



Prediction of Lung Diseases based on Supervised and Unsupervised Approach

Aishwarya H S¹, Sujatha B R²,

Post Graduation Student, Department of Electronics and Communication Engineering, Malnad College of Engineering, Hassan, India¹

Professor, Department of Electronics and Communication Engineering, Malnad College of Engineering, Hassan, India²

Abstract: Technology has played a vital role in health sector for past decades whether in detection of various diseases from early stage or in case of targeted treatment. In many Lung disease cases it's the same scenario the early stage detection of the diseases increases the survival rate of patient In this project the first step used in order to detect the lung disease based on unsupervised learning algorithm modelling is done by using LocNet and the optimized CNN is used for image dataset to compare. For the second step the images are differentiated based on whether lung tumor is either malignant or benign using supervised learning algorithm. In the project here two sets of images with data are considered to estimate the diseases where one set of data contains around 5,606 images whose size 1024 x 1024 pixel with height and width there are 15 classes and another set of data and images is the "full-dataset" that as 112,120 total images whose size 1024 x 1024 pixel with height and width.

Keywords: Optimized CNN, Supervised learning, unsupervised learning, Lung diseases

I. INTRODUCTION

The world is changing with great pace thus the pressure on human physical and mental health as increased, the global warming as led to various climatic changes, so there has been lot of various environmental effects, and this also increases the risk of diseases for people. One of the issues that we will focus on in this project is lung diseases. About 14.8 million adults have been diagnosed with (COPD) Chronic obstructive pulmonary disease in 2020 caused mainly by either smoking or air pollution, genetic factors. There are many lung diseases people can get, but there are few lung diseases whose affecting intensity can be reduced by early stage diagnosis. With the help technological developments now it much easier than before to detect disease in early stage which helps to precisely save the lives of people and it will lessen the constraint on the health system.

Approximately 94.1 per 100,000 males and 103.6 per 100,000 females will be the diagnosed of cancer in India in 2020. The cancer mortality rate is more in men compared to that of women that is around 189.5 per 100,000 and 135.7 per 100,000. The Lung cancer cases constitute around 6.9% in India which is common considering among all other cancer. One of the major cases of cancer deaths are correlated to lung cancer it only as the survival rate of around 56% for total cases detected within the lungs but unfortunately only about 16% of lung cancer cases will be detected in early stages. Recognition and consideration of attributes of the lung tumors will help in early detection, thus increasing the survival chance of the patient by accurate use of targeted treatment and therapy.

Typically, to increase an accuracy of the diagnosis conducted it's prominent that the nuclear medicine physicians and Radiologists to be aware of the condition by correlating both Positron emission tomography (PET) and Computed tomography (CT) to avoid misdiagnosis over the staging of diseases. To prevent false positives and false negatives this can have adverse effect on the patient's both mental and physical health. Many automatic detecting and diagnosing techniques have evolved for tumors in human body parts such as, bones, brain, lung, kidney, pancreas, liver, breast, prostate, and others. And the Computer-aided diagnosis (CAD) includes pre-processing and feature extraction, selection, segmentation of nodules followed by a classification step. In cases where a large number of named training examples are available however, end-to-end training can be used in deep learning approaches.

II. RELATED WORK

The lung nodules categorization plays vital role for early stage diagnosis of the cancer that helps the patients to increase the mortality rate.

According to proposed paper [3] it shows that by using the lung nodules with 3D surfaces analysis is done using Spherical harmonics and how they have utilized the K-nearest neighbour algorithm to differentiate between the benign and malignant tumors. The other approach [4] is by two dimensional (2D) texture features analysis which is done with



help of Haralick, Gabor and local binary patterns on the tumors for verifying malignant and benign nodules using database that publicly available, where Haralick gives the highest feature method.

In paper [5] for the nodule grouping based on the CT scans available for developing a CAD system. The segmentation of an image was done using 3D active contours. One can get to know that by [6] that both using both Genetic Algorithm (GA) and Random subspace method (RSM) to increase the efficiency of the classifier system. In order classify the lung nodules and to obtain the deep features present in scans a CAD system is The nodules shape models was developed and used the variable with Spherical Harmonics and their surfacing characteristics were obtained by applying deep CNN and for malignancy feature random forest classifier is used [8].

