Lung Cancer Detection using Machine Learning

Vaishnavi. D[1], Arya. K. S[2], Devi Abirami. T[3], M. N. Kavitha[4]

[1, 2, 3] B.E-CSE, Builders Engineering College, Kangayam, Tirupur, Tamil Nadu, India.

[4] Department of CSE, Builders Engineering College, Kangayam, Tirupur, Tamil Nadu, India.

Abstract -Automatic defects detection in CT images is very important in many diagnostic and therapeutic applications. Because of high quantity data in CT images and blurred boundaries, tumor segmentation and classification is very hard. This work has introduced one automatic lung cancer detection method to increase the accuracy and yield and decrease the diagnosis time. The goal is classifying the tissues to three classes of normal, benign and malignant. In MR images, the amount of data is too much for manual interpretation and analysis. During past few years, lung cancer detection in CT has become an emergent research area in the field of medical imaging system. Accurate detection of size and location of lung cancer plays a vital role in the diagnosis of lung cancer. The diagnosis method consists of four stages, pre-processing of CT images, feature, extraction, and classification, the features are extracted based on DTCWT and PNN. In the last stage, PNN employed to classify the Normal and abnormal.

Index terms – lung cancer, Dual-Tree Complex wavelet transformation, probabilistic neural network.

I. INTRODUCTON

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Lung cancer was the most common cancer in worldwide, contributing 2,093,876 of the total number of new cases diagnosed in 2018.

The incidence rate has been declining since the mid-1980s in men, but only since the mid-2000s in women, because of gender differences in historical patterns of smoking uptake and cessation.

From 2005 to 2015, lung cancer incidence rates decreased by 2.5% per year in men and 1.2% per year in women. Symptoms do not usually occur until the cancer is advanced, and may include persistent cough, sputum streaked with blood, chest pain, voice change, worsening shortness of breath, and recurrent pneumonia or bronchitis.

Cigarette smoking is by far the most important risk factor for lung cancer; 80% of lung cancer deaths in the US are still caused by smoking. Risk increases with both quantity and duration of smoking. Cigar and pipe smoking also increase risk. Exposure to radon gas released from soil and building materials is thought to be the second-leading cause of lung cancer in the US.

Other risk factors include occupational or environmental exposure to secondhand smoke, asbestos (particularly among smokers), certain metals (chromium, cadmium, arsenic), some organic chemicals, radiation, air pollution, and diesel exhaust. Some specific occupational exposures that increase risk include rubber manufacturing, paving, roofing, painting, and chimney sweeping. Risk is also probably increased among people with a history of tuberculosis. Genetic susceptibility (e.g., family history) plays a role in the development of lung cancer, especially in those who develop the disease at a young age.

We can cure lung cancer ,only if you identifying the yearly stage. So here, we use machine learning algorithms to detect the lung cancer. This can be made faster and more accurate. In this study we propose machine learning strategies to improve cancer characterization. Inspired by learning from CNN approaches, we propose new algorithm, proportion-PNN, to characterize cancer types.

II. RELATED WORKS

The work published by Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, Mathild**e [**1**]**explained about the proposed architecture which comprises multiple streams of 2-D ConvNets, for which the outputs are combined using a dedicated fusion method to get the final classification but the morphological variation of nodules is often greater than what a single candidate detection algorithm.

In this paper “Unsupervised Deep Embedding for Clustering Analysis”[2] it explains about clustering

.Clustering is central to many data-driven application domains and has been studied extensively in terms of distance functions and grouping algorithms. Relatively little work has focused learning representations for clustering. However misclassification of any images doesn't gives the approximate result.

The work published by Mario Buty1, Ziyue Xu1 , Mingchen Gao[3] it tells about Computed tomography imaging. It is a standard modality for detecting and assessing lung cancer. In order to evaluate the malignancy of lung nodules, clinical practice often involves expert qualitative ratings on several criteria describing a nodule's appearance and shape, but these features are mostly subjective and arbitrarily-defined.

The model proposed by Alan L. Yuille4 [4] it explains if it is done under a reasonable transformation function, our approach can be factorized into two stages, and each stage can be efficiently optimized via gradient back-propagation throughout the deep networks. We collect a new dataset with 131 pathological samples, which, to the best of our knowledge, is the largest set for pancreatic cyst segmentation. Without human assistance, our approach reports a 63:44% average accuracy measured by the Dice-Sorensen Coefficient (DSC), which is higher than the number (60:46%) without deep supervision. But in this process it gives less accuracy.

