IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 3, June, 2020, pp. 351-359
Journal homepage: http://iieta.org/journals/ria

An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study

Eali Stephen Neal Joshua[1*], Midhun Chakkravarthy[1], Debnath Bhattacharyya[2]

[1] Department of Computer Science and Multimedia, Lincoln University College, Kuala Lumpur 47301, Malaysia
[2] Department of Computer Science and Engineering, K L Deemed to be University, KLEF, Guntur 522502, India

Corresponding Author Email: stephen_neal@lincoln.edu.my

https://doi.org/10.18280/ria.340314

Received: 24 March 2020
Accepted: 7 May 2020

Keywords:
lung cancer, machine-learning, ensemble-learning, classification, back-propagation algorithm

ABSTRACT

The Main Objective of this research paper is to investigate the accuracy levels of various machine learning algorithms. To find out the accuracy levels of various classifiers we evaluated the various models employed by researchers were listed out and they have few limitations and their drawbacks were listed out. After a systematic literature study, we found out that some classifiers have low accuracy and some are higher accuracy but not reached nearer of 100%. Therefore, we need to employ a more strategic way for the better classification of Lung cancer nodule. Through a systematic literature survey, the low accuracy levels were due to improper dealing of Dicom images. After an extensive study, we found that ensemble classifier was outperformed when compared with the other machine learning algorithms. Thus, by taking consideration of all the classifiers. The findings we drew was that major machine learning algorithms gave accuracy which was not close to 90%. A better model needs to be employed to increase the accuracy level, and need to be revised should be reliable and meaningful draw insights for the tumour diagnosis, so it reflects our better understanding of the classification of lung cancer. And, lastly, intensive research should be done on the field of Oncology for the better classification of benign and malignant tumours.

1. INTRODUCTION

World-leading health issue -CANCER-an approx. of 13.9 million situations and also 8.29 million fatalities [1] because of on discovery at a beginning of blemishes. In the year 2018, U.S.A. brand-new cancer cells instances of 1,682,210 and also fatalities of 595,690. Lung cancer cells one of the most of all cancers cells leading the fatalities to 158,080 as well as of brand-new instances 224,391 in the year 2016. As a result of the impact of lung cancer cells or otherwise acknowledgement of lung cancer cells at the very early action the survival price is a lot less than various other cancers cells. According to the American Cancer cells organization the public, American lung association, as well as Globe Wellness Company factsheet, there is a boost of survival price of 54.4% from 17.7%, when it is discovered at onset as well as if cancer cells are restricted as well as about 16% is identified at the onset.

1.1 Causes of cancer

In case lung blemishes are identified at a starting phase there's a distinction of most likely boost the diagnosis thus lives could be safeguarded and even lower the opportunity of fatalities a year. Blemishes - denseness with a size about in between 2.9 in addition to 30mm and also could be even more selections of the location of theirs, setting, sizing, form, juxta pleural connected to the lung parenchyma, strong, non-solid, sub strong. A great deal of these might be identified in CT. For the initial exploration of the blemishes, CT is abandoned. LCST (LUNG CANCER CELLS TESTING ROUTES) reveals that lung cancer cells minimized by twenty % and also for around 5-year death for lung cancer cells is reduced in CT than upper body X-ray scanning. Essential to the sensible use of CT testing is the radiologist that is entrusted with determining doubtful sores in the sort of lung blemishes in the CT details. The size of this particular work could be substantial, specifically for little lung blemishes. To the moment of the discovery of theirs on occurrence displays in the National Lung Testing Test, 30 5% of lung cancers cells had sizes which were 10 mm in addition to a lot less. A CT checks gotten from the whole of the lungs as well as re-build with one mm secure areas has about 9,000,000 lung voxels. Lung blemishes with sizes in between 4 additionally as 10 mm inhabit seventy-seven to 1200 voxels or possibly 0.00086% to 0.014% of the lung level, testing radiologists to find out every one of them within search dimensions of in between 2 likewise like 5 minutes under excellent situations. The objective of this short article is to assess today's understanding of lung blemish situating of CT checks as we transfer to the period of typical CT based lung cancer cells watching. For a patient after performing a CT scan, radiologists need to analyze the information in the type of pictures based on nodules morphology and method this ought to be in following the clinical methods straight consequences info with the elements as fatigue, etc. or maybe misinterpretation of information. For the improvement of the information in the type of pictures or maybe the advance picture analysis calls for to the radiologists in the interpretation of the information diagnosed Based on the detection velocity for lung cancer using CT is 2.6 ten times greater than utilizing analogue radiography. To conquer the

issues as well as to bring down the workload the methods recognized as the COMPUTER-AIDED DETECTION i.e. CADe methods are centred on the diagnosed information imaging progression [2] as well as to sense the latent lesions in health. There are several more devices as CADx that are utilized primarily in the distinction of the likely lesions in health.

CADe strategy is designed to the very first detection [3] of the chance lesions at a much better accuracy in addition to lower interpretation time, higher sensitivity with a false positive velocity, low automation, low cost of setup also as to determine the different size, shape, place, and role, to avoid the possible attacks with the usage of the program.

