseline methods in feature extraction ability and discrimination ability, which can effectively lay the foundation for the follow-up detection of doctors and achieve the purpose of assisting doctors.

Then we compare the detected lung nodule with them in the TCIA/TCGA dataset. We selected four similar cases in the TCIA/TCGA dataset for the chosen lung cancer and displayed them in Fig. 9. In the figure, the first row and column that are border in green is the case image we randomly selected from the test set of the LIDC-IDRI dataset. We call it target case. Then the other ones are the images chosen from the TCIA/TCGA dataset that is similar to it. The numbers in the matrix represent the cosine similarity between the image in the corresponding row with the image in the corresponding column. As we can see, target case has a

Table 1
Accuracy with different tolerance for false positive.

False Positive Tolerance (#)	Faster RCNN	Multi View CNN	Our Proposed Method
0	0. 640	0.715	0.742
1	0.696	0.763	0.796
2	0.753	0.798	0.816
3	0.786	0.823	0.861
4	0.824	0.837	0.893
5	0.861	0.884	0.924
6	0.876	0.913	0.935
7	0.894	0.924	0.948
8	0.906	0.930	0.952
9	0.913	0.939	0.959
10	0.920	0.941	0.956

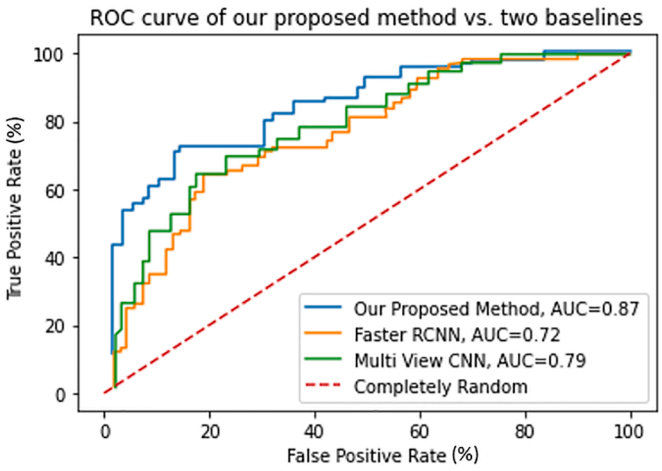

Fig. 8. ROC curve of our proposed method vs. two baseline methods.

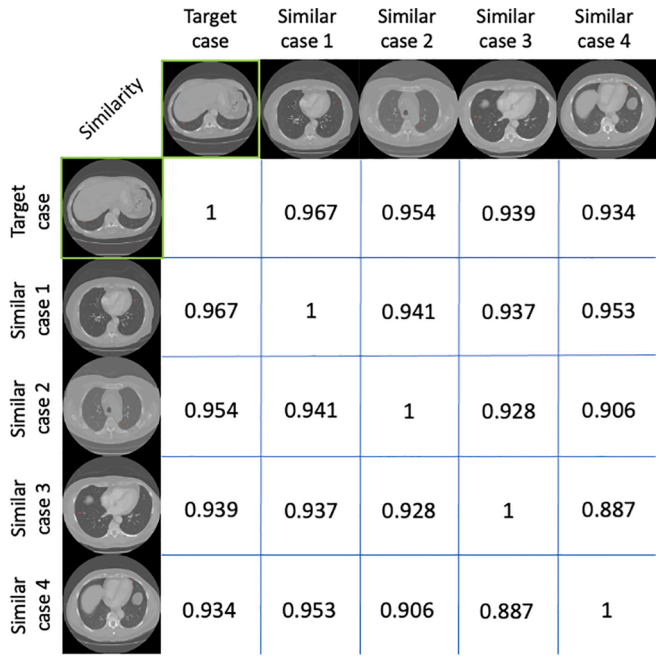

Fig. 9. The sample target lung cancer from the LIDC-IDRI dataset (top left with green border) and its top four similar cases from the TCIA dataset, with similarity matrix showing the similarity between target case and similar cases.

similarity to itself that is infinitely close to 1, which is expected. Then the other cases selected almost all have a high similarity over 0.9.

