Optimization of Machine Learning Model by Applying a Random Projection Algorithm for Breast Lesion Classification

Sonali Gupta,
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract
Malignant and benign lesion classification is a challenging process that requires the best possible fusion of several imaging parameters in relation to tissue density heterogeneity, prediction of lesion boundaries, and change of surrounding tissues. Recent research has shown that important picture properties such as intensity, energy, homogeneity, entropy, and statistical moments, among others, may be modelled using statistics and texture features. As a result, this method was created to make early predictions about the identification of breast lesions in digital image processing. Preprocessing, features extraction, and classification are among the phases that make up the project. The main goal of the suggested technique is to categorise lesions into benign and malignant categories. Additionally to enhance feature categorization.

DOI Number: 10.48047/nq.2021.19.6.NQ21087 **NeuroQuantology2021;19(6): 200-205**

1. INTRODUCTION

To evaluate a person's health and detect illnesses, a blood test known as a complete blood count (CBC) is performed including leukaemia, infection, also anaemia. Whole blood count is crucial in medical diagnosis. Red blood cells (RBCs), white blood cells (WBCs), platelets, and plasma are the four primary types of cells. These categories may be distinguished based on the nucleus and cytoplasm's size, colour, texture, and morphology. Determine the immunity and capacity of the bodily system by counting the number of cells. The existence of illness is indicated by an aberrant cell count, and the patient requires medical attention. Research is now being done on how to automatically count RBCs & WBCs from images of blood.

Leukocytes are another name for WBCs. These cells important function in the immunological system. They safeguard the body by getting rid of germs and viruses there. Leukopenia is the medical word used to denote low count. Leukopenia is a sign of an infection. Leukocytosis is the medical word used to denote a high count. Leukocytosis is a sign of infection, leukaemia, or tissue deterioration. Erythrocytes are another name for RBCs. RBCs transport oxygen to the body's cells and collect carbon dioxide from the lungs. Hemoglobin is a protein found in them. Blood's red hue is caused by the presence of protein in both the inner and outer layers. Oxygen is transported via haemoglobin.

Anemia is caused by an irregular RBC count and manifests as mental fatigue, sickness, weakness, and dizziness. If it is not treated right away, it might lead to more serious symptoms including leukaemia and starvation. RBC indices provide details on cell size and shape and are helpful in identifying different kinds of anaemia. Thrombocytes are another name for platelets. The platelets' job is to clump and clot blood vessel damage in order to halt bleeding. Thrombocytopenia is a low platelet count. It makes a person bleed and prevents blood from clotting. Thrombocytosis is a high platelet count. Blood clots inside blood vessels as a result, preventing normal blood flow. Consequently, platelet count has to be within normal range for optimal blood flow.

Blood is a bodily fluid that transports nutrients and oxygen while generating blood clots to stop excessive blood loss. Moreover, it is essential for controlling body temperature and defending the immune system. Plasma, red blood cells, white blood cells, along with platelets most of the blood is composed of. Leukocyte and erythrocyte counts

can be used to identify undetected medical disorders, concealed infections, and immune system health. Leukocytes, also known as white blood cells, are crucial components of the human immune system that defend the body from illnesses and external invaders. In a microliter of blood, there are 3,500 to 10,500 leukocytes, which barely make up 1% in healthy individuals.Hemoglobin, a unique protein found in red blood cells or erythrocytes, aids in the transportation of oxygen from the lungs to all living tissues also helps to expel carbon dioxide by breathing. Erythrocyte counts in healthy individuals typically vary from 4.7 to 6.1 million in males and 4.2 to 5.4 million in females per microliter of blood.