1.2 Machine learning approaches

More than likely one of the most generalized of the collection of the details at the type of data sources [3, 4] with formulas. Info bases are lots of kinds consisting of the personal or maybe the nearby centers in addition to the general public data sources like the lung photos repository corporation , early cancer cells activity strategy, data sources of Japanese culture of radiological modern technology (jsrt), automated blemishes discovery 2009 (anode09), lung photo database consortium along with photo data source effort (lidc idri). Cade along with cadx items for the searching for and also service for the most essential notes for locating the lung cancer cells elements of browsing lately. However, these components currently lugged for different job. Cade strategies do not provide the radiologist's originality of growths, along with cad methods do not discover blemishes and also do not have exceptional ph degrees of computerization. Because of this, these strategies are not yet extensively utilized in scientific approaches. To produce an excellent means of discovery and also medical diagnosis of lung blemishes on ct images, organizing them straight right into a solitary telephone system for the recognition and also characterization of the blemishes to enhance the amount of automation. The blog post likewise provides as payments making use of landmark as well as pie chart of focused slope methods (hog) for differentiating the attainable blemishes from a few other structures along with capability removal for lung blemishes, specifically. For the medical diagnosis, it is based upon the probability of hatred permitting a whole lot even more help in the decision making by the radiologists. An innovative classifier as well as additionally assistance vector maker (svm), artificial semantic network [4] have been used to remove incorrect positives. The relational db made use of in this certain evaluation contains 520 instances acquired arbitrarily from numerous public domain datasets. The technique which we used below is the division with a precision of 97% in addition to an excellent system discovery formula with a level of sensitivity of 94.4% with 7.04 wrong positives a circumstance. Various type of blemishes (separated, juxta pleural, juxta vascular additionally as ground glass) with sizes in between 3 mm additionally as 30 mm have been understood. We have taken the roc as well as aoc to discover the hatred of the cancer cells cell: 0.96 for blemishes unlikely of being deadly, 0.80 for blemishes reasonably unlikely of being deadly, 0.72 for blemishes with indeterminate deadly cells, 0.67 for blemishes reasonably questionable of being spiteful and also 0.83 for blemishes anxious of being evil-minded.

2. MOTIVATION

Based on the 10000 projects in 2017 alone, approximately 18.1 million rare new types of disease cancer occurred globally, and this caused the deaths approximately 15.6% of the death [4]. 520 individuals having between one as well as eight pulmonary nodules. Some of them, thirty-one with nodules very improbable of being malignant, [1] from the Figure 1 we can clearly state that 64% with nodules sensibly improbable of being malignant, 149 with nodules with undetermined malignant cells, 78% with nodules [4] moderately suspicious of being malevolent and 62% with nodules very suspicious of being malevolent.

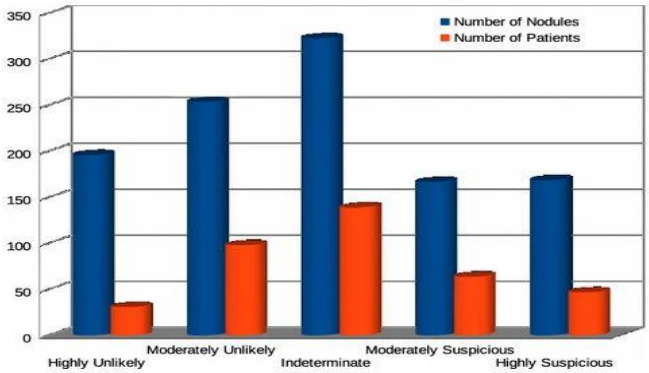

Figure 1. Likelihood of cancer malignancy [4]

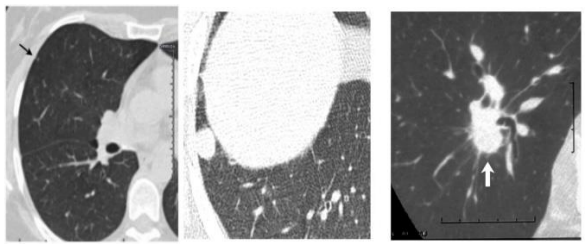

Figure 2. CT scan section of lung nodule

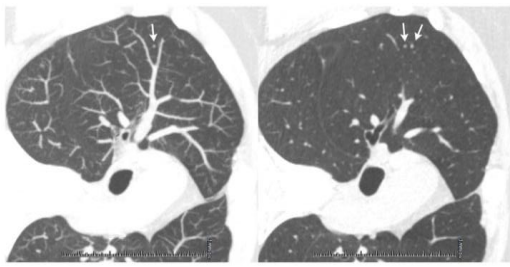

Figure 3. 1.25 mm heavy transverse CT

The 1.23 mm large CT section shows an approximately 2.6 mm long Nodule. Nodule with the larger malignant tumours can be found normal at any normal CT scan received from the previous young adulthood. From Figure 2 we can clearly state from the above assumptions that lump of 2.6 mm Long Nodule can ultimately lead to relevance to their identification. There may be a chance that this lump might be spread to the other body parts or new findings in a patient [5] with the proven document results. Hence there is no scope of finding a lung cancer screening from the image. From the above two images clearly, I can state that (12.9 MM lung images) the CT scan image on the left side was visible and we can predict that the

person is suffering from lung cancer. The nodule [6] which was on the right side clearly states that pulmonary blood vessels. Although it is significantly larger compared with the image on the left side with this conclusion we can differential normal lump nodule with abnormal growth of lump nodule

In the Figure 3, 1.27 mm serious slanting CT section revels 2.8 mm rounded opacities appearing as lung nodules. An eight mm stable TS MIP centred on the slanting segment above reveals that only the posterior of the two opacities is a lung nodule (arrow in the bottom image) along with the forward clarity corresponds to a part of a regular blood vessel. Detection [6, 7] of suspicious nodule candidates. TM can be used to detect the pulmonary nodules efficiently, but at the same time, many vessels are also detected as nodules in this manner. The intensity value of nodules usually has a Gaussian-like distribution:

$$q(r)=q_{\max} * e^{-(r/\rho)2}, 0\leq r\leq R \quad (1)$$

where,

$$\rho= R\ (\ln(q_{\max}) - (\ln(q_{\min}))^{-0.5} \quad (2)$$

in which, $q_{\min}$ and $q_{\max}$ are the minimum and maximum of nodule intensity, $q(r)$ is the intensity of nodule at space from the centroid, of the nodule and R is the radius of the nodular template.

TM method computes the normalized cross-correlation (NCC) of the image and the template in the equation:

$$\Gamma=\frac{\sum_{x,y}[f(x,y)-f_{u,v}][t(x-u,y-v)-t]}{\sqrt{\sum_{x,y}f(x,y)-f_{u,v}]^2\sum_{x,y}[t(x-u,y-v)-t]^2}} \quad (3)$$

where,

$$\overline{f_{u,v}}=\frac{1}{N_x N_y}\sum_{x=u}^{u+N_x-1}\sum_{y=V}^{v+N_y-1}f(x,y)$$

$f(x, y)$ and $f(u, v)$ are the input picture and the mean value of it's in the region under template placed at (u, y), t(x, v), is the nodule style, as well as it's mean great, is t̄. In (two), the correlation coefficient γ varies between one and' one. In the event the correlation coefficient is γ above an optimistic threshold amount, the point (u, v) is approved to belong to a doubtful nodule candidate. This particular threshold amount based on the database of ours is experimentally set to 0.6.