5.2. Feature importance

We then want to explore how the detection model makes its decisions based on the features of the nodules. In addition to the 3D image features, each of the nodules also has its own corresponding features that are marked by the radiologists. We aggregated the results of in the test set in the LIDC-IDRI dataset, and for every nodule of each patient, we plot how they impact the model result. We show the results in Fig. 10. As we can see from the figure, volume, calcification, diameter, spiculation, and surface area are the top five features that impact the model output. In comparison, the sphericity, texture, and internal structure have a much smaller impact on the model output. This is in accordance with the understanding of lung cancer nodules. Then to further investigate the effectiveness of our model, we also zoom in on one randomly selected

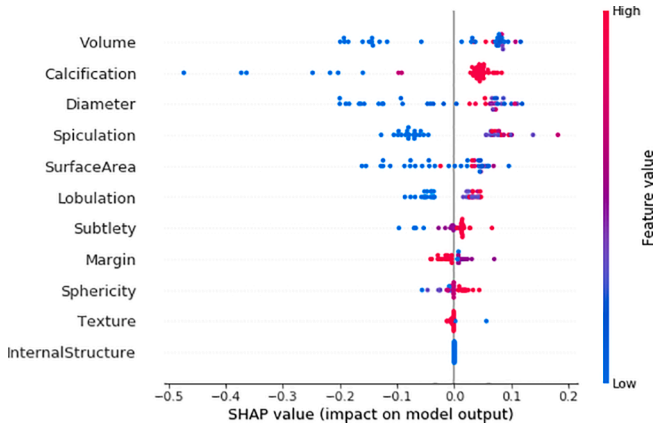

Fig. 10. SHAP value swarm plot of each feature's impact on the model outcome.

individual nodule. In Fig. 11, we show the waterfall plot for the impact of each feature on the decision made for that particular nodule, help explain how much each factor contributed to the model's prediction. Large positive/negative values indicate that the feature had a significant impact on the model's prediction. In our case, the x-axis shows the malignancy, y-axis shows the value of each feature, and the red/blue arrow shows how each feature positively/negatively impacts the malignancy level. For instance, for this particular nodule, its texture is 4, and the texture increased the malignancy by 0.03. As we can see from the figure, the results are roughly similar to those shown in Fig. 10. Calcification, spiculation, and volume are the top three features impacting the model output.

5.3. Overall running efficiency

The model we selected for lung cancer detection is relatively large and thus requires a lot of resources for training on a single machine. We used cloud-based multi-processing and batch prediction to reach its running efficiency requirement during the feature extraction process. The model training is distributed over several GPU-equipped instances, which is more efficient than CPU instances. By parallelizing the process, the training process runs in parallel, thus much faster. For our case, we used 4 GPUs in cloud computing instances to train the model, which improved the speed by 320%. The efficiency increase with the number of GPUs is not proportional but strongly correlated. This is due to unavoidable overhead in processing the data, data ingestion, etc. Then for

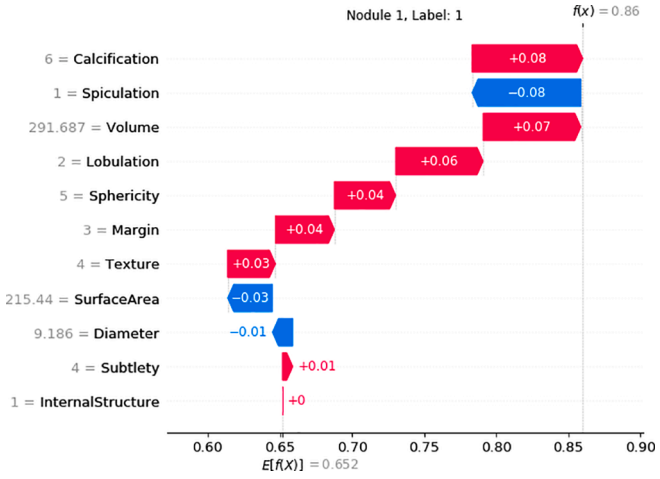

Fig. 11. Individual nodule level feature importance exploration with waterfall plot.

similarity analysis, we used a matrix-wise operation to avoid the low-efficiency condition of Python.

