IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 3, March 2025

Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches. (Python)

Hrutvij Thorat[1], Ankit Sangale[2], Gokul Mahajan[3], Aryan Kale[4] ,Prof. Miss. T. S. Pawar[5]

Students, Computer Engineering[1,2,3,4]

Guide. Computer of Engineering[5]

S.P.I.T. Polytechnic, Kurund, Ahmednagar, Maharashtra, India.

Abstract*:* *Risk stratification (characterization) of tumors from radiology images can be more accurate and faster with computer-aided diagnosis (CAD) tools. Tumor characterization through such tools can also enable non-invasive cancer staging, prognosis, and foster personalized treatment planning as a part of precision medicine. In this papet, we propose both supervised and unsupervised machine learning strategies to improve tumor characterization. Our first approach is based on supervised learning for which we demonstrate significant gains with deep learning algorithms, particularly by utilizing a 3D convolutional neural network and transfer learning. Motivated by the radiologists' interpretations of the scans, we then show how to incorporate task-dependent feature representations into a CAD system via a graph-regularized sparse multi-task learning framework. In the second approach, we explore an unsupervised learning algorithm to address the limited availability of labeled training data, a common problem in medical imaging applications. Inspired by learning from label proportion approaches in computer vision, we propose to use proportion-support vector machine for characterizing tumors. We also seek the answer to the fundamental question about the goodness of "deep features" for unsupervised tumor classification. We evaluate our proposed supervised and unsupervised learning algorithms on two different tumor diagnosis challenges: lung and pancreas with 1018 CT and 171 MRI scans, respectively, and obtain the state-of-the-art sensitivity and specificity results in both problems.*

Keywords: Risk stratification

I. INTRODUCTION

The advent of deep learning has significantly advanced the medical field, particularly in the early detection and characterization of tumors. This paper focuses on the use of deep learning techniques to enhance the characterization of lung and pancreatic tumors, two of the most challenging and prevalent cancers. Accurate and timely identification of these tumors is crucial for effective diagnosis, treatment planning, and patient prognosis.

Current methods primarily rely on supervised learning approaches, which require labeled data for training models. However, unsupervised learning techniques are emerging as valuable alternatives, providing insights from unlabeled data, which is abundant in medical imaging. By combining both supervised and unsupervised learning methods, the goal is to improve tumor detection, reduce the reliance on large labeled datasets, and enhance the accuracy and efficiency of diagnosis.

The integration of these advanced techniques has the potential to not only increase diagnostic accuracy but also enable personalized treatment plans, ultimately improving patient outcomes. Through innovative deep learning models, this research aims to revolutionize cancer detection and pave the way for more automated and reliable healthcare solutions.

DOI: 10.48175/IJARSCT-23918

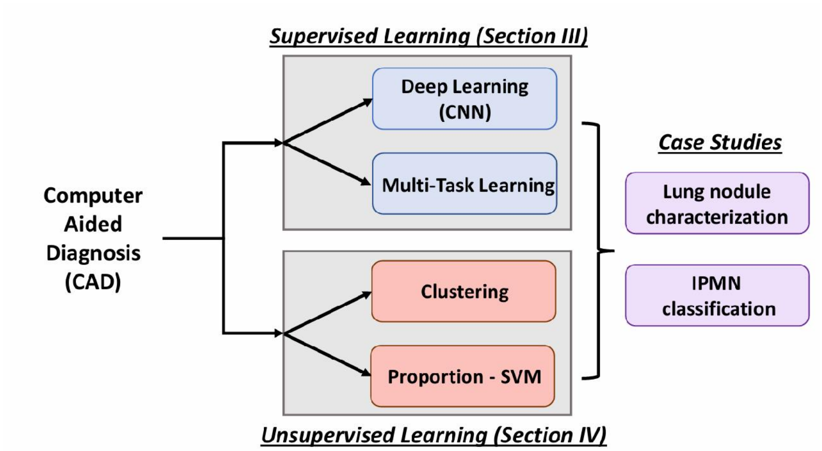

Supervised Learning (Section III)
Deep Learning (CNN)
Multi-Task Learning
Computer Aided Diagnosis (CAD)
Clustering
Proportion - SVM
Unsupervised Learning (Section IV)
Case Studies
Lung nodule characterization
IPMN classification

Figure 1: A block diagram that demonstrates the many plans, techniques, and experimental case studies presented in this paper. To describe tumors, we create both supervised and unsupervised learning systems. We conduct inspections for lung nodule characterization from CT scans and suggest a novel 3D CNN architecture based on Graph Regularized Sparse Multi-task learning for the supervised learning scheme.