From the experiments of ours, we utilized a library of semi-circular and circular guides with 90° rotations with a Gaussian division of grayscale intensity. The hostile diameter of nodules in the database of ours is 8.5 mm with a regular deviation of 3.6 mm. appropriately, we chose a diameter of 8.5 mm for our semi-circular and circular guides the just like. To use this technique, numerous regions are tagged, but the real good speed is high. Figure three shows a sample of this particular strategy. The template impression is viewed as Figure 4(a) After eliminated interference, the pulmonary parenchyma impression is fallen by best industry segmentation and viewed as Figure 4(b). The normalized cross-correlation of the template as well as the picture is viewed as Figure 4(c) and the original areas of interest are revealed as Figure 4(d).

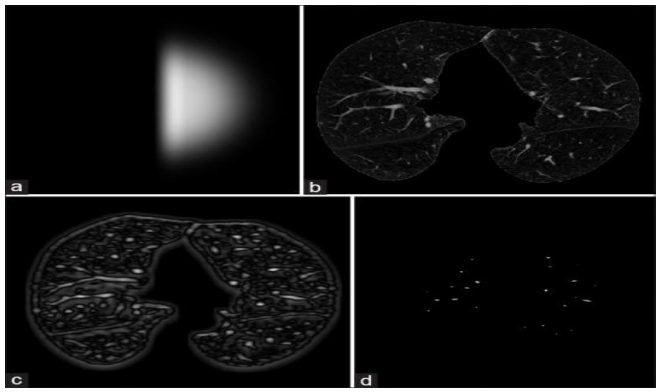

Figure 4. Suspicious nodule candidate detection. (a) Template, (b) original image, (c) normalized cross-correlation of image and template, (d) result of template matching method

It should be noted that in the above procedure, some vessels are also detected as suspicious nodule candidates. 2D processing is frequently not able to remove the interference of vascular, leading to a high false-positive rate. To separate the nodules from vessels in the next section, nodule candidates are segmented, and then these segmented nodule candidates are classified using some features.

2.1 Segmentation process

Accurate segmentation of suspicious nodule candidates, that could be a nodule or a vessel, is very important. Segmentation of nodule candidates is challenging because of the following three effects: noise, lump that are attached to blood vessel or attached to lung barrier, and low contrast of intensity values between nodules and other structures in CT images [8], due to low-dose CT imaging.

The essential thought in speed curve models is to let a curve to deform to reduce a given cost function to provide the desired segmentation results. In general, there are two main classifications of active curve models, Side-based with location-based.

Side-based active curve models mainly utilize the edge and gradient information to drive the contours to identify object boundaries. The result of image segmentation by these models is highly dependent on the initial contour placement and very sensitive to image noise.

Compared with edge-based active contours, region-based active contours model the foreground and background regions statistically and optimize a global energy cost function. These models are less receptive to initialization and picture noise. Region-based active contours focal point on a localized power function base on piece-wise stable model of Chan and Vese (C-V model), which can be written as:

$$F(u,v,\phi)=\mu\int_{\Omega}|\nabla H(\phi)|dx+v\int_{\Omega}H(\phi)dx+\lambda_1\int_{\Omega}H(\phi)(I-u)^2dx+\lambda_2\int_{\Omega}(1-H(\phi))(I-v)^2dx \quad (4)$$

where, I am a given picture defined on domain Ω, υ, and are v the worldwide mean intensities of the interior and exterior regions. λ_1, λ_2, v, and μ are positive constant coefficients and is $H(\Phi)$ Heaviside function.

Optimizing global statistics usually is not ideal for segmenting heterogeneous objects. Therefore, to precisely part, these objects, a new model of the active contour is developed

which utilizes local information.

Localized active contours are capable of segmenting objects with heterogeneous feature profiles. In this approach, segmentation is not based on global section models. The average intensities in interior and exterior areas of a mask $B(x, y)$ are computed at each point. To optimize [9] the total energy of the contour, the mask is considered in each point separately, and the point is moved to decrease the energy function. This energy function is defined as follows:

$$E(\phi) = \int_{\Omega} \delta\phi(x) \int_{\Omega} B(x,y).F(I(y),\phi(y))dxdy + \lambda \int_{\Omega} \delta\phi(x) \| \nabla\phi(x) \| dx \quad (5)$$

F is general inner product energy, and $\delta\Phi(x)$ is the Dirac function. Using $B(x, y)$, F operates only on local image information in the neighbourhood of (x, y), and λ is a smoothing parameter. Using the calculation of variation results:

$$\frac{\partial\phi}{\partial t}(x) = \delta\phi(x) \int_{y} B(x,y)\delta\phi(y).((I(y)-u_x)^2 - (I(y)-v_x)^2)dy + \lambda\delta\phi(x)div(\frac{\nabla\phi(x)}{|\nabla\phi(x)|}) \quad (6)$$

The local equivalents of and that defined in terms of the $B(x, y)$ function, u_x and v_x:

$$u_x = \frac{\int_{\Omega} B(x,y)H\phi(y)I(y)dy}{\int_{\Omega} B(x,y)(1-H\phi(y)dy} \quad (7)$$

$$v_x = \frac{\int_{\Omega} B(x,y)(1-H\phi(y)).I(y)dy}{\int_{\Omega} B(x,y)(1-H\phi(y))dy} \quad (8)$$

Local active contour has three parameters, maximum iterations which are set to 200, local radius ($B(x, y)$) set to 1, and smoothing parameter (λ) set to 0.2. The parameters are chosen according to nodule specifications experimentally.

Figure 4 compares the results of global and local active contours applied to the segmentation of a nodule. It can be seen that local active contour successfully segmented the nodule candidates with concave and low contrast boundaries.

