Diagnosis of Starting Stage Lung Cancer Detection using Machine Learning

Mrs.A.NANDHINI MCA, (Ph.D)
Assistant Professor,
Department of Computer Applications, Nehru College of Management, Coimbatore.

Ms MOUNIKA R. IIMCA,
Department of Computer Applications, Nehru College of Management, Coimbatore.

ABSTRACT—The exact lung cancer identification is a critical problem that has attracted the researchers' attention. The practice of multiview single image and segmentation has been widely used for the last 2 years to improve the identification of lung cancer disease. The utilization of machine learning (ML) and deep learning (DL) techniques can significantly expedite the process of cancer detection and stage classification, enabling researchers to study a larger number of patients in a shorter time frame and at a reduced cost applying the image segmentation approach herein, the multiresolution rigid registration mechanism is applied to enhance the segmentation further. Techniques like principle component averaging and discrete wavelet transform are verified for the image fusion development. To review the performance of the suggested technique, the image database resource initiative-based lungs image database consortium is tested in this paper which includes 4,682 computed tomography scan images of 61 patients with nodules sizes from 3 to 30 mm. According to the study finding, the outperformed results of our model are obtained in terms of feature mutual information, and peak signal-to-noise ratio, which were recorded at 0.80 and 19.25, respectively. Moreover, the detection and stages of cancer (STG-1, STG-2, STG-3, and STG-4) of lung nodules are also assessed by using the ResNet-18 convolutional neural network classifier. With only 1.8 FP/scan, the achieved accuracy and sensitivity for detection are 98.2% and 96.4%, respectively. The study's findings show that our proposed strategy outperforms existing models significantly. Therefore, the proposed models have the potential to be implemented in clinical settings to provide support to doctors in the early diagnosis of cancer, while minimizing the occurrence of false positives in scans.

INDEX TERMS Early detection, lung cancer, CNN,ANN and RCNN

I INTRODUCTION

Cancer is one of most dangerous disease that causes deaths. Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. Cancer can start almost anywhere in the human body, which is made up of trillions of cells.Lung cancer is the leading cause of cancer deaths worldwide.When cancer starts in the lungs, it is called lung cancer. People who smoke have the greatest risk of lung cancer, though lung cancer can also occur in people who have never smoked.

Treatment and prognosis depend on the histological type of cancer, the stage (degree of spread), and the patient's performance status. Possible treatments include surgery, chemotherapy, and radiotherapy survival depends on stage, overall health, and other factors, but overall only 14% of people diagnosed with lung cancer .The two general types of lung cancer include: (1)Small cell lung cancer, (2)Non-small cell lung cancer. SMALL CELL LUNG CANCER is almost related with smoking and grows more quickly and form large tumors that can spread widely through the body. cnNON-SMALL CELL LUNG CANCER (NSCLC) is the most common type of lung cancer. It accounts for over 80% of lung cancer cases.The general symptoms of lung cancer include coughing up blood, chest pain, weight loss and loss of appetite, shortness of breath and feeling weak. Each type of lung cancer grows and spreads in different ways, and is treated differently.

Analyzing the exponentially growing cancer-associated databases poses a major challenge to researchers.Image processing is the process of transforming an image into a digital form and performing certain operations to get some useful information from it. The image processing system usually treats all images as 2D signals when applying certain predetermined signal processing methods. Image processing methods such as noise reduction, feature extraction, identification of damaged regions, and maybe a comparison with data on the medical history of lung cancer are used to locate portions of the lung that have been impacted by cancer.

Recent progress in machine learning (ML) and deep learning (DL) techniques has resulted in a significant shift towards computer-aided detection (CAD) systems for lung cancer detection. Machine learning algorithms are molded on a training dataset to create a model. As new input data is introduced to the trained ML algorithm, it uses the developed model to make a prediction.

Deep learning uses artificial neural networks to perform sophisticated computations on large amounts of data.It is a type of machine learning that works based on the structure and function of the human brain. There exist some of the traditional ML-based techniques in the literature aiding lung cancer detection and classification, for example, Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors(KNN). These techniques perform manual feature extraction, and then the classifier is trained using extracted features.