2. LITERATURE SURVEY

This research presents a concentrated literature review on current neural network advancements in medical picture segmentation including edge detection, visual content analysis, plus registration using computer-aided diagnosis. Generally speaking, supervised and unsupervised learning are the two learning techniques for neural networks used in medical image processing. A network is trained through a set of inputs and outputs (targets) in supervised learning with the aim of minimising the network's total output error over all training examples. In unsupervised learning, a function is developed that assesses the network's appropriateness or accuracy and the training dataset is devoid of any goal data. By automating numerous design decisions, Group Method of Data Handling (GMDH) neural networks may help users with design decisions.The process of combining many sources of data into a single coordinate system is called image registration. Neural network applications have been divided into four main groups. The registration of a patient's information to an anatomical atlas can also be done through non-rigid registration of medical pictures. It can be challenging to defend the usage of neural networks in circumstances when it is necessary to express the process converting inputs to output values simply and concisely. Neural networks are famously difficult to read and analyse [1].

Procedures in medical information known as computer-aided detection (CADe) and computer-aided diagnosis (CADx) help clinicians understand medical pictures. Segmentation algorithms use methods including thresholding, region expanding, deformable templates, as well as pattern recognition to act on the intensity or texture changes of an image. The most popular segmentation approaches include statistical, geometrical, structural, model-based, signal processing, spatial domain, Fourier domain, Gabor, and wavelet filters, as well as discriminating functions dependent upon Mahalanobis distance. A diagnostic system's evaluation, which is based on statistical decision theory, provides estimates of the likelihood that a certain course of action will be taken. The effectiveness of a CAD system is assessed using performance metrics such the true positive percentage, false-positive fraction, sensitivity, specificity, and Receiver Operator Characteristics (ROC).In order to aid radiologists in the identification of various illnesses on medical pictures, CAD generates a result that serves as a "second opinion." The diagnostic procedure was developed by combining CAD and applied mathematics. The benefits of CAD for aiding radiologists in lung cancer identification. Many prospective clinical trials conducted in recent years have shown the benefits of CAD for aiding radiologists in the diagnosis of lung cancer, and researchers are also working hard to expand the use of CAD. The segmentation of medical pictures is difficult, though, as a ground truth is sometimes absent [2].

The medical background history of TB disease in chest X-rays is covered in this work, along with a study of the many methods for TB detection and categorization. Conventional chest X-rays (CXRs) are a low-cost method of TB screening, but mass screening of a large population is a laborious and time-consuming task. In order to detect tuberculosis infections in chest X-rays, computer-aided diagnostic systems (CAD) rely on radiologists and have the ability to reduce the likelihood of TB detection errors. For medical imaging modalities such as X-ray, MRI, ultrasound, etc., image segmentation is a crucial step. Image segmentation is used in various picture preparation processes, including object recognition, object occlusion, boundary estimate, editing, or image database searches, image security, and image compression.This research examines the examination of a CAD system for the automated analysis of chest x-rays for pulmonary TB detection. It places emphasis on outlining the X-ray image processing procedures for TB classification, such as feature extraction and pre-processing [3].

This paper describes an automated method for quantifying the amount of patterns of diffuse parenchyma lung disease's interstitial pneumonia (IP) (DPLD). The lung field is divided into lung parenchyma (LP) volume using this method, which segments the data from a multi detector CT (MDCT) scan. A three-class pattern classification of LP into normal, ground glass, and reticular patterns is how IP patterns are identified and described. Using volume overlap, the suggested scheme's performance was assessed. Accurate vascular tree structure segmentation is needed to lower false-positive detection rates. K-NN

classification is one of the most straightforward supervised classification techniques in the field of statistical pattern recognition. A tablet having a 305 mm active area, 0.2 lp/mm resolution, with 0.25 mm accuracy was employed (Wacom Intuos3 Tokyo, Japan). To minimise the feature vector's dimensions, the stepwise discriminant analysis (SDA) statistical method was used (130). Between all three classes, a covariance matrix was created for each of the characteristics.After the feature extraction using the Co-occurrence matrix 3 for 2D pictures, the discussion turned to how a similar method might be used to create a co-occurrence matrix for volumetric data. It provides a thorough grasp of the extraction of co-occurrence features. The primary flaw in the traditional FCM method is that it treats each picture as a separate set of points since the fuzzy function does not account for spatial dependency [4].