For the purpose of analysing the border attributes of nodules in the CT scans with high resolution to enhance the efficiency image detection [11]. By using a multi-task regression (MTR) method which supports diverse computational characteristics obtained by deep learning models of Stacked Denoising Auto-Encoder (SDAE) and Convolutional neural network (CNN), Haar-like features to detain eight interpreting features of lung CT nodules. The data is obtained by Lung Imaging Database Consortium (LIDC) for the vast interpretations; there the nodules are quantitatively estimated to the semantic attributes by numerous radiologists [13].

The method to use the CT images from the Lung image database consortium-Image database resource initiative (LIDC-IDRI) for lung nodule screening and annotations are given using which a multiscale CNN to record tumors diversity by obtaining its features [14]. For lung cancer the radiologists consider the CT scans then come up with the following check-up plans based on nodule size and type detected. So in this [15] raw CT scans which as nodules without extra information of size or segmentation will be analysed using 3D and 2D views. Then the deep learning trained model is used for screening trail independent of datasets available.

The CAD system proposed [16] where the trained data as previously learnt by the discerning attributes for the pulmonary nodules which use the multi view Convolutional networks. The network contains the nodule data by three detectors that are designed for large, solid and medium solid nodules. To every set of 2D patches on different surface planes and then outputs are merged with fusion to obtain final grouping.

A. Lung Diseases

Lung cancer starts in lungs and then spreads across lymphatic nodes or to further parts of body. Same way cancer in other body part can also spread across lungs. The cancer tissue that spreads from one part of body to another is known metastases. The different types of lung disease X-ray images used in project are Hernia, Pneumonia, Pleural thickening, Edema, Emphysema, Cardiomegaly, Infiltration, Consolidation, Pneumothorax, Fibrosis, Mass, Nodule, Atelectasis, Effusion.

B. Malignant and Benign

Usually the benign nodules don't affect the normal function of body; a big concern of benign tumors is separating it from malignant nodules which are cancerous and harmful. The mortality rate of the lung cancer patient is more if it is detected and treatment is done in the early days. Attributes of different types of tumors are Proportions, Growth rate, Repetitive, Interfere, Risk of life, Position, Age factor.

III. PROBLEM STATEMENT AND NEED FOR THE PROJECT

Based on the labelled lung disease data and X-ray image data available on Kaggle publicly is a large dataset that is used in the implementation of the project.

For the here first the review and survey of the data set is done, later the Machine Learning and Deep Learning techniques will be applied in order see whether the patient is suffering from any lung disease or not, and so what type of lung disease is present. In this project primary categorization of input is done on patient's data such as their age, gender, X-ray images, View Position and output is know whether the diseases is found. The difficulty is it's a new dataset, and the analysis is about a large dataset that has not filtered fully, here the data contains lots of noise, and using only X-ray images is not enough to determine the patient's illness.

In the project it uses both machine learning and deep learning techniques to analyse the data and to create a model for the diagnosis patients. The important aspect here is the combination of processed patients information along with data from X-rays is needed, and then by using CNN with a pre-trained model, first time using the LocNet network for data this form.

As per previous analysis there is high mortality rate of lung cancer patients, so the key to increase their survival rate lies in early diagnosis. Early detection is necessary because abnormal tissue which is spotted initially is easier to treat otherwise it may spread to other parts of body and lead to difficulty in treatment. Now in the era where data is largely available it will be easier to detect the diseases based on different algorithms present using machine learning and deep



learning techniques. Thus with the power of technology as well as the large amount of data being available to the public, this is a good time to solving this problem.

- The objective here is to classify the lung diseases with proper labels using optimized CNN.
- And also to improve the efficiency of the classifying model.

A. Datasets and Inputs

The project uses a large dataset that is public on Kaggle, because the dataset is so large; first a test will be conducted on the sample data set.

- For the project in first case only a random sample of 5% is considered having 5,606 lung images along with data about the patients like age, gender, patient data, X-ray of lung diseases such as Hernia, Edema, Pneumonia, Fibrosis, Emphysema, etc.
- In the second case total of around 112,120 lung images with class labels and patient data of 14 diseases is taken.

B. Data Analysis and Data Processing

Data processing such as standardized age in digital form per year, age noise filtering, and one hot attributes such as gender, snapshot of the image as well as the specific type of illness that the patient suffers.