A k-Nearest Neighbors (k-NN) algorithm is first developed to predict lung cancer in its early stage. As the feature selection algorithm can affect the performance of the KNN model, a genetic algorithm (GA) is utilized to optimize the model used to predict. In the prediction of lung cancer, the Random Forest Algorithm showed improved accuracy compared with other methods. This predictive model will help health professionals in predicting lung cancer at an early stage.Support Vector Machine(SVM) classifier is a supervised machine learning algorithm used as a tool for data classification with advantages in handling data with high dimensionality and a small sample size. The performance of the SVM is observed for each feature as input.

Hence, a lung cancer detection system that employs Image Processing Techniques is used to detect the presence of lung cancer in CT- images. The prediction accuracy of this data fetched decent numbers close to ninety-six percentage

II LITERATURE REVIEW:

Bhatia at al- To detect lung cancer from CT scans using deep residual learning. The feature set is fed in to multiple classifiers, and Random forest, and the individual predictions are ensemble to predict the likelihood of a CT scan being cancerous. The accuracy achieved is 84% on LIDC-IRDI outperforming previous attempts[1].Nithila and Kumar-Accurate classification and prediction of lung cancer using technology that is enabled by machine learning and image processing. For the classification, ANN, KNN, and RF are some of the machine learning techniques that were used. It is found that the ANN model is producing more accurate results for predicting lung cancer[2].Lakshmanaprabu et al-created OODN (Optimal Deep Neural Network) by lowering the number of characteristics in lung CT scans and comparing it to other classification algorithms. The ODNN is applied to CT images and then, optimized using Modified Gravitational Search Algorithm (MGSA) for identify the lung cancer classification The comparative results show that the proposed classifier gives the sensitivity of 96.2%, specificity of 94.2% and accuracy of 94.56%.[3]

M.F.Abdullah et al-The categorization of lung cancer stages from CT scan images using image processing and k-Nearest Neighbor and results show that the KNN method has a high accuracy[4].Mr.Vijay et al- MATLAB have been used through every procedures made and process such as image pre-processing, segmentation and featureextraction have been discussed in detail to get the more accurate results[5].Anjali Kulkarni-Classify the stages of lung cancer, image processing technique is developed.In this work, new algorithm is developed using image processing technique to detect the cancer at early stage with more accuracy[6].P. M. Shakeel- Applying deep learning instantaneously trained neural network for predicting lung cancer. Eventually, the system is examined by the efficiency of the system using MATLAB based simulation results. The system ensures that 98.42% of accuracy with minimum classification error 0.038[7].

III METHEDOLOGY

The methodology adopted in this paper is carried out in three different methods, shown in Fig. 1. This study's problem is diagnosing whether the patient has cancer in the early stage. As is clear from the research purpose, the target variable is defined as discrete, so we need to use classification algorithms to identify the target variable.

According to Fig. 1, the first method comprises one fundamental building phase called image classification. It means, in this method, the raw CT images were given to CNN followed by ANN without any pre-processing(Raw CT images went through the third phase – the blue rectangular – immediately). The second method includes three primary building phases: image pre-processing, image segmentation, and image classification (According to Fig. 1, Raw CT images went through first (the yellow rectangular), second (the green rectangular), and third (the blue rectangular) phases, respectively). Finally, the third method comprises seven fundamental phases: image pre-processing, image segmentation, image feature extraction, building a numerical dataset, dimensional reduction, feature selection, and Although these methods have fundamental phases in common, they are entirely different methods implemented on the same lung CT scan images.

The pre-processing image phase of the study itself is composed of two parts: image resizing and image Denoising. Initially, raw lung CT images are resized, and subsequently, the median filter is applied to denoise them. The watershed segmentation algorithm identified the most critical objects in the CT images in the image segmentation phase to make the following steps more reliable. In the image classification phase, raw CT images in method one and segmented CT images in method two are used as input. The CNN and ANN algorithms are applied to the CT images to classify whether the images belong to a cancerous or noncancerous patient.