In order to categorise lung pictures in a series of computed tomography (CT) images, this research suggests a locality-constrained linear coding (LLC) based method. Studies show that this method is superior to baseline classification approaches that use histogram and VQ similarity. Spatial pyramid matching and feature context are two feature representations that are evaluated in this study. According to experimental findings, the integrated techniques perform at the cutting edge on the lung imaging dataset. SPM and FC outperform a pure histogram in performance.SPM+LLC+KNN produces the greatest classification rate of 89.2%. Representative photos of 30 participants were taken, including 9 people with CLE, 6 people with PSE, 7 people with ILD, and 8 people who were healthy controls (NLs). Three representative slices of each subject's top, middle, and lower lung were used to pick textured tissue samples. 200 lung ROI pictures with a 1.5 mm and 5 mm spatial resolution made up the final datasets.The pathological severity of emphysema may be evaluated using it. We integrate texture characteristics with feature representations, which, in contrast to conventional approaches, captures spatial distribution [5].

3. PROPOSED SYSTEM

The majority of cancer fatalities are caused by inadequate first diagnosis, necessitating computer-aided diagnosis (CAD). In order to provide radiologists with an expert level preliminary diagnosis for cancer patients, digital pathology is a technology. Identifying cell areas from micro-slide pictures is a difficult stage in digital pathology, but it is essential for subsequent procedures like identifying tumour subtypes using a convolutional neural network.Just 27% of cells are utilised in the detection of malignant cells. Because there is a human picture segmentation method involved, the segmentation of the images will not be exact.

An established method for dimension reduction and feature extraction is linear discriminant analysis (LDA). It has been extensively employed in several high-dimensional data application fields, including face recognition and picture retrieval. Nevertheless, because an eigen-decomposition involving the scatter matrices is required, PCA+LDA has substantial time and space costs. This prospective, multicenter research assessed the capacity of fluorine-18 deoxyglucose positron emission tomography (FDG-PET) to distinguish between benign and malignant lung nodules. FDG-PET was used to assess 89 individuals with newly discovered indeterminate SPNs on chest radiography and CT.

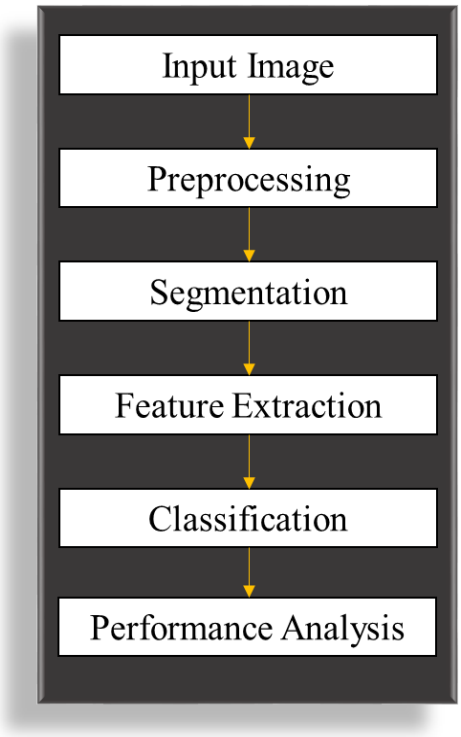

Fig 1: Block Diagram

PET data were assessed semi-quantitatively using a visual scoring approach in addition to standardised uptake values (SUVs) as a measure of FDG accumulation. PET exhibited an overall sensitivity as well as specificity of 92% and 90% for the identification of malignant nodules, compared to visual examination, which had a little greater sensitivity of 98% but not statistically significantly superior. The method for identifying Common Imaging Signs of Lesions (CISLs) in lung CT images is suggested in this research. In order to describe areas of interest (ROIs) in lung CT scans, it incorporates the bag-of-visual-words predicated upon this Histograms of Oriented Gradients (HOG)

as well as the Local Binary Pattern (LBP).The thickness of the cavity wall's spicular edge, a diameter of more than 3 cm, a benign development rate, as well as a benign pattern of calcification were the most crucial radiographic indicators of malignant nodules. The most crucial radiographic features for benign nodules were a benign growth rate and a benign pattern of calcification.