We provide a novel clustering technique, SVM, for unsupervised learning schemes and evaluate it for the classification of pancreatic cysts (IPMN) from MRI images and lung nodules from CT scans.

II. RELATED WORK

The developments in machine learning applied to CAD and medical imaging systems created for the diagnosis of lung cancer are compiled in this section. Relevant works are primarily chosen from the clinical trials because the automatic characterization of IPMN from MRI scans has not been thoroughly investigated in the literature. In this sense, our work is the first.

Imaging characteristics and Classifiers: Traditionally, nodule segmentation, the calculation and selection of low-level characteristics from the image, and the application of a classifier/regressor may be necessary for the risk stratification (classification) of lung nodules.

In the method by class labels were produced using Artificial Neural Networks and other physical statistics, such as intensity measures, were recovered. In spherical harmonics were used for shape analysis after appearance-based models were used to segment lung nodules. Classification based on k-nearest neighbors was the final step. Local Binary Patterns, Gabor, and Haralick are examples of 2D texture characteristics that were extended to 3D via another approach.. The last stage involved classification using Support Vector Machines (SVM). In a separate investigation, Way et al. used 3D active contours to segment nodules before applying the rubber band straightening transform. To obtain class labels, a Linear Discriminant Analysis (LDA) classifier was used.A feature selection-based method using imaging and clinical data was presented by Lee et al. A combination of the random subspace approach and evolutionary algorithms was used to measure the information content and feature relevance. The final classification of the condensed feature set was then obtained using LDA. A recent study combined deep learning features with spherical harmonics features. after which lung nodule characterization was done using RF classification. CNN has only been used in 2D space up to this point which prevents it from adding crucial volumetric and contextual information for nodule characterisation. CNN was used in a different method by Shin et al. to classify lung nodules. In addition to not being fully 3D CNN, the method required training an off-the-shelf classifier like RF and SVM and ignored high-level nodule properties.Lung nodule malignancy characterisation has been demonstrated to benefit from the knowledge of many high-level imaging features. In a study examining the relationship between nodule characteristics and malignancy, discovered that in a given dataset, 82% of the lobulated, 93% of the ragged, 97% of the heavily spiculated, and 100% of the halo nodules were cancerous. In the automatic identification of lung nodule characteristics and kinds was investigated. The aim was to classify six distinct nodule types, including calcified, perifissural, solid, non-solid, part-solid, and spiculated nodules.However, the approach is based on 2D CNN and fell short The detection of lung nodules has also been a topic of active interest among researchers and A brief but instructive review of the most recent detection studies can be found in . Additionally, 66% of the round nodules were found to be benign when estimating the malignancy of lung nodules.

Pancreatic Cysts (IPMN): Despite significant advancements in automatic methods for segmenting the pancreas and its cysts there is limited use of sophisticated machine learning algorithms for fully automatic risk-stratification of IPMNs. Hanania et al. examined the impact of 360 imaging features, such as intensity, texture, and shape, in classifying subjects as having low or high-grade IPMN. In a different example, Gazit et al. extracted texture and features from the solid component of segmented cysts, followed by a feature selection and classification scheme. Both of these methods required segmentation of the cysts or pancreas and were assessed solely using CT scans.

Unsupervised Learning: Labeled data (supervision) is usually used to solve the visual identification and classification tasks. However, the use of unsupervised learning techniques is very beneficial for applications where it is time-consuming and costly to manually generate labels matching to huge datasets.

Problems in a variety of fields, including object classification speech processing and audio classification have been resolved through the application of unsupervised techniques. Traditionally, these techniques used supplementary information that was included with the data to enhance learning, which would not be accessible for a number of medical imaging classification problems.