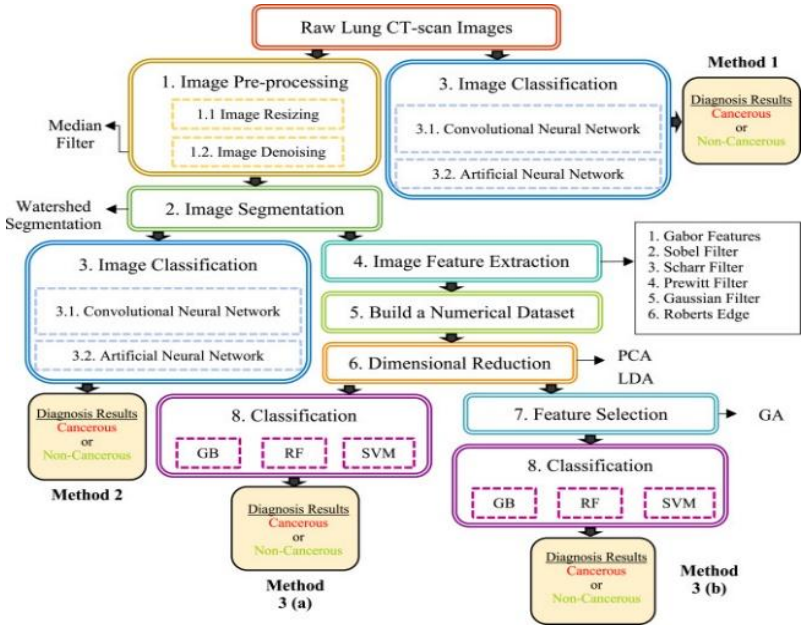


Fig. 1. The proposed framework, in comparison to other methods

So far, the fundamental phases that are implemented in methods one and two are described. In the third method, the image feature extraction phase is applied to segmented lung CT images to extract possible numerical features from each image's pixels. The extracted statistical data for each image is stored in a dataset, so building a numerical dataset phase is completed. We obtained a vast dataset as we extracted any possible mathematical features from the CT images to improve diagnosing cancerous patients from noncancerous o
In part one, classification algorithms – GB, RF, and SVM – are applied to both LDA and PCA datasets separately to classify the data into two groups based on extracted feature.

The same classification algorithms are used to classify the data into two groups

based on the selected features. In the following subsections, we will describe each fundamental phase in detail.

3.1. Image pre-processing

The term pre-processing belongs to a series of tasks needed for enhancing the quality of raw images, increasing the performance of the subsequent phases such as image segmentation, image feature extraction, and classification. The primary objectives in this study phase are to apply image resizing and denoising.
3.1.1. Image resizing
In the image resizing step, pixels are either added to the image or removed. Since medical images have many details that may be effective, no pixel removal is performed in the current research. On the other hand, the CNN algorithm's input images should preferably be in square sizes for a better diagnostic process. Therefore, all images used in this study have dimensions of 512 × 512 pixels.
3.1.2. Image denoising
Image denoising is a process of applying filter(s) to reduce image noise. It should be noted that CT images have the lowest noise level among the medical images, and it is practically unnecessary to perform this step. In any case, to prevent image distortion, the median filter is applied to cancel any possible noise of lung CT.

3.2. Image segmentation

The image segmentation tries to help the algorithm to diagnose as best it can by removing unnecessary parts. In fact, at this point, the algorithm can only focus on the lung region, which will improve the classification performance. Watershed segmentation is used in the suggested model [32]. The watershed technique is utilized when segmenting complicated pictures since basic thresholding and contour detection will not produce accurate results. The watershed method is built on capturing specific background and foreground information. Markers are then used to run watersheds and determine the precise borders. Markers can be defined by users, e.g., manually, or defined by some algorithms, e.g., thresholding operation — we used thresholding operation in our analysis.

3.3. Image classification

Image classification is the primary domain in which deep neural networks play the most critical role in medical image analysis. The image classification accepts the given input images and produces output classification to identify whether the disease is present [35]. The image classification phase is composed of two parts: CNN and ANN.

3.4. Image feature extraction

The following phases are performed in the third proposed method. After the raw CT scan images are processed and segmented, several numerical features are extracted from the images in the image feature extraction phase. Each feature will be obtained from each pixel in a single image and then stored in a dataset.