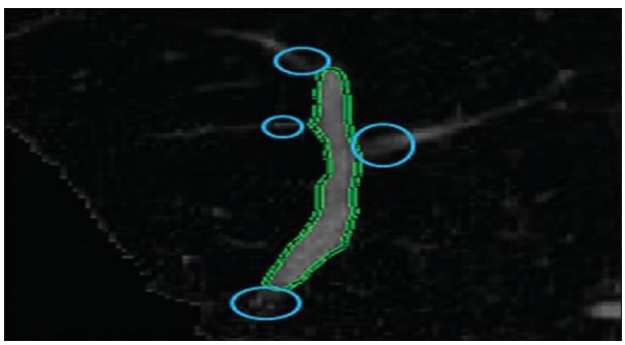

Figure 5. (a and c) Initialization (b) result with worldwide energies, (d) final result with limited energies

This model was tested on different kinds of nodule candidates including both nodules [10] and vessels. The experiments show the robustness and reliability of the model. As shown in Figure 5, this method successfully segmented nodule candidate regions from adjacent structures in the majority of cases but failed to discriminate between background and some vessels. Figure 5 shows a vessel segmented by localized active contour. As it is shown in this figure, some gaps are generated within vessels. It should be noted that such a failure makes no problem in the final results because vessels are identified from nodules and omitted in later stages.

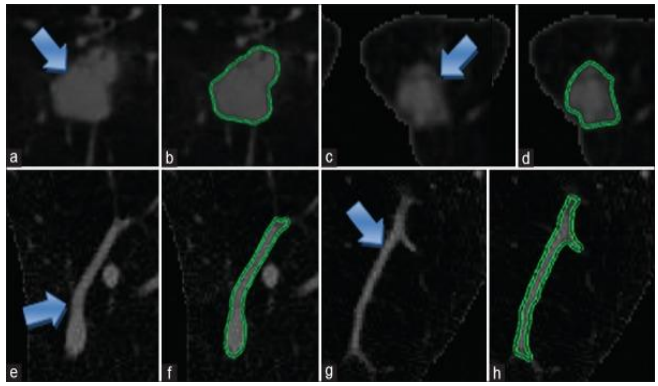

Figure 6. Segmentation results. (a and c) are nodule and (e and g) are the vessel that is shown by blue arrows, (b, d, *f* and h) are the contour of the segmentation

In each CT cross-section, some vessels may resemble nodules [11] causing false nodule candidates that must be identified and omitted. The following three steps do the job. Non-nodule objects can be discriminated from nodule candidates using the information of nodules such as size, volume, sphericity, and compactness.

From the Figure 6 Detection of vessels parallel to computed tomography cross-plane geometrical shapes of pulmonary nodules and vessels are spherical and cylindrical, respectively. A slice in a sphere, semi-sphere, vertical, or nearly vertical cylinder is circular or nearly circular. As the cylinder inclines more parallel to CT cross-plane, the difference between the long diameter and the short diameter increases. The long diameter is the distance between two points of nodule boundary [12] with maximum distance, and the short diameter is the longest chord perpendicular to the long diameter.

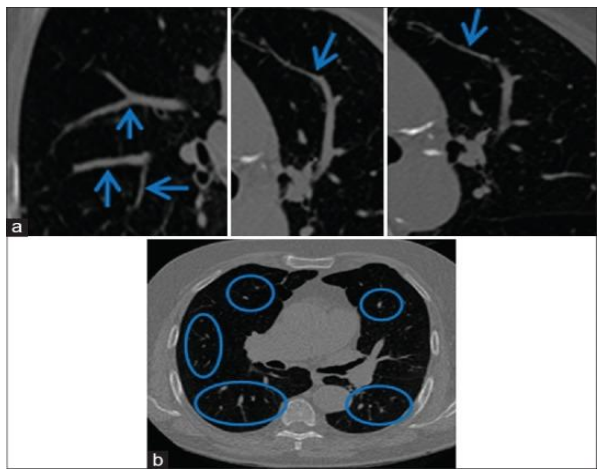

Figure 7. (a) Blue arrows show vessels detected in section 2.4.1(b) Blue circles show vessels which resemble nodules but not detected in section 2.4.1 Detection of oblique blood vessels

If we consider the centre of gravity of a nodule or vertical vessel in two successive slices [13], there would be a small

displacement between these two centres in Figure 7, whereas this displacement would be greater for an oblique vessel. The diagrams in Figure 9 show the distributions of these displacements for nodules and vessels. Hence, several oblique blood vessels in nodule candidates can be identified using a threshold on these displacements. The diagram vertical axis is the number of nodules/vessels with specific displacements of the horizontal axis. The threshold value is experimentally set to 0.75.

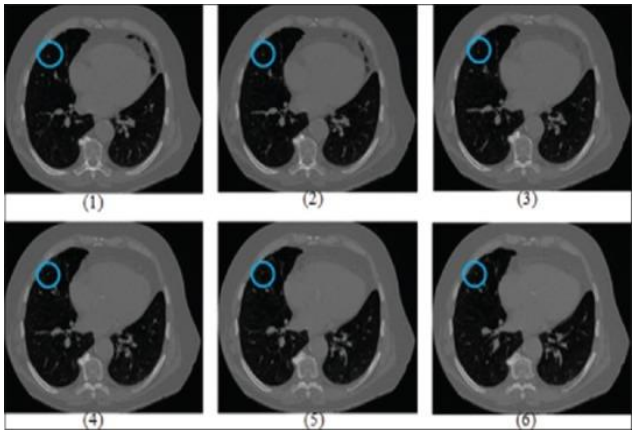

Figure 8. Vessel shown inside the blue circles in consecutive six computed tomography slices, vertical to computed tomography cross-plane detection of vessels vertical to computed tomography cross-plane

From the Figure 8 in the previous two steps, oblique and parallel [14, 15] vessels were removed from the suspicious nodule candidate list. In the current step, we will only consider the remaining suspicious nodule candidates. A vessel is not a compact object and continues to be connected in consecutive slices of CT images. On the contrary, a nodule as shown in Figure 11 is a compact and sphere object, so its cross-section on a CT slice is nearly a semi-circle.