Various methods for detecting and diagnosing issues in medical imaging have made use of unsupervised learning. Stacking au to encoders was the method employed by Shin et al. to detect various organs in MRI data. A brain tumor segmentation technique using a stacked denoising autoencoder was demonstrated by Vaidhya et al. and assessed on multi-sequence MRI images. In a study by Sivakumar et al. , unsupervised clustering techniques are used to segment lung nodules. Kumar et al. classified lung nodules using autoencoder characteristics in a different research. Nevertheless, these auto-encoder methods failed to produce adequate categorization outcomes. In addition to this, unsupervised deep learning has been investigated for breast density segmentation and mammography risk prediction .

III. METHODOLOGY

A. LIDC Dataset: Extraction of Data

A database containing the thoracic CT scans of 1010 patients with lung malignancies and annotations (nodule outlines) from up to four radiologists is publicly accessible thanks to the Lung Image Database Consortium (LIDC). Although more than a thousand patients have annotations accessible, only 157 patients have diagnostic data that includes nodule ratings (0-Unknown, 1-Benign, 2-Primary Malignant, 3-Malignant (metastatic)). The ratings were acquired through biopsy, surgical resection, progression, or radiological image assessment to demonstrate the nodule status throughout two years at two levels: first at the first diagnostic at the nodule level and then at the patient level. Thoracic CT scans for 1010 people were gathered over an extended period of time using a variety of scanners to create the LIDC database. A varied number of slices may be linked to a certain nodule study because a single study may have several nodules associated with it. Additionally, up to four radiologists' annotation data for a nodule may be included in each slice.

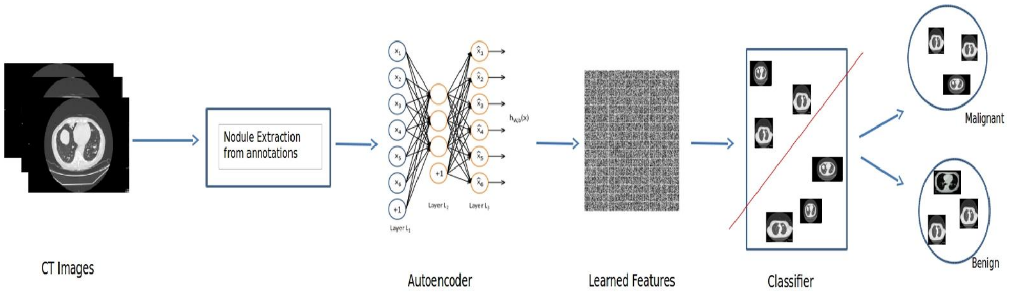


Figure 2. Illustration of different modules of the proposed CAD system.

Only nodules larger than 3 mm in size have these annotations available in the LIDC dataset. We decided to utilize the ratings from diagnostic data as the ground truth for training the classification system and assessing the outcomes rather than utilizing the radiologists' ratings in the dataset because the diagnostic data is the sole way to determine the certainty of malignancy.

B. Autoencoder: Extraction of Features

An autoencoder represents a two-layer network that "encodes" data to a latent space using a linear or non-linear transformation after receiving an input f [01]d. It reconstructs the data on the output layer by a "decoding" execution. Specifically stating that we wish to pick up a representation:

$l(f(i); Wb) = (Wf(i)+b)$ (1)

such that $(WTl(f(i);Wb)+c)$ is approximately $f(i)$.

$\min \; W,b,c\sum i=1n\|\phi(WT\phi(Wf(i)+b)+c)-f(i)\|22$ (2)

The l-2 norm is used to penalize the inaccuracy between the input and output. Sigmoid or hyperbolic tangent functions are typically employed as the autoencoder's activation function. Accurately determining the cost function in an autoencoder opens the door to the use of more sophisticated optimization techniques, like L-BFGs , for network training.

Pre-training the filters greedily enables a method to evade the optimization objective by beginning from a point that is probably closer to the optima, since the optimization function in deep neural networks may be non convex .TheThe choice of network setup settings (batch size, iteration set, number of hidden layers, etc.) appears to be fairly ad hoc, and study on the subject is currently ongoing.