3.5. Building a numerical dataset

In the previous step, many features were extracted from each pixel in a CT image. For example, an image with 256 × 256 dimensions has 65,536 pixels, so if we extract 40 features from each and store them in a dataset, we will have 1 row for the image, and 40 × 65,536 columns. On the other hand, adding a target column should not be forgotten to determine whether the input image belonged to a cancerous patient or a noncancerous one. This procedure is continued until all the images' features are extracted and stored in a dataset.

3.6. Dimensional reduction

In the previous phase, an extensive dataset consisting of many features was obtained. However, implementing classification on this extensive dataset is time-consuming and not efficient. Implementing dimensional reduction algorithms on large data is one of the most critical steps. PCA and LDA are two-dimensional reduction algorithms used in this paper.

3.7. Feature selection

Feature selection methods have become an unavoidable part of the machine learning process to deal with high-dimensional data. Feature selection can identify related features and eliminate unrelated and repetitive ones to observe a subset of attributes that best describe the problem.

The first goal of the proposed feature selection method is to reach the same accuracy rate as the exclusive features. The second goal is to improve the accuracy rate. Here, gathering extensive information on the features costs too much, both in time and money, and new information is wasted in classifying and diagnosis. Reducing the dimension in terms of the number of features is recommended to get a better response and find a better correlation between the features and the outcomes.

A GA is a technique to select the best features. This technique generates a binary random vector consisting of the features using Eq. (1).

Vector(sj):sj=Yi;Yi;

Yi ={1 if Vectorsj contains feature i(1)

0 otherwise

An objective function based on the misclassification performance criterion is defined for any selected combination of the features. This objective function is a penalty function that should be minimized to find the best features. Here, the misclassification rate is simple and is obtained using Eq. (2)

$$mcr=\sum aij-[\sum aij;(i=j)]\sum aij/;i,j=1,2,\ldots,m \qquad (2)$$

The number of classification targets is the number of cases, the target is classified as the target using the classification method. The elements that construct the matrix in (3) form the so-called confusion matrix that depends on the problem as well as the dataset.

a11...a1m

: :

am1…amm

Now, the objective function to be minimized is a weighted sum of the mcr and nf (number of selected features) defined as:
$MinZ=w1*mcr+w2*nf$ (4)
Dividing both sides of Eq. (4) by w1, we will have:
$MinZ=mcr+w2w1/*nf.$ (5)
Assuming w2w1/=W, the objective function becomes:
$MinZ=mcr+W*nf.$ (6)

Now, W is defined as
$W\propto mcr \rightarrow W=\beta*mcr \rightarrow MinZ=mcr+\beta*mcr*nf$(7)

Finally, the objective function will be:
$MinZ=mcr(1+\beta*nf),$ (8)
where β is defined as a penalty for having an additional feature ($0\leq\beta\leq1$). Using this objective function, the GA finds the best combination of the features with the minimum number of features that minimize both the cost and the misclassification rate. Here, the stopping criterion to end the iterations in GA is chosen to be a predefined number of iterations

IV EXPERIMENTAL EVALUATION

This section discusses the way the data is collected, the implementation results of the proposed three methods on the data, and the analysis of the results.

A.DATA COLLECTION

KAGGLE DATA SET
Open Images 2019 - Object Detection
Computer vision has advanced considerably but is still challenged in matching the precision of human perception.Open Images is a collaborative release of ~9 million images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. This uniquely large and diverse dataset is designed to spur state of the art advances in analyzing and understanding images.This year's Open Images V5 release enabled the second Open Images Challenge to include the following 3 tracks:Object detection track for detecting bounding boxes around object instances, relaunched from 2018.
Visual relationship detection track for detecting pairs of objects in particular relations, also relaunched from 2018.Instance segmentation track for segmenting masks of objects in images, brand new for 2019.Google AI hopes that having a single dataset with unified annotations for image classification, object detection, visual relationship detection, and instance segmentation will stimulate progress towards genuine scene understanding.
The total number of CT scan images used in this paper is 364, of which 238 are cancerous images, and the rest (126) belong to noncancerous images. All these images are collected with the help of a pulmonologist to skip any probable error in classifying images. Some of the CT images of the lungs acquired from the hospital database are shown in Fig. 2.