In the continuation the data analysis uses same steps then the values of data are sorted out and described based on the available information. Then the factor plotting and subplotting of available data is done as shown in Figure 1. Analysis of data such as age, gender as shown in Figure 1 and photo-taking will affect the likelihood that a patient will develop a specific disease; image processing resizing of images to its size, and converting them into black and white or the colour images with parallel processing. After pre-processing is done three image files for train, valid and test data is formed in both gray scale and RGB which is used in further processing local network for convoluting.

Then the data is divided into training set - validation set - testing set at different rates: in the sample dataset will divide proportionally 60% training set – 20% validation set – 20% testing set because this dataset is small with 5,606 samples, and full dataset have proportionally 80% training set – 10% validation set– 10% testing set because this is huge 112,120 samples.

C. Modelling with CNN

Test the CNN with the original architecture to make sure CNN is catching the pattern of the diseases, then accelerate the convergence by using a pre-trained model. Optimize CNN with momentum and the spatial transform. From testing multiple values using CNN with a variety of diseases and models to identify patients with the disease. For the given dataset sample some method is used to increase the data. After that the best flow to train on the full dataset is chose. Modelling with LocNet and comparing it to Optimized CNN. Following the same steps for the CNN section just instead of applying CNN here it uses the LocNet network.

D. Unsupervised Learning

Unsupervised Learning is a type in ML where the models do not require the user's supervision. Rather it gives model a work to identity its own hidden patterns and information that was not detected before. This type of models usually deals with unnamed data values.

This technique is used to perform complex processing tasks compared to supervised learning. Unsupervised learning can be uncertain when compared with other natural learning methods. Unsupervised algorithm is used in a model to train it with unlabeled data values that do not have any supervision. In the all the lung X-ray images are trained then the output is a certain lung disease among the 14 types of images used.

According to paper [24] they are using the unlabeled audio speech data to which the unsupervised algorithm is applied to categorize the audio based on language directly into word units by a Bayesian model the input data is transcribed. Error rates are recognized based on the mapping the output that is unsupervised with true transcription available.



Figure.1 Data analysis plots based on Age and Count for different lung diseases

Multimodal medical images provide much information for tissues diagnosis which can help in automated radiotherapy planning, automated diagnosis and medical image retrieval. [26] By the usage of unlabeled multimodal DCE-MRI datasets the organ identification in MRI images is done by deep learning methods which by visual and hierarchical features categorisation of object classes. In this project the unsupervised method used is clustering of the X-ray images along with extra patient data available to differentiate malignant and benign tumors in lungs.

E. Supervised Learning

Supervised learning in creating artificial intelligence (AI) models, where a computer algorithm will be trained based on input dataset which is labelled to get the suitable labelled output. The model is trained up to a time where it can detect the primary patterns and relationships between the input data and the output labels, allowing it to give accurate labelling results when presented with never-before-seen data. Supervised learning is good at classification and regression problems, in the project it used to classify different chest X-ray and CT based on the labelling and to produce disease output predicted which is also labelled. In this project it is used to classify the lung X-rays into benign or malignant category.

F. Optimized CNN

Typically after survey of image size is done it is seen that 64*64 pixel image is both small and good for model to maintain its pattern of an image. By using the spatial transformer with simple localization network model to get the key feature from an image where some of the layers are supported in front as lamda layer. So to overcome this problem in the project optimized CNN based image recognition model is used. Optimizing is a method used to change the attributes of a neural network that can be either weights or learning rate to decrease the losses in the system. Critical data values are tested in the architecture models in many places where always classification is a first step. Then by adjusting the attributes as follows for optimizing the system they are threshold value, precision, recall, and Fbeta score. Later by refining the index of the dropout layer in the classification is done.

IV. BLOCK DIAGRAM OF THE ARCHITECTURE

For the supervised learning method classification is necessary, here by the optimization of CNN is done by enhancing CNN features. The block diagrams contain 5 layers as shown in Figure 2. Whereas the Programming flowchart for predicting the Lung disease in Figure 3.