The next part provides a detailed explanation of how to extract deep features using an autoencoder.

IV. CAD SYSTEM OVERVIEW

The literature has a full description of our CAD system. 17 Here is a brief synopsis, and Fig. 3 displays the f lowchart. First, the CT scan's VOI-containing nodule was recognized by the radiologist. For every scan, we produced isotropic voxels using linear bilinear interpolation in the axial plane otherwise, or interpolation in the z direction if the slice interval exceeded the pixel size.

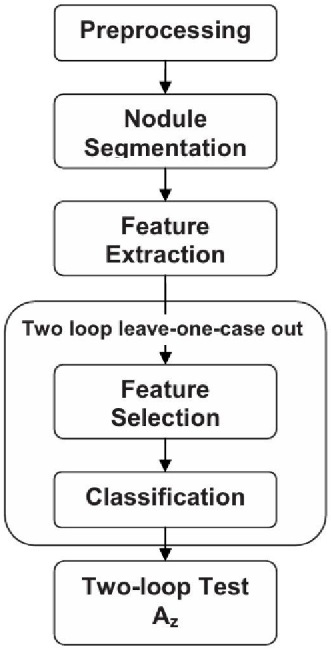


FIG. 3. A schematic showing the major image processing steps of the CAD system. The two-loop resampling scheme is described in Fig. 4.

Facilitating initial contour generation and segmentation is the aim of interpolation. Spatial resolution would not be enhanced by the interpolation. Voxels that were not a part of the chest wall or mediastinum in the VOI were allocated to the object or background class using k-means clustering in order to create the initial contour. As previously mentioned, a lung mask was used to remove the mediastinal or pleural voxels from this and future feature extraction procedures.

A spherical structuring element with an autonomously calculated size dependent on the size of the clustered item was used to conduct morphological opening. The object's blood vessels and other attachments may be removed by the morphological opening. The nodule in the VOI was then segmented using a 3D AC model. Using classification performance as the figure of merit from our earlier study, we computed the weights for the 3D AC using the optimization method. For the performance comparison detailed below, the segmentation was optimized independently using the feature space with and without the new picture features. Following optimization, all of the nodules for the specified feature space were segmented using the same set of weights.

As explained in Sec. II D, we added additional features to the feature space for this investigation. The best features for the classification job were then chosen by applying a feature selection approach to the multidimensional feature space. The chosen features were used to train a feature classifier. Test cases were used to assess the trained classifier's performance, and ROC analysis was used to measure the classification accuracy.

Demographic features:

We looked into the impact of patient variables, such as gender and age at scan time, as supplementary data for the CAD system. Gender was represented as 1 for males and 0 for females, and age was an integer value.

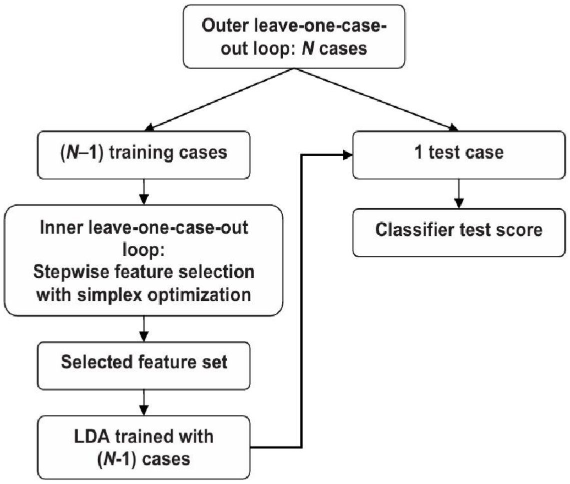

Outer leave-one-case-out loop: N cases
(N–1) training cases
1 test case
Inner leave-one-case-out loop: Stepwise feature selection with simplex optimization
Classifier test score
Selected feature set
LDA trained with (N-1) cases

FIG. 4. In the outer leave-one-case-out loop, the data set is divided into N−1 training cases and 1 test case. For each N−1 case cycle, an LDA classifier is designed from a set of selected features as a result of an inner leave-one-case-out training and testing scheme. After each case is left out in turn, the two-loop test Az is calculated from the malignancy scores of N test cases. women. Despite the fact that the majority of CAD systems simply employ picture features, it has been discovered that using demographic data is advantageous. This potentially helpful information cannot be included in our data set because we did not regularly collect smoking histories in the patient files.