This work has introduced automatic lung cancer detection method to increase the accuracy and yield and decrease the diagnosis time .In MR images, the amount of data is too much for manual interpretation and analysis. The diagnosis method consists of four stages. In this Probabilistic Neural Network employed to classify the Normal and abnormal.

III. EXISTING SYSTEM

CNN is a class of deep neural network ,but it is done only with the collection of data and it is not labeled. It is most commonly applied to analyze visual imagery. CNN use relatively little pre-processing compared to other image classification algorithm. But ,it is difficult to get accurate results . Not applicable for multiple images for Lung detection in a short time .CNN uses relatively little pre-processing compared to other image classification algorithm.

IV. PROPOSED SYSTEM

This summarizes the advances in machine learning applied to PNN for the development of lung cancer diagnosis.CNN is a class of deep neural network, but it is done only with the collection of data and it is not labelled. It is most commonly applied to analyze visual images.CNN use relatively little pre-processing compared to other images classification algorithm. since the CNN algorithm takes lot many images as a data to calculate. It creates lagging of time and doesn't gives accuracy compared to PNN.

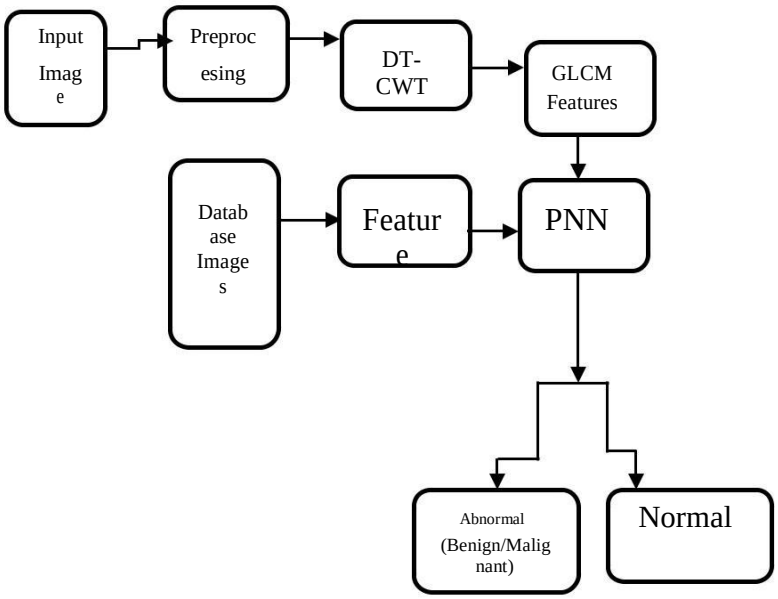

Fig 1 System flow diagram

A.PREPROCESSING:

Most appropriate input data has been selected, it must be pre-processed otherwise, the neural network will not produce accurate result. This reduces the number of inputs to the network and helps it learn more easily. It removes unwanted signals from the CT

images. It converts color images to grey-level coding.

B.DUAL TREE COMPLEX WAVELET TRANSFORMATION

The Dual-tree complex wavelet transform (DTCWT) calculates the complex transform of a signal using two separate DWT decompositions (tree *a* and tree *b*). If the filters used in one are specifically designed different from those in the other it is possible for one DWT to produce the real coefficients and the other the imaginary.

In n numerical analysis and functional analysis, a Dual-Tree Complex wavelet transformation is any wavelet transform for which the wavelet are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transform is temporal resolution. It captures both frequency *and* location information (location in time). The Haar wavelet transform may be considered to pair up input values, storing the difference and passing the sum.

This process is repeated recursively, pairing up the sums to prove the next scale, which leads to differences and a final sum.