- In Lambda input layer the transfer of routing feature is done that will have an average value of image as zero.
- Batch normalization of the lung images is done to reduce the training time at the input layer of network.
- Spatial transformation is used here for lung image to change the co-ordinate system as necessary.
- VGG16 is used for creating a model by training it with lung images and data's. Then the pre-trained VGG which 13 layers of convolution network is used for further processing with pixel varying.
- At last is Dropout layer which as an input pool of convoluted images. It has two parts flatten with extra inputs dropouts, dense Relu dropout.
- Finally output will be obtained based on the lung image input given.

First step for any raw dataset is to pre-process the data values for any missing, null values and sometimes noise is also present which is also removed in this step. Then Keras pre-processing is applied for data pre-processing and data augmentation of modules which is used in deep learning it can be used in all kinds of datasets available whether image data, text data or sequence data. Where by improving the precision value there is a decrease in the recall value; the precision value is always obtained by true positives divided by true positive and false positive; whereas the recall value is obtained by true positive positives divided by true positive and false positive; whereas the recall value is obtained by true positive divided by true positive and false negative. After the Localized network is created by training the models obtained data. Then the input is given to it in the form of X-ray image to which the output whether the disease is present or not, which disease it is or is it benign or malignant type will be known.

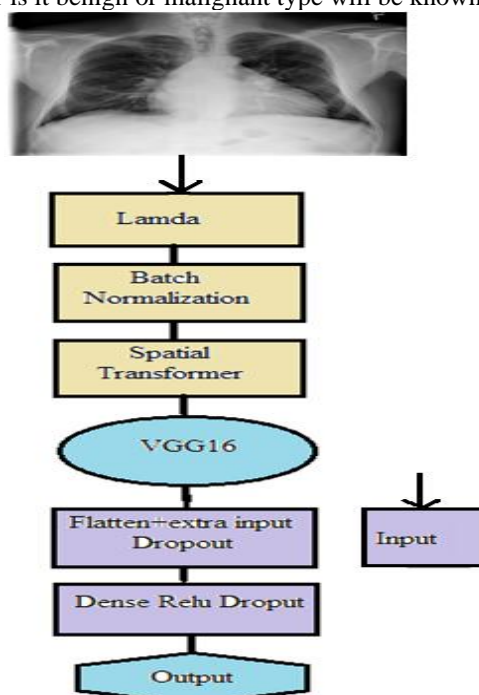


Figure.2 Block diagram of Architecture

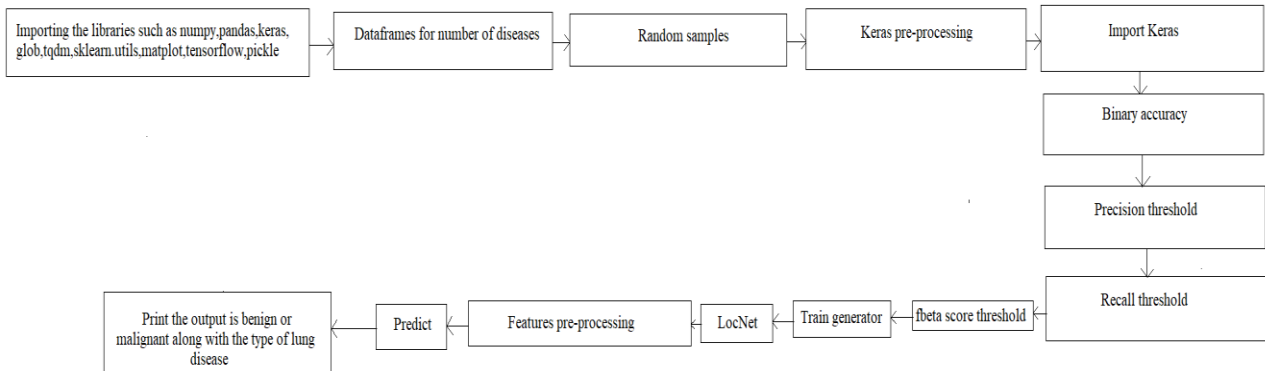


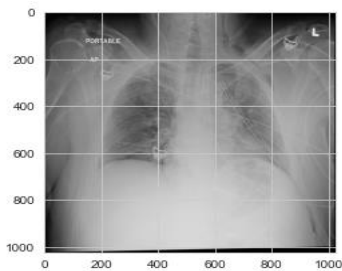
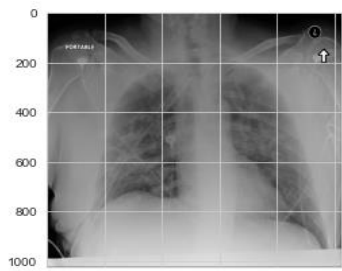
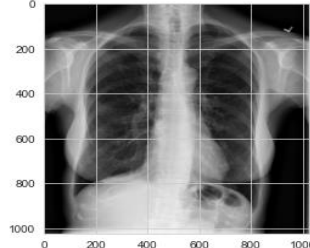
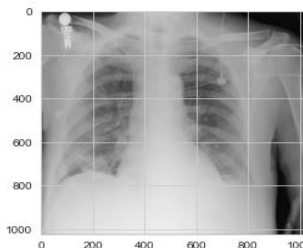
Figure.3 Flowchart for predicting the lung disease using deep learning using Python