V. DATASET

Collection of Images: A CNN model was created utilizing 3494 CT images collected from 222 patients with pathologically confirmed pancreatic cancer and 3751 CT images collected from 190 people with healthy pancreas between June 2017 and June 2018. Figure 5 displays the sample photos.

As indicated in Table 1, we separated these images into three phase-based datasets and used tenfold cross validation to assess the approach for binary classification (i.e., cancer or not).

Biomarkers in urine contain: The potential use of urinary biomarkers in the identification of pancreatic cancer has been studied. With few treatment choices and a dismal prognosis, pancreatic cancer is frequently discovered at an advanced stage. Finding trustworthy and non-invasive biomarkers for early detection is therefore essential to enhancing results. The possibility of a number of urine indicators to identify pancreatic cancer has been investigated, including:

1.CA 19-9: A glycoprotein that is presently employed as a biomarker in clinical practice and is frequently raised in individuals with pancreatic cancer.

2. MUC1: Pancreatic cancer has been shown to overexpress this transmembrane mucin protein.

3. Osteopontin: A glycoprotein that can be found in urine and has been demonstrated to be overexpressed in pancreatic cancer.

4. Tumor-associated trypsin inhibitor (TATI): This protein has been studied as a possible biomarker and is frequently high in patients with pancreatic cancer.

5. Human epididymis protein 4 (HE4): A glycoprotein that can be discovered in urine and has been shown to be over expressed in pancreatic cancer.

TFF1 (Trefoil factor 1), REG1B (Regenerating islet-derived protein 1 beta), REG1A (Regenerating islet-derived protein 1-alpha), and LYVE1 (Lymphatic Vessel Endothelial Hyaluronan Receptor 1) biomarkers are used as datasets in this study.

Since LYVE1 expression has been shown to be raised in a number of tumor types, including breast, lung, and pancreatic cancer, it has been studied as a possible target for cancer therapy. According to a 2014 study that was published in the journal Pancreas, individuals with pancreatic cancer had considerably higher serum levels of REG1B than did patients with pancreatitis and healthy controls. REG1B levels were higher in the urine of patients with pancreatic cancer than in healthy controls and patients with chronic pancreatitis, according to a 2017 study that was published in the same journal. TFF1 levels were considerably greater in the serum of patients with pancreatic cancer than in healthy controls and patients with pancreatitis, according to a 2017 study published in the journal PLOS ONE. TFF1 levels in the urine of patients with pancreatic cancer were higher than those of healthy controls and patients with chronic pancreatitis, according to a different 2018 study that was published in the journal Oncotarget.

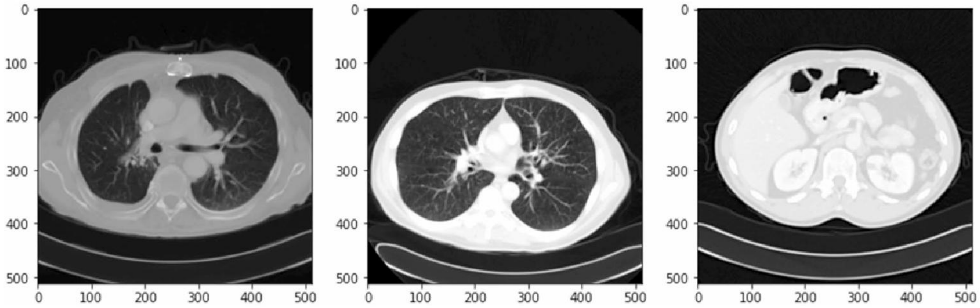


Figure 5. Sample CT scan images from the dataset.

Particulars	Data
Total pancreatic cancer images	3494
Patient with pancreatic cancer	222
Total healthy pancreas images	3451
Patient with healthy pancreas	190
Thickness	5.0 mm

Table 1. Image dataset information.