Then for accessing our model and collecting results, we have an ingestion system that takes standardized CT scan DICOM files as input, process the CT scan image series, and input into the model. Then if the detection model determines positive for cancer, the most similar TCIA/TCGA case IDs are retrieved and returned. The case IDs are used to retrieve diagnosis information, treatment plan, genomic data, etc., that can be used to help give better treatment plans.

6. Conclusion

In this paper, to leverage the rapid development of massive medical data and computer-assisted detection as an opportunity in the medical industry 4.0, we propose a data integration system that includes cloud-based deep learning model for lung cancer detection and historical similar case finding. Our proposed system takes CT scans as input and uses deep learning to detect whether patients have lung cancer. Then if the case is deemed positive, the system will search for similar cases in the TCIA dataset. With the case in TCIA dataset, we can find its history and a large amount of information in the corresponding TCGA dataset. With the cloud-based system, even doctors with limited computational resources and training can get recommendations on their patients, refer to historical cases and propose more fact-based diagnoses and treatment plans for their patients. From end-to-end, data is integrated so that heterogeneous data can be extracted to produce standardized results. This process proves that data integration between different datasets and different data types is very important. Then a lot of experiments are used to prove the effectiveness of the detection system and present the similarity comparison. The results show the proposed system has a strong capability.

However, we know that although the use of cloud-based deep learning can fully capture the time and space feature information in the volume data of medical images, at the same time, they have some limitations. As the dimensionality increases, the model parameter also significantly increases. If the sample size is insufficient to support the parameter, over-fitting is prone to happen. Therefore, future work will explore more advanced methods to alleviate the possible over-fitting issue during the experiment. In addition, in this paper, we focus on the supervised learning type of deep learning, using CT images to identify whether the patient has lung cancer. But for other potential use cases, unsupervised learning can see great potential when dealing with an unlabeled dataset as well. This mainly focuses on identifying the structure of the unknown dataset, while grouping similar data points together. It can be very helpful in tasks like gene expression identifications or cancer molecular subtyping comparisons. For future work, integrating this type of data will be another field that can greatly benefit the researchers and help improve the performance of models.

Credit authorship contribution statement

Chang Gu: Conception and design; Collection and assembly of data; Data analysis and interpretation; Manuscript writing; Final approval of manuscript.

Chenyang Dai: Collection and assembly of data; Data analysis and interpretation; Manuscript writing; Final approval of manuscript.

Xin Shi: Collection and assembly of data; Data analysis and interpretation; Manuscript writing; Final approval of manuscript.

Zhiqiang Wu: Data analysis and interpretation; Manuscript writing; Final approval of manuscript.

Chang Chen: Administrative support; Conception and design; Manuscript writing; Final approval of manuscript.

Declaration of Competing Interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Data Availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81802256 and 81902335), Shanghai Science and Technology Fund (20S11900600), Shanghai Rising Star Program (20QA1408300) and Clinical Research Plan of SHDC (SHDC2020CR4028).