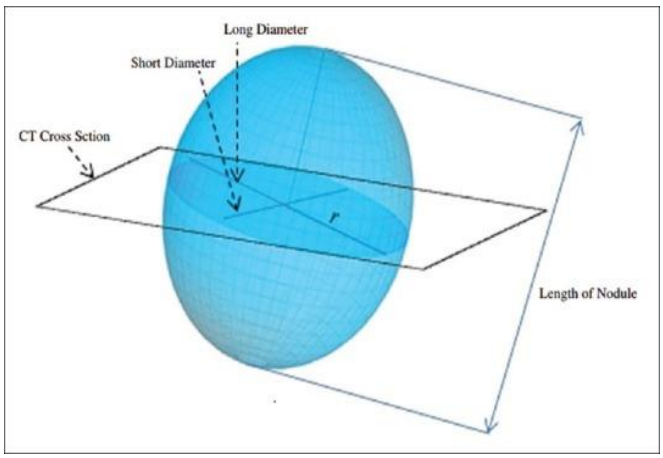

Figure 9. Feature parameters of a nodule

Experimental results indicated that there is a small difference between the length and short diameter of a nodule, whereas the difference would be higher for vascular areas [16, 17] because vascular areas continue to be connected in successive slices of 2D CT images [9]. Thus, the length and short diameter of remaining suspicious nodule candidates are extracted as features. Furthermore, the length of nodule candidates is calculated as follows:

$$\text{Length} = n \times \text{slice thickness} \quad (9)$$

n is the number of slices that transverses a nodule candidate. In this study, a slice thickness of data is mm. That we can clearly observe in Figure 4 and 9.

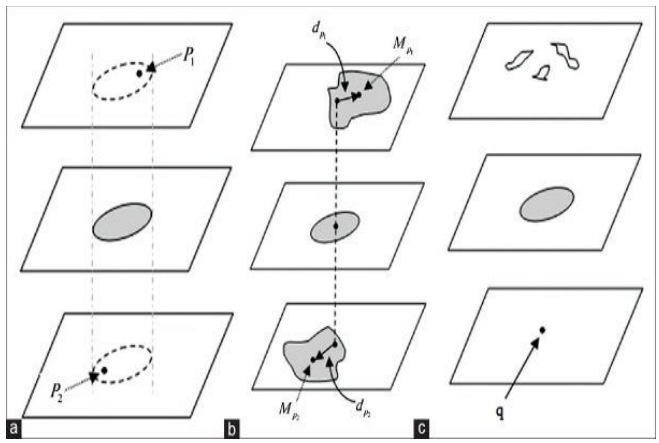

Figure 10. (a) Active contour of the current slice and the two corresponding regions in previous and next slices. (b) Single regions segmented by local active contours in previous and next slices. (c) Multiple regions or one pixel

(1) Three slices, the current slice, and slices before and after it are considered

(2) The cross-section area of the nodule [18] is already segmented by local active contour

(3) Its relative area in previous and next frames are marked, and the brightest pixels in these marked areas are determined [19] and labelled by points P_1 and P_2, respectively

(4) P_1 and P_2 are selected as the centre of the initial local active contours

(5) The radius of the active contours is 2.

The following cases may occur as shown in Figure 10.

(1) Each contour in the previous/next slice has just one region. $M_{p1} \mid M_{p2}$ is the centre of gravity in this region. If the following two conditions are fulfilled, the candidate area is considered to belong to the current nodule candidate [20]

a. The average intensity in this region be greater than a threshold

b. The displacement of its centre of gravity concerning the centre of gravity of the current slice, shown with vectors $d_{p1} | d_{p2}$ be smaller than a threshold

(2) If the region segmented in the previous/next slice has multiple areas or has just one pixel, as in Figure 10c, the previous/next slice does not belong to the current nodule candidate.

According to the above algorithm, the total number of slices belonging to the nodule n and so its length is determined.

3. LITERATURE SURVEY

A thorough literature survey is described in Section 3. Section 4 deals with the performance evaluation of models. Results metrics are included in Section 5. Future Enhancements are distinguished in Section 7.

Syed et al. [1] proposed their work on "Comparative Analysis of learning Algorithms for Lung Cancer

Identification". In this work, the Authors worked on two types of features i.e.., benign vs Malignant. The proposed work tried to automate the entire process of detecting the tumour automatically. Authors conducted the experiments on the two phases:

First Phase: Identification of the most significant features from the CT Scan Images and performing the mapping.

Second Phase: Machine learning algorithms are applied.

Dash et al. [3] recommended their job "Multi-Classifier Structure for lung cells category". In this job, the Authors have made use of High-Resolution Computed Tomography (HRCT) checks for the discovery of lung cancer cells. The classifier uses the Discrete Wavelet Transform and also Several Classifier to obtain the first choice on the input picture. From the input photo, they have drawn out the functions and also were provided as input to the Semantic network Classifier and also Ignorant Bayes Classifier and afterwards Choice Combination was used leading to the Accurate Choice.

Günaydin et al. [5] proposed their work on "Comparison of Lung Cancer Detection Algorithms". In this work, Authors worked on Lung cancer occurrence on both men and woman, Authors used various classifiers like Principal Component Analysis, K-Nearest Neighbors, Support Vectors Machines, Naïve Bayes, Decision Tree, and Artificial Neural Network and various machine learning methods to detect anomalies. He calculated the Accuracy, sensitivity, and specificity by applying the machine learning algorithms. The authors didn't apply noise removal methods on the image and used lung part radiography (Table 1-3).

Makaju et al. [8] proposed their work on "Lung Cancer Detection using CT Scan Images". In this work, the Authors stated that instead of working on various filters like Gabor Filter, Median Filter, and Gaussian Filter they used Watershed Segmentation (Table 2, 3).

Radhika et al. [9] proposed their work on "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms". In this work, the Authors worked on four classifiers viz SVM model, Naïve Bayes model, Decision Tree model, Logistic Tree model, and applied these classifiers on BRATS Dataset, OASIS Dataset, NBTR Dataset are taken from UCI Machine learning repository. And compared the accuracy levels among these algorithms in Table 1.