It has been demonstrated that REG1A is a predictor of pancreatic cancer prognosis, with worse outcomes linked to higher levels of REG1A. The biomarker data from the used dataset are displayed in Fig. 6.

They gathered biomarkers from three different patient groups' urine:

- Indicators of health
- Patients with malignant pancreatic environment and pancreatic ductal adenocarcinoma

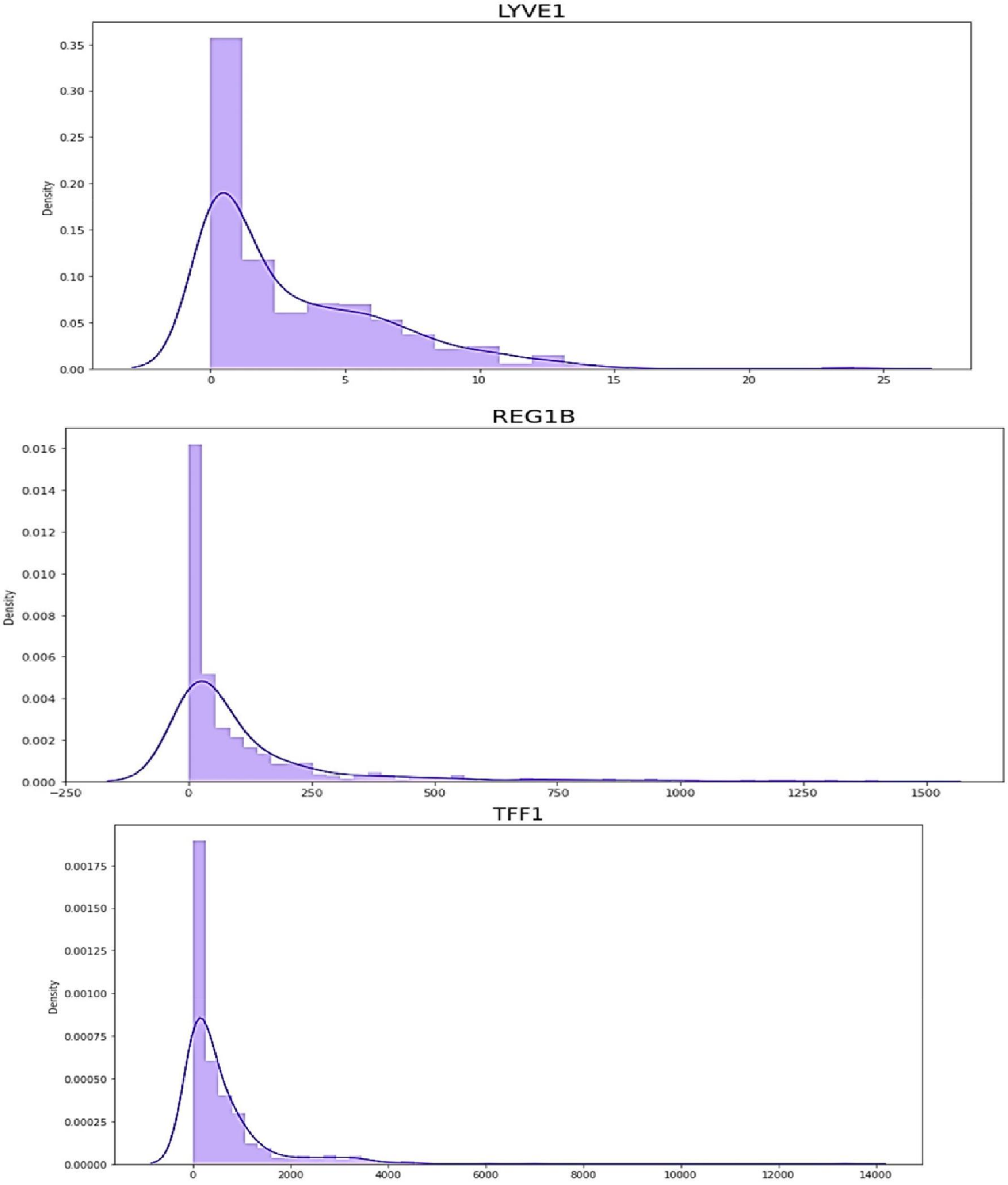


Figure 6. the circulation of data in biomarkers dataset.