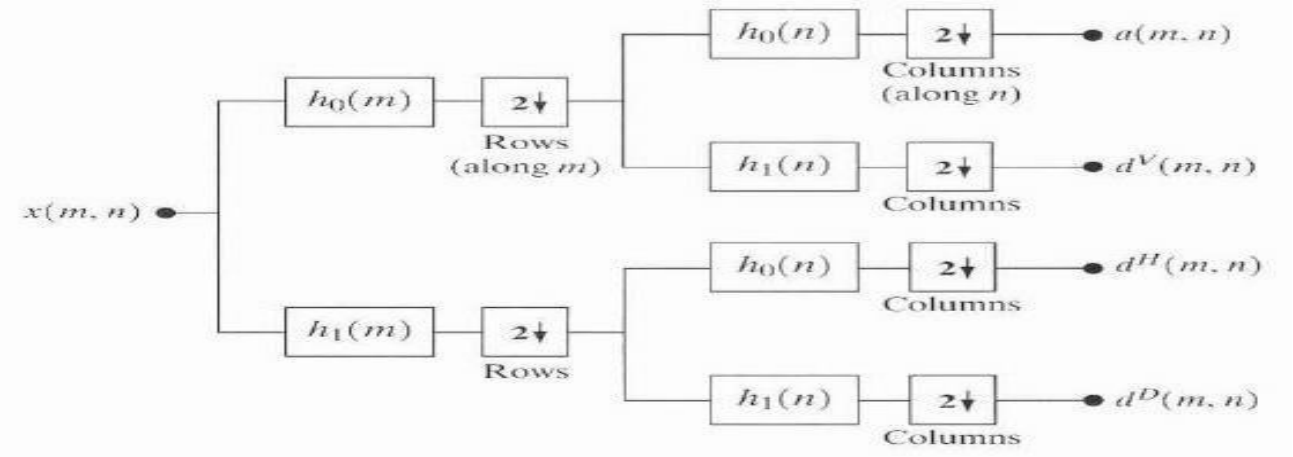

Fig 2Diagrammatic representation of dual tree complex wavelet transformation

C.GRAY LEVEL CO-OCCURRENCE MATRIX

It referred as co-occurrence distribution. It is the most classical second-order statistical method for texture analysis. An image is composed of pixels each with an intensity (a specific gray level), the GLCM is a tabulation of how often different combinations of gray levels co-occur in an image or image section. Texture feature calculations use the contents of the GLCM to give a measure of the variation in intensity at the pixel of interest. GLCM texture feature operator produces a virtual variable which represents specified texture calculation on a single beam echogram.

Steps for virtual variable creation:

Quantize the image data: Each sample on the echogram is treated as a single image pixel and its value is the intensity of that pixel. These intensities are then further quantized into a specified number of discrete gray levels, known as Quantization.

Create the GLCM*:* It will be a square matrix N x N in size where N is the Number of levels specified under Quantization.

Steps for matrix creation are:

Let s be the sample under consideration for the calculation. Let W be the set of samples surrounding sample s which fall within a window centered upon sample s of the size specified under Window Size. Now define each element i, j of the GLCM of sample present in set W, as the number of times two samples of intensities i and j occur in specified Spatial relationship. The sum of all the elements i, j of the GLCM will be the total number of times the specified spatial relationship occurs in W. Make the GLCM symmetric, Make a transposed copy of the GLCM

and add this copy to the GLCM itself. This produces a symmetric matrix in which the relationship i to j is indistinguishable for the relationship j to i.

Due to summation of all the elements i, j of the GLCM will now be twice the total number of times the specified spatial relationship occurs in W. Normalize the GLCM, Divide each element by the sum of all elements and GLCM may now be considered probabilities of finding the relationship *i, j (or j, i)* in W. Calculate the selected Feature. This calculation uses only the values in the GLCM are energy, entropy, contrast, homogeneity, correlation. The sample s in the resulting virtual variable is replaced by the value of this calculated feature.

GLCM directions of Analysis are Horizontal (0^0) **,**Vertical (90^0)**,**Diagonal. In diagonal it has *Bottom left to top right (-45 ^{0}),Top left to bottom right (-135 ^{0})***.**Denoted P_0, P_{45}, P_{90}, & P_{135} Respectively.

Ex. P_0 (i, j)

GLCM of an image is computed using a displacement vector d, defined by its radius δ and orientation θ.

Consider a 4x4 image represented by figure 1a with four gray-tone values 0 through 3. A generalized GLCM for that image is shown in figure 1b where #(i,j) stands for number of times i and j have been neighbors satisfying the condition stated by displacement vector d.

ENERGY

It is also called Uniformity or Angular second moment. Measure the textural uniformity that is pixel pair repetitions. Detects disorders in texture. Energy reaches a maximum value equal to one.

ENERGY=$\sum_i\sum_j P_{ij}^2$

ENTROPY

Measure the disorder or complexity of an image. The entropy is large when the image is not texturally uniform. Complex textures tend to have high entropy. Entropy is strongly but inversely correlated to energy.

ENTROPY (ent)=$-\sum_i\sum_j P_{ij}\log_2 P_{ij}$

D.PROBABILITY NEURAL NETWORK

Performance of the PNN classifier was evaluated in terms of training performance and classification accuracies. This network is a kind of radial basis network and It gives fast and accurate classification and is a promising tool for classification of the defects from quality material. Existing weights will never be alternated but only new vectors are inserted into weight matrices when training. So it can be used in real-time. Since the training and running procedure can be implemented by matrix manipulation, the speed of PNN is very fast.