Roy et al. [10] proposed their work on "A Comparative study of lung cancer detection using supervised neural network". In this work, the authors have done extensive research on classification methods used for the diagnosis and detection of lung carcinoma. He used Bio-medical images and machine learning algorithms to classify. The approach he followed is (Table 1, 3).

Step 1: Converting RGB image to Grayscale
Step 2: Applying Linearization Techniques
Step 3: Detecting and segmentation the Image
Step 4: Extracted the Features
Step 5: Cross-Validation

Song et al. [11] proposed their work" Prognosis of Stage I Lung Cancer Patients through Quantitative Analysis of Centrosomal Features", this gives nodules of cancer area with marks. And also, they used addition features like Centroid, Diameter, and pixel means intensity have been extracted for the detection of the cancer nodule. The strength of this work is:

(1) The accuracy of detecting lung cancer nodule has increased.
(2) Detects whether the cancer cell is benign or malignant.
(3) False detection parameters like salt-pepper noises and speckle noise were removed.

Tripathi et al. [12] proposed their work on "A Comparative Analysis of segmentation Techniques for Lung Cancer Detection". In this work, the authors proposed the various image segmentation work like simple thresholding, Otsu thresholding, Edge Thresholding, DE based Segmentation, Marker Controlled Watershed Segmentation. In recent years there are enough techniques were developed for the identification of lung cancer. They combined various classifiers techniques with different segmentation algorithms for the identification of lung cancer nodule using image processing. Here in this paper he used SVM classifier and applied on various image segmentation techniques in Table 1.

4. COMPARATIVE STUDY AND PERFORMANCE METRICS

Lung Nodules Detection Techniques with Machine Learning Algorithms Table 1 shows the various models employed by the authors.

Table 1. Various Machine Learning Algorithms deployed for classification

Author	Technique Applied
(Tripathi, Tyagi, and Nath 2019)	Image Segmentation Technique
(*Radhika, Nair, and Veena 2019*)	Decision Trees, Logistic Regression, Support Vector Machines
(*Roy et al. 2019)*	Random Forest, Support Vector Machines
(*Abbas Ali et al. 2018*)	Support Vector machines
(*Günaydin, Günay, and Şengel 2019*)	Artificial Neural Network, Naive Bayes, Decision Tree
(*Makaju et al. 2018*)	Support Vector Machines
(*Dash et al. 2014*)	Neural Network Classifier, Naive Bayes, Decision Fusion
(*Song et al. 2012*)	Linear Discriminator Analysis (LDA), Support Vector Machines
(*Vijai Anand 2010*)	Back Propagation Network
(*Lynch et al. 2017*)	Artificial Neural Network
(*Günaydin, Günay, and Şengel 2019*)	Back Propagation Network (BPN)

Table 2. Comparative study of classification accuracy

Algorithm	Accuracy	Sensitivity	Specificity
Image Segmentation Technique	59.12%	65.4%	65.4%
Decision Trees	58.11%	72.67%	72.67%
Logistic Regression	65.4%	84.34%	96.66%
Random Forest	62.5%	64.28%	64.28%
Support Vector machines	62.5%	88.5%	88.5%
Artificial Neural Network	65.4%	65.4%	65.4%
Naive Bayes	62.5%	62.5%	62.5%
Ensemble Classifier	88.5%	72.67%	65.4%
Decision Fusion	72.67%	74.78%	62.5%
Linear Discriminator Analysis (LDA)	74.78%	62.5%	91.66%
Back Propagation Network	72.67%	72.67%	72.67%
K-Nearest	72.24%	94.12%	53.78%

Table 3. Comparison of various machine learning classifiers

	S. Deviation Train	S. Deviation Test	ROC	AUC	Gmean	Precision	Recall	F1_score
AdaBoost	0.15726	11.74342	0.78319	0.50000	0.65094	0.73469	0.83721	0.78261
Decision Tree	0.15726	9.18525	0.79361	0.50000	0.66322	0.75000	0.83721	0.79121
Extra Forest	0.63311	7.90518	0.80644	0.50000	0.67843	0.74510	0.88372	0.80851
Extra Tree	0.15726	6.78843	0.83648	0.50000	0.71894	0.80435	0.86047	0.83146
Gaussian Naive Bayes	0.63547	5.58525	0.88857	0.50000	0.80128	0.90244	0.86047	0.88095
Gaussian Process	1.05988	6.81512	0.86531	0.50000	0.76870	0.89744	0.81395	0.85366
k-Nearest Neighbors	0.81683	7.24203	0.84448	0.50000	0.73553	0.85366	0.81395	0.83333
LDA	1.06127	8.35666	0.82243	0.50000	0.70560	0.82927	0.79070	0.80952
Linear SVM	0.15726	11.74342	0.78319	0.50000	0.65094	0.73469	0.83721	0.78261
Logistic Regression	1.33854	7.25686	0.85489	0.50000	0.75185	0.87500	0.81395	0.84337
Multilayer Perceptron	0.67524	5.97748	0.88736	0.50000	0.80237	0.92308	0.83721	0.87805
QDA	0.76297	6.91503	0.85611	0.50000	0.75074	0.85714	0.83721	0.84706
Random Forest	0.62897	6.18669	0.80281	0.50000	0.67604	0.77778	0.81395	0.79546
RBF SVM	0.86054	7.49009	0.88614	0.50000	0.80388	0.94595	0.81395	0.87500
SGD Classifier	9.98671	10.26020	0.68629	0.50000	0.61694	0.59155	0.97674	0.73684
Support Vector Machine	0.86054	7.49009	0.88614	0.50000	0.80388	0.94595	0.81395	0.87500