For unsupervised algorithm all the lung X-ray images are trained using VGG16 and optimized CNN for convolution. Then when a set of image input is given the supervised output is obtained for two category one for benign and other for malignant. Based on the confidence value obtained after epoch =5 if the value is greater than 0.5 then the output predicted is near to true value. More the confident value the output can be considered true value for the given input value along with it unsupervised algorithm is used to find the lung disease based on the trained image files.

A. Result

Here the number of input X-ray image is given based on which the output is obtained has shown in Table I.

TABLE I . Output for the input images given

 <p>Unsupervised Output: Edema, confident: 0.9538707 The given Image for Supervised Output is Malignant</p>	 <p>Unsupervised Output: No finding, confident: 0.92624587 The given Image for Supervised Output is Benign.</p>
 <p>Unsupervised Output: Pneumothorax, confident: 0.9491018 The given Image for Supervised Output is Malignant</p>	 <p>Unsupervised Output: Mass, confident: 0.9682551 The given Image for Supervised Output is Benign</p>

	
<p>Unsupervised Output: Nodule, confident: 0.96917266 The given Image for Supervised Output is Benign</p>	<p>Unsupervised Output: Effusion, confident: 0.98763937 The given Image for Supervised Output is Malignant</p>

V. CONCLUSION

As science and technology is getting developed time and again new discoveries and inventions are made each day. In that case machine learning and Deep learning plays a vital role in the future era because the amount data available will be lot more which is used in analysis of more disease studies of medical images. Here only 2D images are used to improve system 3D images can be used for overall viewing of a position and for more accuracy. This lung disease predicted can be further studied in future for more accuracy to help doctors make analysis. Same programming can be used with certain changes and training for other organs of body like in case of pancreatic cancer. Here the confident values for epoch=1 was in the range of 0.23-0.4 but when training is done for epoch =5 the confident value rises from 0.83-0.96 so for all the medical image processing using deep learning need lot of training to get the high accuracy to get the prediction output which is near to true value.