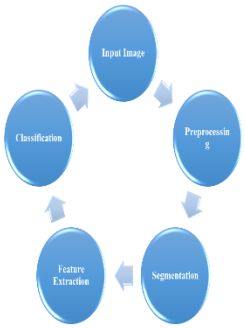

Fig 2: System Architecture

The category of the imaging sign present in each ROI was identified using the Max-Min posterior Pseudo-probabilities (MMP) learning approach. The suggested method yielded average results of 91.8% sensitivity, 98.5% specificity, and 98% accuracy. Each HOG or LBP as well as the combination of LBP with intensity histograms were outperformed by HOG-LBP features, and MMP performed better than Support Vector Machines (SVMs).Nevertheless, there are drawbacks, such as the manual picture segmentation procedure and the utilisation of just 27% of the cells to detect malignant cells.

In this process, we look at the viability of using a random projection approach to create an ideal feature vector from the enormous feature pool that was initially generated by CAD and enhance the performance of the machine learning model. We put together a retrospective dataset of 1,487 mammograms, 644 of which had verified malignant mass lesions, and 843 of which had benign lesions. To first segment mass areas and compute 181 initial features, a CAD approach is used. In order to forecast the risk that a lesion would be malignant, support vector machine paradigms integrated with a variety of feature dimensionality reduction techniques are created. The method is more trustworthy. By using clustering, the cells are automatically segmented. The approach is more effective since the feature categorization is stable.

The next section provides an explanation of the many phases that are involved in putting the suggested technique into practise:

1. Input image

A rectangular array of numbers makes up a picture (pixels). Each pixel is a measurement of a different aspect of a scene over a limited region. There are several ways to measure the characteristic, but often we either measure the average brightness (one number) or the brightest areas of the image after applying red, green, plus blue filters (three values). An eight bit integer is often used to represent the values, offering a brightness range of 256 levels. When we talk about an image's resolution, the terms "pixel count" and "brightness values" are used.

2. Pre-Processing

The technique of reducing noise from a signal is known as noise reduction. Analog and digital recording equipment alike have characteristics that make noise intrusion a possibility. Random or incoherent white noise, as well as coherent noise brought on by a device's mechanism or processing algorithms, are all types of noise. Hiss is a common type of noise in electronic recording equipment brought on by haphazard electrons that deviate from their intended route under the strong impact of heat. These errant electrons alter the output signal's voltage and produce audible noise as a result.

3. Segmentation

Image segmentation is a method for breaking a digital image up into separate pieces (sets of pixels, also known as super-pixels). Segmentation's goals include simplifying an image's representation and/or making it more pertinent and intelligible. Morphology is a broad category of image processing techniques that alters pictures according to their forms. Morphological processes may be applied to grayscale pictures as well, although they are less effective on absolute pixel values since they only have confidence in the related ordering of pixel values so instead of their numerical values, which makes them more focused on binary images.

4. Feature Extraction

Starting with a collection of measured data, the feature extraction method is used in machine learning, pattern recognition, and image processing to produce derived values (features) that are intended to be informative and non-redundant. This process speeds up the learning and generalisation processes and, in some cases, improves human interpretations. Dimensionality reduction and feature extraction are connected. In order to do the intended job employing this reduced representation rather of the whole starting data, it is assumed that the selected features will contain the pertinent information from the input data. In this procedure, the feature values from the picture are extracted utilising the two different types of feature extraction methods, including GLCM and PCA algorithms.