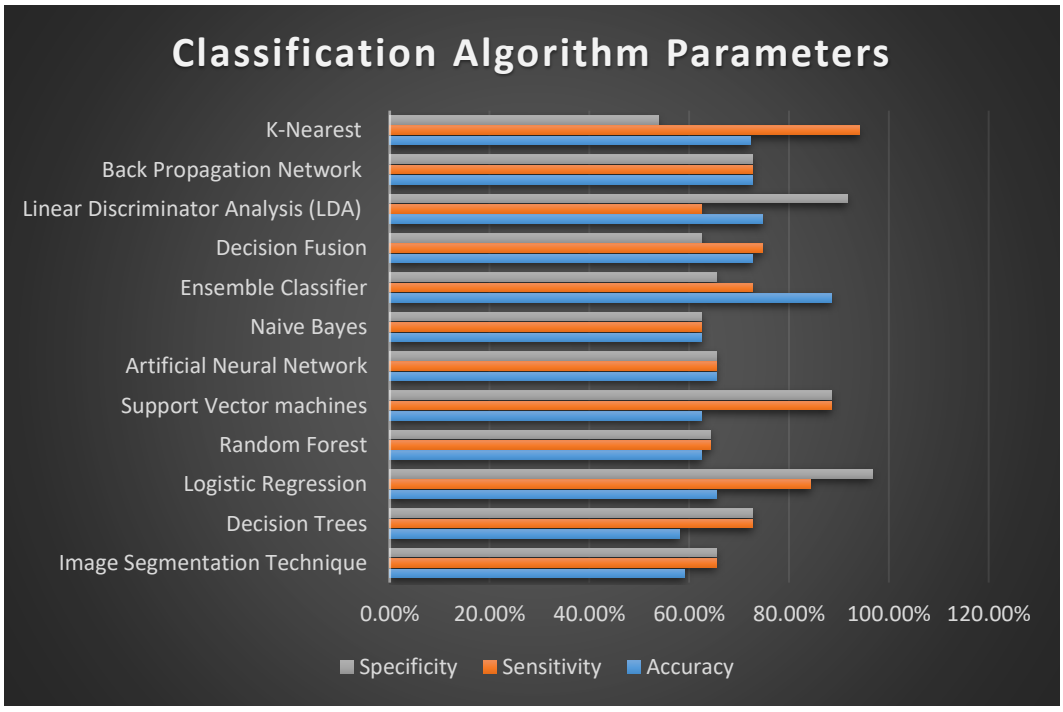

Figure 11. The accuracy levels of various machine learning algorithms for lung cancer classification

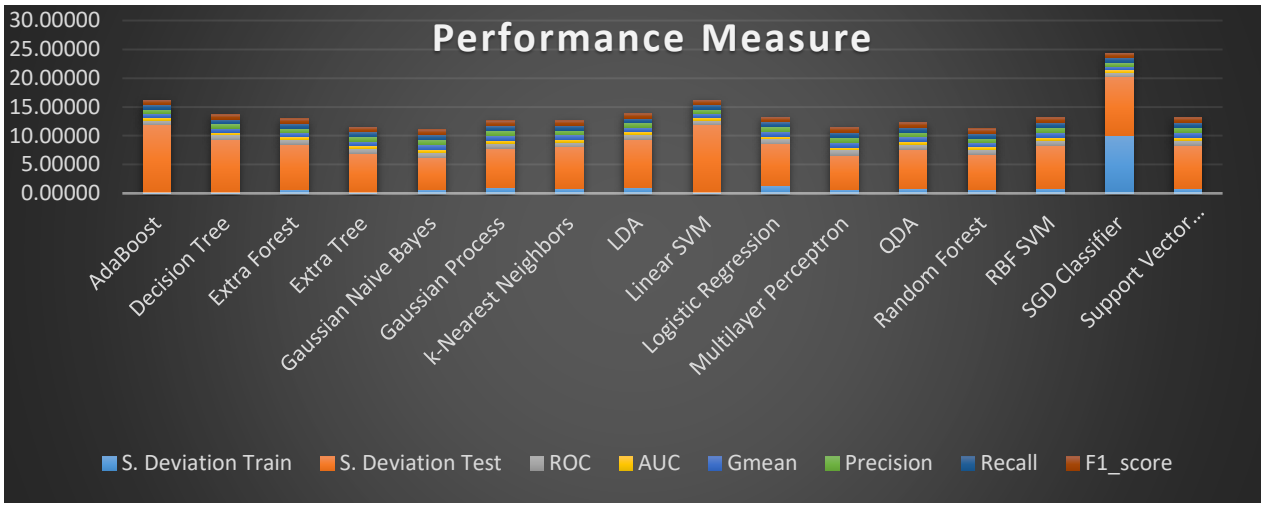

Figure 12. The performance levels of each classifiers

5. RESULTS DISCUSSIONS

Table 2 shows the performance metrices of various machine learning algorithms, by taking considerations of Standard deviation, ROC, AUC, Accuracy, Sensitivity and Specificity parameters. After a thorough research we came to conclusion that all the classifiers are not close to the 100% accuracy. Out of all the classifiers ensemble outperformed well. In the Table 3 we have tabulated the various parameters to determine the algorithms efficacy for the medial images from systematic literature survey. And the Figure 11 and 12 projects the performance metrices of the various machine algorithms and their efficacy.

6. CONCLUSION

This paper summarizes on extensive study on various machine learning algorithms to classify malignancies of the lung cancer detection. After a systematic literature review on a various research article, it was found that the classification of the benign and malignant tumour was not predicted accurately. Majorly we identified three main gap identification with a thorough literature survey. In, that the first one is, the accuracy of the various classifier mentioned in the article was not satisfactory. Secondly, novel techniques have to employed for the medical image's especially Dicom(Digital Imaging and Communications in Medicine) and MRI(Magnetic Resonance Imaging) images for noise reduction. Thirdly, earlier machine learning algorithms like SVM(Support Vector Machine), Naive Bayes, K-Nearest Neighbours, Decision Trees, used to take raw images into the account due to which we will miss few hidden patterns. Lastly, we are working with medical informatics a small minute error can result in erroneous results, so proper classifier with improvised algorithmic technique should be employed for the better accuracy. Intensive research has to be performed in the field of cancer studies. So, we can minimize the cause of death due to cancer.