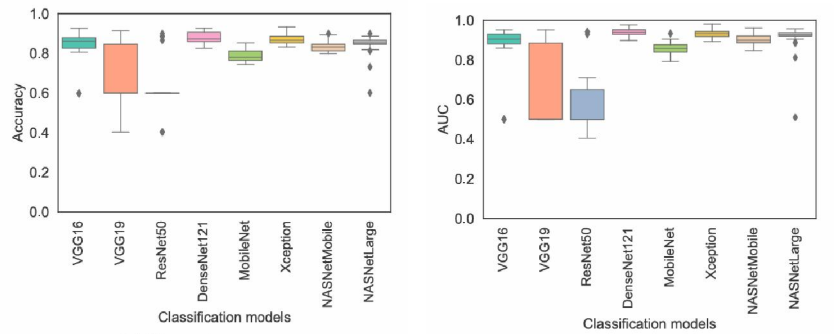
- When precise, they were arranged by sex and age. Finding a trustworthy technique to identify pancreatic cancer was the goal. The dataset data is shown in Fig. 7. This figure displays the dataset, which includes 590 patients' sample_id, patient_cohort, sample origin, age, sex, diagnosis, 199 stages, 208 diagnosis samples, 350 plasma tests, 590 cretinine, and biomarkers.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

This study modifies and applies the most advanced CNN models—such as VGG16, VGG19, ResNet50, DenseNet121, MobileNet, Xception, NASNetMobile, and NASNetLarge—to the lung nodule diagnostic problem. In particular, a fully connected layer replaces the last layers of the models because the original models were developed for standard image datasets, which have different input sizes than the 50×50×5 lung nodule picture size. Additionally, the Z-axis slices are used as the input images' channels when applying the models. Since the Z-axis data is shallower than the X- and Y-axis data, it is possible to classify 3D images by channelizing the Z-axis without significantly raising the computing cost.The CNN models are trained using adaptive moment estimation (Adam) as the parameter optimization algorithm and binary cross-entropy as the loos function. Furthermore, 500 epochs with a batch size of 16 are used to train the models. Five replications and five-fold cross-validation are used to represent the CNN models' classification ability. The dataset's photos are separated into five groups. Four of the partitions are utilized as training datasets and the fifth is used as testing datasets for every validation. Additionally, 12.5% of the training dataset is used for validation.AUC, accuracy, sensitivity, specificity, and precision are used to assess the diagnosis outcomes. The computational specifications for all experiments are 64-bit Windows 10 with an Intel i7 CPU (4.00 GHz), 32 GB of random-access memory (RAM), and an NVIDIA GeForce GTX 1080.

The mean and standard deviation (SD) for the training time and performance outcomes are displayed in Table 1. It is clear from the diagnosis results that DenseNet121 and Xception produce more accurate outcomes than other CNN models. In particular, Xception achieves the maximum sensitivity, whereas DenseNet121 achieves the highest accuracy, specificity, precision, and AUC.On the other hand, whereas other CNN models, such ImageNet [8], may successfully classify ordinary nature photographs, they are unable to accurately identify cancerous lung nodules. Espe-

Classification model	Accuracy (SD)	Sensitivity (SD)	Specificity (SD)	Precision (SD)	AUC (SD)	Training time (minutes) (SD)
VGG16	82.33%	70.73%	90.14%	69.61%	84.94%	12.21
	10.48%	32.01%	5.07%	31.19%	15.77%	0.17
VGG19	71.04%	42.93%	89.94%	40.57%	69.04%	15.64
	14.29%	42.54%	19.83%	40.65%	20.37%	0.29
ResNet50	62.42%	32.80%	82.29%	23.84%	59.78%	26.24
	15.09%	44.93%	36.82%	35.42%	18.19%	0.51
DenseNet121	**87.77%**	80.93%	**92.38%**	**87.88%**	**93.79%**	35.14
	2.88%	6.89%	2.74%	3.67%	2.01%	2.52
MobileNet	78.79%	71.13%	83.95%	75.19%	85.85%	9.06
	3.04%	6.59%	4.47%	4.75%	3.12%	0.29
Xception	87.03%	**82.73%**	89.92%	84.97%	93.24%	21.84
	2.53%	6.23%	3.79%	4.44%	2.11%	1.01
NASNetMobile	83.62%	77.73%	87.58%	80.97%	90.30%	60.42
	2.81%	6.49%	3.31%	3.81%	2.73%	6.41
NASNetLarge	84.13%	73.93%	90.98%	81.91%	90.51%	110.94
	5.95%	17.94%	5.18%	18.11%	8.69%	7.23