References

[1] Y. Adie, D.J. Kats, A. Tlimat, et al., Neighborhood disadvantage and lung cancer incidence in ever-smokers at a safety-net healthcare system: a retrospective study, Chest (2019), https://doi.org/10.1016/j.chest.2019.11.033.
[2] xxx 2022 https://blog.cambridgesemantics.com/can-graph-integrate-data-at-scale-hint-yes-but-the-answer-isnt-what-you-think.
[3] L. Wang, D. Rajan, An image similarity descriptor for classification tasks, J. Vis. Commun. Image Represent. 71 (2020) article. 102847.
[4] S.G. Armato, G. McLennan, L. Bidaut, et al., The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans, Med. Phys. 38 (2) (2011) 915–931.
[5] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, J. Digit. Imaging 26 (6) (2013) 1045–1057.
[6] B. Golosio, G. Masala, A. Piccioli, P. Oliva, et al., A novel multithreshold method for nodule detection in lung CT, Med. Phys. 36 (8) (2009) 3607–3618.
[7] W. Wang, et al., Nodule-plus R-CNN and deep self-paced active learning for 3D instance segmentation of pulmonary nodules, IEEE Access 7 (2019) 128796–128805.
[8] Y. Xie, et al., Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT, IEEE Trans. Med. Imaging 38 (4) (2019) 991–1004.
[9] H. Dhayne, R. Haque, R. Kilany, Y. Taher, In search of big medical data integration solutions - a comprehensive survey, IEEE Access 7 (2019) 91265–91290.
[10] G. Aceto, V. Persico, A. Pescapé, Industry 4.0 and health: internet of things, big data, and cloud computing for Healthcare 4.0, J. Ind. Inf. Integr. 18 (2020) article. 100129.
[11] M. Tsiknakis, et al., A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of postgenomic clinical trials on cancer, IEEE Trans. Inf. Technol. Biomed. 12 (2) (2018) 205–217.
[12] A.M. Fathollahi-Fard, M.A. Dulebenets, M. Hajiaghaei–Keshteli, R. Tavakkoli-Moghaddam, M. Safaeian, H. Mirzahosseinian, Two hybrid meta-heuristic algorithms for a dual-channel closed-loop supply chain network design problem in the tire industry under uncertainty, Adv. Eng. Inform. 50 (2021) article. 101418.
[13] X. Jiang, Z. Tian, W. Liu, et al., Energy-efficient scheduling of flexible job shops with complex processes: a case study for the aerospace industry complex components in China, J. Ind. Inf. Integr. 27 (2021) article. 100293.
[14] H. Zhao, C. Zhang, An online-learning-based evolutionary many-objective algorithm, Inf. Sci. 509 (2020) 1–21.
[15] R. Alkurd, I.Y. Abualhaol, H. Yanikomeroglu, Personalized resource allocation in wireless networks: an AI-enabled and big data-driven multi-objective optimization, IEEE Access 8 (2020) 144592–144609.
[16] R.J. Chen, M.Y. Lu, T.Y. Chen, et al., Synthetic data in machine learning for medicine and healthcare, Nat. Biomed. Eng. 5 (6) (2021) 493–497.
[17] J. Pasha, M.A. Dulebenets, M. Kavoosi, O.F. Abioye, H. Wang, W. Guo, an optimization model and solution algorithms for the vehicle routing problem with a "factory-in-a-box, IEEE Access 8 (2020) 134743–134763.
[18] D. Ardila, A.P. Kiraly, S. Bharadwaj, et al., End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nat. Med. 25 (6) (2019) 954–961.
[19] P. Huang, C.T. Lin, Y. Li, et al., Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method, Lancet Digit. Health 1 (7) (2019) e353–e362.
[20] Y. Guo, Q. Song, M. Jiang, et al., Histological subtypes classification of lung cancers on CT images using 3D deep learning and radiomics, Acad. Radiol. 28 (9) (2021) e258–e266.
[21] T. Lustberg, J. van Soest, M. Gooding, et al., Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer, Radiother. Oncol. 126 (2) (2018) 312–317.
[22] M. Tortora, E. Cordelli, R. Sicilia, et al., Deep reinforcement learning for fractionated radiotherapy in non-small cell lung carcinoma, Artif. Intell. Med. 119 (2021) article. 102137.

[23] Y. Xiao, J. Wu, Z. Lin, Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data, Comput. Biol. Med. 135 (2021) article. 104540.
[24] L. Ubaldi, V. Valenti, R.F. Borgese, G. Collura, et al., Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples, Phys. Medica 90 (2021) 13–22.
[25] T.T. Ho, T. Kim, W.J. Kim, et al., A 3D-CNN model with CT-based parametric response mapping for classifying COPD subjects, Sci. Rep. 11 (1) (2021) 1–12.
[26] A. Kumar, S. Dyer, J. Kim, C. Li, et al., Adapting content-based image retrieval techniques for the semantic annotation of medical images, Comput. Med. Imaging Graph. 49 (2016) 37–45.
[27] xxx 2022 https://www.cancer.gov/tcga.