7. FUTURE ENHANCEMENTS

Potential Work could help to find suitable inputs from very few selected models. While SVM, Decision tree as well as Back Propagation Network classifiers had been found the least precision results. For the greater accuracy amounts rather than based on the single classifier for the prediction of malignant and benign tumors it will be suggestible to make use of the ensemble method with Extensive Deep Neural Network for more effective yielding of the outcome. A brand new attempt to assess every entity criterion and just how it pertains to patient survival, particularly in the situation of longer-lived individuals, may be crucial to far more correct as well as prognostic co-relational,semi-supervised machine learning algorithms. Due to the very poor overall performance in the longer survival times, separating the issue into individuals that survive under thirty-five weeks and greater than thirty-five weeks might deliver enhanced designs, which includes much more correct classifiers in the crucial subset with less time of survival. Focusing on much less survivable people might help to find the clinically evident arguments that enhanced predictions for these individuals are more crucial than for longer-lived individuals. Evaluation of scientifically substantial cutoff survival times relatively associated to the attention of the entire variety might aid focusing the classification issue as well as acquiesce an easy to predict subject. The terrible overall accuracy in the higher withstanding period furthermore implies that the collected variables might be insufficient to anticipate long-expression endurance. Further labor can check out the skills of unit deviations apart from those evaluated in this specific study and various sets of years.

ACKNOWLEDGMENT

I sincerely thank Dr. Debnath Bhattacharyya for guiding through out the process. This work was supported by the Research Lab of Vignan's Institute of Information Technology, Visakhapatnam, India.

REFERENCES

[1] Syed, A.A., Fatima, W., Wajahat, R. (2018). Comparative analysis of learning algorithms for lung cancer identification. Indian Journal of Science and Technology, 11(27): 1-9.

https://doi.org/10.17485/ijst/2018/v11i27/130707
[2] Cruz, C.S.D., Tanoue, L.T., Matthay, R.A. (2011). Lung cancer: Epidemiology, etiology, and prevention. Clinics in Chest Medicine, 32(4): 605-644. https://doi.org/10.1016/j.ccm.2011.09.001
[3] Dash, J.K., Mukhopadhyay, S., Garg, M.K., Prabhakar, N., Khandelwal, N. (2014). Multi-classifier framework for lung tissue classification. 2014 IEEE Students' Technology Symposium, Kharagpur, India, pp. 264-269. https://doi.org/10.1109/TechSym.2014.6808058
[4] Firmino, M., Angelo, G., Morais, H. (2016). Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. BioMedical Engineering Online, 15(1): 1-17. https://doi.org/10.1186/s12938-015-0120-7
[5] Günaydin, Ö., Günay, M., Şengel, Ö. (2019). Comparison of lung cancer detection algorithms. 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, Istanbul, Turkey, pp. 1-4. https://doi.org/10.1109/EBBT.2019.8741826
[6] Kadir, T., Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. Translational Lung Cancer Research, 7(3): 304-312. https://doi.org/10.21037/tlcr.2018.05.15
[7] Lynch, C.M. Abdollahi, B., Fuqua, J.D., de Carlo, A.R., Bartholomai, J.A., Balgemann, R.N. van Berkel, V.H., Frieboes, H.B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. International Journal of Medical Informatics, 108: 1-8. http://dx.doi.org/10.1016/j.ijmedinf.2017.09.013
[8] Makaju, S., Prasad, P.W.C., Alsadoon, A., Singh, A.K., Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125: 107-114. https://doi.org/10.1016/j.procs.2017.12.016
[9] Radhika, P.R., Nair, R.A.S., Veena, G. (2019). A comparative study of lung cancer detection using machine learning algorithms. 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, pp. 1-4. https://doi.org/10.1109/ICECCT.2019.8869001
[10] Roy, K., Chaudhury, S.S., Burman, M., Ganguly, A., Dutta, C., Banik, R. (2019). A comparative study of lung cancer detection using supervised neural network. 2019 International Conference on Opto-Electronics and Applied Optics (Optronix), Kolkata, India, pp. 1-5. https://doi.org/10.1109/OPTRONIX.2019.8862326
[11] Song, D., Zhukov, T.A., Markov, O., Qian, W., Tockman, M.S. (2012). Prognosis of stage I lung cancer patients through quantitative analysis of centrosomal features. 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, pp. 1607-1610. https://doi.org/10.1109/ISBI.2012.6235883
[12] Tripathi, P., Tyagi, S., Nath, M. (2019). A comparative analysis of segmentation techniques for lung cancer detection. Pattern Recognition and Image Analysis, 29(1): 167-173. https://doi.org/10.1134/S105466181901019X
[13] Vijai Anand, S.K. (2010). Segmentation coupled textural feature classification for lung tumor prediction. 2010 International Conference On Communication Control And Computing Technologies, Ramanathapuram, pp. 518-524. https://doi.org/10.1109/ICCCCT.2010.5670607
[14] Xiao, Y., Wu, J., Lin, Z., Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. Computer Methods and Programs in Biomedicine, 153: 1-9. http://dx.doi.org/10.1016/j.cmpb.2017.09.005
[15] Thakur, S.K., Singh, D.P., Choudhary, J. (2020). Lung cancer identification: A review on detection and classification. Cancer Metastasis Reviews. https://doi.org/10.1007/s10555-020-09901
[16] Richter, A.N., Khoshgoftaar, T.M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artificial Intelligence in Medicine, 90: 1-14. https://doi.org/10.1016/j.artmed.2018.06.002
[17] Nishio, M., Sugiyama, O., Yakami, M., Ueno, S., Kubo, T., Kuroda, T., Togashi, K. (2018). Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. PLoS ONE, 13(7): e0200721. https://doi.org/10.1371/journal.pone.0200721
[18] Kadir, T., Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. Transl Lung Cancer Res, 7(3): 304-312. https://doi.org/10.21037/tlcr.2018.05.15
[19] Limkin, E.J., Sun, R., Dercle, L., Zacharaki, E.I., Robert, C., Reuzé, S., Schernberg, A., Paragios, N., Deutsch, E., Ferté, C. (2017). Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Ann Oncol., 28(6): 1191-1206. https://doi.org/10.1093/annonc/mdx034
[20] Binson, V.A., Subramoniam, M. (2018). Advances in early lung cancer detection: A systematic review. 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, pp. 1-5. https://doi.org/10.1109/ICCSDET.2018.8821188