(a) Diagnosis accuracy.

(b) Diagnosis AUC.

Figure 7. Lung nodule diagnosis accuracy and AUC

DOI: 10.48175/IJARSCT-23918

VII. CONCLUSION

Using a machine learning algorithm, we were able to identify the liver and lung as early and late recurrence locations, respectively, with high probability. This model's ability to predict recurrence areas should be helpful for PC patients' future care.

REFERENCES

[1] Sarfaraz Hussein, Pujan Kandel, Candice W. Bolan, Michael B. Wallace, and Ulas Bagci , Senior Member, IEEE.Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning ApproachesAccepted for publication in IEEE Transactions on Medical Imaging 2019

[2]Devinder KumarAlexander WongDavid A. ClausiLung Nodule Classification Using Deep Features in CT Images978-1-4799-1986-4/15 $31.00 © 2015 IEEE DOI 10.1109/CRV.2015.25

[3] M. G. Dinesh 1, Nebojsa Bacanin 2*, S. S. Askar 3 & Mohamed Abouhawwash 4,5*Diagnostic ability of deep learning in detection of pancreatic tumour| https://doi.org/10.1038/s41598-023-36886-8

[4] Lung Nodule Diagnosis on 3D Computed Tomography Images Using Deep Convolutional Neural Networks Qianqian Zhang, Haifeng Wang, Sang Won Yoon, Daehan Won, and Krishnaswami Srihari Department of Systems Science and Industrial Engineering, State University of New York at Binghamton, Binghamton, NY 13902, United States

[53D Shape Analysis for Early Diagnosis of Malignant Lung Nodules Ayman El-Baz1, , Matthew Nitzken1, Fahmi Khalifa1, Ahmed Elnakib1, Georgy Gimel'farb2, Robert Falk3, and Mohammed Abo El-Ghar4 1 Bioimaging Laboratory, Bioengineering Department, University of Louisville, Louisville, KY, USA Tel.: (502)-852-5092 aselba01@louisville.edu 2 Department of Computer Science, University of Auckland, Auckland, New Zealand 3 Department of Radiology, Jewish Hospital, Louisville, KY, USA 4 Urology and Nephrology Department, University of Mansoura, Mansoura, Egypt

[6] Lee, M., Boroczky, L., Sungur-Stasik, K., Cann, A., Borczuk, A., Kawut, S., Powell, C.: Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. Artificial Intelligence in Medicine 50(1), 43–53 (2010)

[7]Chen, S., Ni, D., Qin, J., Lei, B., Wang, T., Cheng, J.Z.: Bridging computational features toward multiple semantic features with multi task regression: A study of CT pulmonary nodules. In: MICCAI. pp. 53–60. Springer (2016)

[8] Chen, S., Ni, D., Qin, J., Lei, B., Wang, T., Cheng, J.Z.: Bridging computational features toward multiple semantic features with multi task regression: A study of CT pulmonary nodules. In: MICCAI. pp. 53–60. Springer (2016)

[9]Vaidhya, K., Thirunavukkarasu, S., Alex, V., Krishnamurthi, G.: Multi modal brain tumor segmentation using stacked denoising autoencoders. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 181–194. Springer (2015)

[10] Kallenberg, M., Petersen, K., Nielsen, M., Ng, A.Y., Diao, P., Igel, C., Vachon, C.M., Holland, K., Winkel, R.R., Karssemeijer, N., et al.: Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE transactions on medical imaging 35(5), 1322–1331 (2016)

DOI: 10.48175/IJARSCT-23918