5. Classification

A method of classification that utilizes contextual information in images is known as contextual image classification, and it is a topic of pattern recognition in computer vision. "Contextual" suggests that this method focuses on the relationships between the neighbouring pixels, commonly known as the neighbourhood. This method's objective is to categorise the photos using the contextual data.

The challenge of categorising examples into one of the more than two classes in machine learning is known as multiclass or multinomial classification (Binary classification is the process of categorising occurrences into one of two classes). Most classification techniques, however are by inherently multinomial classifiers, others are binary algorithms that may be converted to multinomial classifiers using a number of techniques. Contrary to popular belief, multiclass classification does not require multiple labels to be predicted for every occurrence, as is the case with multi-label classification.

6. Performance Measures

Statistical measurements of the effectiveness of a binary classification test, commonly known as a classification function in statistics, include sensitivity and specificity: Sensitivity is the percentage of positives that are accurately classified as such (often referred to as the recall rate, the true positive rate, or the chance of detection in some sectors) (i.e. the percentage of sick people who are correctly identified as having condition). Specificity measures the percentage of negatives that are accurately classified as such (sometimes referred to as the real negative rate) (Specifically, the proportion of healthy individuals who are appropriately diagnosed as not having the illness.)

True positive: Ill individuals were recognised for what they were.

False positive: Healthy persons were mistakenly labelled as ill.

True negative: Healthy individuals were recognised as such.

False negative: Identification of sick individuals as healthy.

4. RESULTS

This method was created to detect breast lesions in digital image processing early on. Preprocessing, feature extraction, and classification are all involved. A retrospective dataset with 1,487 mammography instances is used to examine the viability of using a random projection approach to create an ideal feature vector from the vast feature pool that was initially created by the CAD. To first segment mass areas and compute 181 initial features, a CAD approach is used. To forecast the chance that lesions would be malignant, SVM models are created and incorporated with various feature dimensionality reduction techniques. Due to the feature classification's stability, the procedure has the advantages of dependability, clustering, and efficiency.

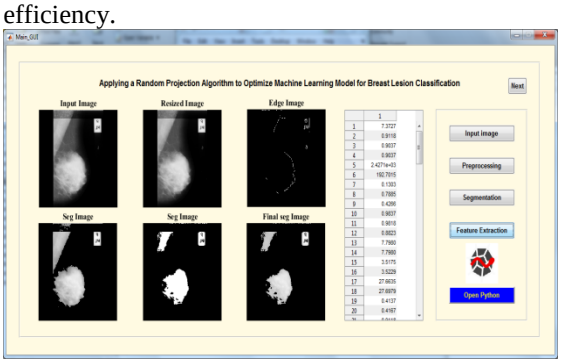

Fig 3: Final Segmented Image

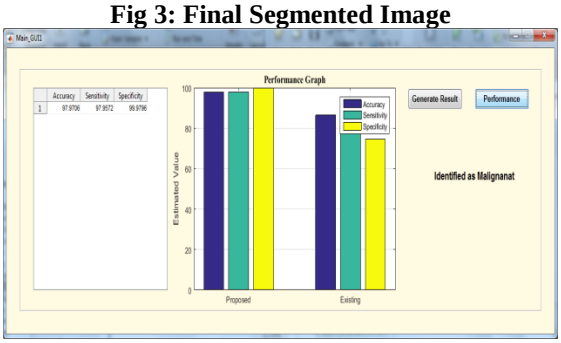

Fig 4: Performance Analysis

5. CONCLUSION

To identify breast lesions through digital image processing early, this method was created. Preprocessing, the extraction of characteristics, and categorization are involved. The viability of using a random projection approach to create an ideal feature vector from the previously created huge feature pool by the CAD is examined using a retrospective dataset of 1,487 instances of mammograms. In order to segment mass areas and first compute 181 features, a CAD approach is used first. The chance of lesions being malignant is predicted using SVM models incorporated with a variety of feature dimensionality reduction techniques. The technique has advantages in terms of clustering, dependability, and efficiency since the feature categorization is stable.