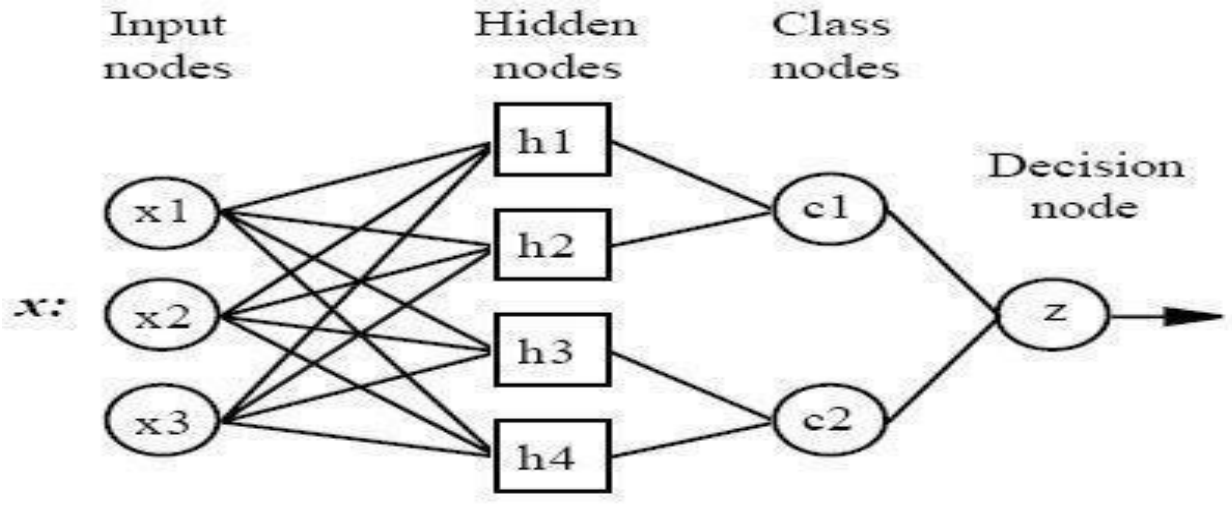

Fig 3 Architecture of probabilistic neural network

LAYERS

PNN is often used in classification problems. When an input is present, the first layer computes the distance from the input vector to the training input vectors. This produces a vector where its elements indicate how close the input is to the training input. The second layer sums the contribution for each class of inputs and produces its net output as a vector of probabilities. Finally, a complete transfer function on the output of the second layer picks the maximum of these probabilities, and produces a 1 (positive identification) for that class and a 0 (negative identification) for non-targeted classes.

1. INPUT LAYER

Each neuron in the input layer represents a predictor variable. In categorical variables, N-1 neurons are used when there are N number of categories. It standardizes the range of the values by subtracting the median and dividing by the inter quartile range.Then the input neurons feed the values to each of the neurons in the hidden layer

2. PATTERN LAYER

This layer contains one neuron for each case in the training data set. It stores the values of the predictor variables for the case along with the target value. A hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the radial basis function kernel function using the sigma values.

3. SUMMATION LAYER

For PNN networks there is one pattern neuron for each category of the target variable. The actual target category of each training case is stored with each hidden neuron; the weighted value coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron's category. The pattern neurons add the values for the class they represent.