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 Approaches (2018).
- [2] Sadot, E., Basturk, O., Klimstra, D.S., Gonen, M., Anna, L., Do, R.K.G., DAngelica, M.I., DeMatteo, R.P., Kingham, T.P., Jarnagin, W.R., et al.: Tumor-associated neutrophils and malignant progression in intraductal papillary mucinous neoplasms: an opportunity for identification of highrisk disease. *Annals of surgery*262(6) ,1102 (2015).
- [3] El-Baz, A., Nitzken, M., Khalifa, F., Elnakib, A., Gimelfarb, G., Falk,R., El-Ghar, M.A.: 3D shape analysis for early diagnosis of malignant lung nodules. In: IPMI. pp. 772–783. Springer(2011).
- [4] Han, F., Wang, H., Zhang, G., Han, H., Song, B., Li, L., Moore, W.,Lu, H., Zhao, H., Liang, Z.: Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *Journal of Digital Imaging* 28(1), 99–115 (2015).
- [5] Way, T.W., Hadjiiski, L.M., Sahiner, B., Chan, H.P., Cascade, P.N.,Kazerooni, E.A., Bogot, N., Zhou, C.: Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. *Medical Physics* 33(7), 2323–2337 (2006).
- [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] Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classification using deep features in CT images. In: *Computer and Robot Vision (CRV)*,2015 12th Conference on. pp. 133–138. IEEE (2015).
- [8] Buty, M., Xu, Z., Gao, M., Bagci, U., Wu, A., Mollura, D.J.: Characterization of Lung Nodule Malignancy Using Hybrid Shape and Appearance Features. In: *MICCAI*. pp. 662–670. Springer(2016).
- [9] Saouli, R., Akil, M., Kachouri, R., et al.: Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images. *Computer methods and programs in biomedicine* 166, 39–49(2018).
- [10] Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk Stratification of Lung Nodules Using 3D CNN-Based Multi-task Learning. In: *International Conference on Information Processing in Medical Imaging*. pp. 249–260. Springer (2017).
- [11] Furuya, K., Murayama, S., Soeda, H., Murakami, J., Ichinose, Y.,Yauuchi, H., Katsuda, Y., Koga, M., Masuda, K.: New classification of small pulmonary nodules by margin characteristics on high resolution CT. *Acta Radiologica* 40(5), 496–504 (1999).
- [12] Uchiyama, Y., Katsuragawa, S., Abe, H., Shiraishi, J., Li, F., Li, Q.,Zhang, C.T., Suzuki, K., Doi, K.: Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography. *Medical Physics* 30(9), 2440–2454 (2003).
- [13] Chen, S., Ni, D., Qin, J., Lei, B., Wang, T., Cheng, J.Z.: Bridging computational features toward multiple semantic features with multitask regression: A study of CT pulmonary nodules. In: *MICCAI*. pp.53–60. Springer (2016).
- [14] Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale Convolutional neural networks for lung nodule classification. In: *IPMI*. pp.588–599. Springer (2015).
- [15] Ciompi, F., Chung, K., Van Riel, S.J., Setio, A.A.A., Gerke, P.K., Jacobs, C., Scholten, E.T., Schaefer-Prokop, C., Wille, M.M., Marchiano, A., et al.: Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific reports* 7, 46479 (2017).
- [16] Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S.J., Wille, M.M.W., Naqibullah, M., Sanchez, C.I., van Ginneken, B.: Pulmonary nodule detection in CT images: false positive reduction using multi-view Convolutional networks. *IEEE TMI* 35(5), 1160–1169 (2016).



- [17] Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis* 42, 1–13 (2017).
- [18] Khosravan, N., Bagci, U.: S4ND: Single-Shot Single-Scale Lung Nodule Detection. *arXiv preprint arXiv:1805.02279* (2018).
- [19] Zhou, Y., Xie, L., Fishman, E.K., Yuille, A.L.: Deep Supervision for Pancreatic Cyst Segmentation in Abdominal CT Scans. *arXiv preprint arXiv:1706.07346* (2017).
- [20] Society, A.C.: *Cancer Facts & Figures*. American Cancer Society (2016).
- [21] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep Convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*(2015)
- [22] Gazit, L., Chakraborty, J., Attiyeh, M., Langdon-Embry, L., Allen, P.J., Do, R.K., Simpson, A.L.: Quantification of CT Images for the Classification of High-and Low-Risk Pancreatic Cysts. In: *SPIE Medical Imaging*. pp. 101340X–101340X. International Society for Optics and Photonics (2017).
- [23] Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: *ICCV*. vol. 1, pp.370–377. IEEE (2005).
- [24] Kamper, H., Jansen, A., Goldwater, S.: Fully unsupervised small vocabulary speech recognition using a segmental Bayesian model. In: *Interspeech* (2015).
- [25] Lee, H., Pham, P., Largman, Y., Ng, A.Y.: Unsupervised feature learning for audio classification using Convolutional deep belief networks. In: *Advances in neural information processing systems*. pp. 1096–1104(2009).
- [26] Shin, H.C., Orton, M.R., Collins, D.J., Doran, S.J., Leach, M.O.: Stacked auto encoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE transactions on pattern analysis and machine intelligence* 35(8),1930–1943 (2013).
- [27] Vaidhya, K., Thirunavukkarasu, S., Alex, V., Krishnamurthi, G.: Multimodal brain tumor segmentation using stacked denoising auto encoders. In: *International Workshop on Brain lesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. pp. 181–194. Springer (2015).
- [28] Sivakumar, S., Chandrasekar, C.: Lung nodule segmentation through unsupervised clustering models. *Procedia engineering* 38, 3064–3073 (2012).
- [29] 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).