REFERENCE

[1] G. Battista, C. Sassi, M. Zompatori, D. Palmarini, and R. Canini, “Ground-glass opacity: Interpretation of high resolution CT findings,”LaRadiolo-gia Medica , vol. 106, pp. 425–442, 2003.

[2] Z. G. Yang, S. Song, and S. Talcashima, “High-resolution CT analysis of small lung adenocarcinoma revealed on screening helical CT,”

Amer. J. Roentgenol. , vol. 176, no. 6, pp. 1399–1407, 2001.
[3] T. Aoki, Y. Tomoda, H. Watanabe, H. Nakata, T. Kasai, H. Hashimoto, M. Kodate, T. Osaki, and K. Yasumoto, "Peripheral lung adenocarcinoma: Correlation of thin-section findings with histologic factors and survival," Radiology, vol. 220, pp. 803–809, 2001.
[4] J. J. T. Owen, D. E. McLoughlin, R. K. Suniara, and E. J. Jenkinson, "The role of mesenchyme in thymus development," Current Topics Microbiol. Immunol. , vol. 251, pp. 133–137, 2000.
[5] M. R. Melamed, B. J. Flehinger, M. B. Zaman, R. T. Heelan, W. A. Perchick, and N. Martini, "Screening for lung cancer: Results of the memorial sloan-kttering study in New York", Chest, vol. 86, no. 1, pp. 44–53, 1984.
[6] C. V. Zwirewich, S. Vedal, R. R. Miller, and N. L. M ¨uller, "Solitary pulmonary nodule: High-resolution CT and radiologic-pathologic correlation," Radiology, vol. 179, no. 2, pp, 469–476, 1991.
[7] S. F. Huang, R. F. Chang, D. R. Chen, and W. K. Moon, "Characterization of spiculation on ultrasound lesions," IEEE Trans. Med. Imag. , vol. 23, no. 1, pp. 111–121, Jan. 2004.
[8] M. Noguchi and Y. Shimosato, "The development and progression of adenocarcinoma of the lung," Cancer Treatment Res., vol. 72, pp. 131–142, 1995.
[9] T. V. Colby and C. Lombard. "Histiocytosis X in the lung," Human Pathol. , vol. 14, no. 10, pp. 847–856, 1983.
[10] V. J. Lowe, J. W. Fletcher, L. Gobar, M. Lawson, P. Kirchner, P. Valk, J. Karis, K. Hubner, D. Delbeke, E. V. Heiberg, E. F. Patz, and R. E. Coleman, "Prospective investigation of positron emission tomography in lung nodules," J. Clin. Oncol. , vol. 16, no. 3, pp. 1075–1084, 1998.
[11] K. S. Lee, Y. Kim, and S. L. Primack, "Imaging of pulmonary lymphomas," Amer. J. Roentgenol., vol. 168, no. 2, pp. 339–345, 1997.
[12] J. W. Gurney, "Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis: Part 1. Theory," Radiology, vol. 186, no. 2, pp. 405–413, 1993.
[13] J. J. Erasmus, H. I. McAdama, and J. H. Connolly, "Solitary pulmonary nodules: Part II. Evaluation of the indeterminate nodule," Radiographics , vol. 20, no. 1, pp. 59–66, 2000.
[14] L. Song, X. Liu, L. Ma, C. Zhou, X. Zhao, and Y. Zhao, "Using HOG-LBP features and MMP learning to recognize imaging signs of lesions," in Proc. Comput.-Based Med. Syst. , 2012, pp. 1–4.
[15] X. Ye, X. Lin, G. Beddoe, and J. Dehmeshki. "Efficient computer-aided detection of ground-glass opacity nodules in thoracic CT images," in Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2007, pp. 4449–4452.