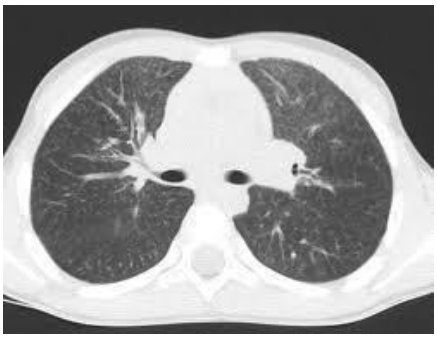

Fig 4 Preprocessing

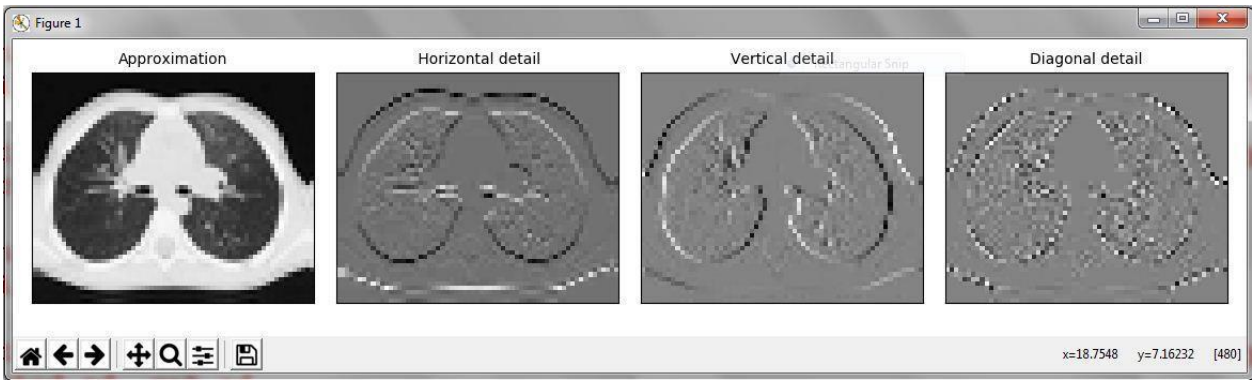

Fig 5 Dual-tree complex wavelet transformation

V. CONCLUSION AND REMARKS

We have presented a PNN algorithm for cancer detection in early stages based on neural network. This gives the good result of accuracy and low computation time make the PNN algorithm highly suited to make decision for screening the lung cancer. Instead of using images, video can be used for the clarity.

REFERENCES

[1] Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, 0DWKL´3XOPRQDU\ QRGXOH GHWHFWLRQ LQ &7 LPDJHV false positive reduction using multi-view FRQYROXWLRQDO QHWZRUNV´

[2] JunyuaQ ;LH 5RVV *LUVKLFN ³8QVXSHUYLVHG 'HHS (PEHGGLQJ IRU &OXVWHULQJ $QDO\VLV´

[3] 0DULR %XW\ =L\XH ;X 0LQJFKHQ *DR´ Characterization of Lung Nodule Malignancy using +\EULG 6KDSH DQG $SSHDUDQFH)HDWXUHV´

[4] $ODQ / <XLOOH ... sion for Pancreatic ... &\VW 6HJPHQWDWLRQ LQ $EGRPLQDO &7 6FDQV

[5] 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)

[6] 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)

[7] 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)

[8] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV. pp. 4489-4497. IEEE (2015)

[9] Lee, C., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-Supervised Nets. International Conference on Artificial Intelligence and Statistics (2015)

[10] Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. Computer Vision and Pattern Recognition (2015)

[11] Milletari, F., Navab, N., Ahmadi, S.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. arXiv preprint arXiv:1606.04797 (2016)

[12] Roth, H., Lu, L., Farag, A., Sohn, A., Summers, R.: Spatial Aggregation of Holistically-Nested Networks for Automated Pancreas Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (2016).

[13] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. &LRPSL ³2ff-the-shelf convolutional neural network features for pulmonary nodule detection in FRPSXWHG WRPRJUDSK\ VFDQV´ LQ , Symposium on Biomedical Imaging, 2015, pp. 286-289.

[14] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. V. Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batrah, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallamm, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, and B. < &URIW ³7KH /XQJ Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung QRGXOHV RQ &7 VFDQV´ 0HGLFDO 3K\ 915-931, 2011.

[15] D. P. Naidich, A. A. Bankier, H. MacMahon, C. M. Schaefer-Prokop, M Pistolesi, J. M. Goo, P. Macchiarini, J. D. Crapo, C. J. Herold, - + $XVWLQ DQG : ' 7UDYLV ³5HF for the management of subsolid pulmonary nodules detected at CT: a statement from the IOHLVFKQHU VRFLHW\´ 5DGLRORJ\ 317, 2013.

[16] D. Manos, J. M. Seely, J. Taylor, J. Borgaonkar, + & 5REHUWV DQG - 5 0D\R ³7KH and data system (LU-RADS): a proposal for FRPSXWHG WRPRJUDSK\ VFUHHQLQJ´ &D Association of Radiologists Journal, vol. 65, pp. 121-134, 2014

[17] D. M. Xu, H. Gietema, H. de Koning, R. Vernhout, K. Nackaerts, M. Prokop, C. Weenink, J. Lammers, H. Groen, M. Oudkerk, and R. van .ODYHUHQ ³1RGXOH PDQDJHPHQW S 1(/621 UDQGRPL]HG OXQJ FDQFHU VF Lung Cancer, vol. 54, pp. 177-184, 2006.