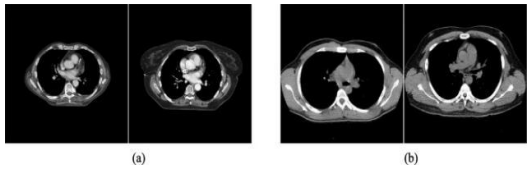


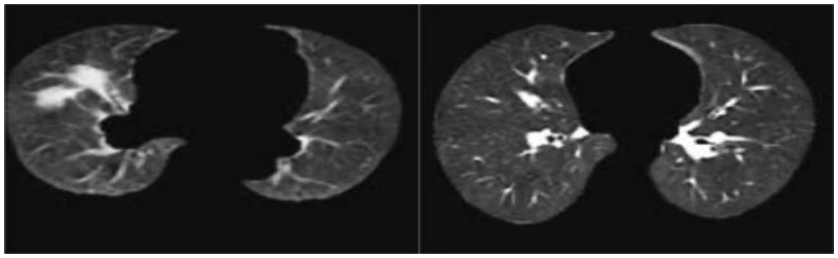
Fig. 2. (a) medical CT images of lung cancer patients, (b) medical CT images of lung patients other than lung cancer.

B. The implementation results of the first method

As seen in Fig. 2, applying any pre-processing or segmentation on the raw images is not needed when implementing CNN and ANN. In other words, the raw lung CT scan images are fed as inputs to the CNN and ANN architecture in the first method. Several different structures are evaluated to obtain the best structure to distinguish cancerous CT images from noncancerous ones. However, the best structure consists of three convolution layers with 64, 64, and 128 feature maps, respectively, in the first, second, and third layers in the convolutional neural network section. The artificial neural network also contains two hidden layers, each containing 128 neurons. This study uses max pooling with dimensions of 2×2 after each convolution layer to maintain the feature maps.

C.The implementation results of the second method

As shown in Fig. 1, image pre-processing and segmentation are used in the second method before running the CNN and the ANN algorithms. This method aims to determine whether performing image pre-processing and image segmentation affects the performance of the first method. In the first step of image pre-processing, all the image sizes are set to 512 × 512 so that all the pixels in an image remain intact. The next step applies the median filter to remove any possible noise of resized lung CT images. Fig. 4 shows the image before and after the mediafilter is performed on both the cancerous and noncancerous lung CT scans. As seen in this figure, the images after the filter are not much different from the images without the filter, which is a characteristic of CT scan images.



The image on the right shows lung cancer after applying the mask, and the image on the left shows noncancerous lung after using the mask.

D. The implementation results of the third method

As masks are placed on filtered images in the image segmentation phase, images' unnecessary parts are covered. Therefore, to reduce the number of columns, the segmented images are resized to 256 × 256. By these dimensions for each segmented image, 2,621,440 data are generated for each image when the features are extracted.The filters/features are applied alone to each image's pixel and store each pixel's calculations in a data frame. To create a numeric dataset, the number of pixels in each image is 65,536, and the total number of filters executed on each pixel is 40. The data values of the original image's pixels and the label of being cancerous (1) or noncancerous (0) are also given in this dataset. Thus, each picture contains 1 row and 40 × 65,536 feature columns plus a label column and the original pixel's value.

V CONCLUSION:

Lung cancer is one of the deadliest types of the disease, claiming the lives of approximately one million people each year.This paper discusses about the automatic Cancer detection and classification of CT Images using deep learning algorithm. The CNN algorithm and Googlenet were chosen for detecting the cancer regions and classifying them into normal and abnormal. For the CNN algorithm implementation, a deep convolution network architecture called VGG-16 was used as base network. The proposed algorithm efficiently identifies the Lung Cancer. Given the current state of affairs in medicine, it is critical that lung nodule identification be performed on chest CT scans. As a result, the use of CAD systems is crucial for the early detection of lung cancer. Image processing is a necessary activity that is employed in a wide range of economic domains. It is used in X-ray imaging of the lungs to find areas of the body that have developed malignant growths. Image processing techniques such as noise reduction, feature extraction, identification of damaged regions, and maybe comparison with data on the medical history of lung cancer are used to locate sections of the lung that have been affected by cancer. This study demonstrates accurate lung cancer classification and prediction using technologies enabled by machine learning and image processing. To begin, photographs must be collected. Following that, the images are preprocessed using a geometric mean filter. This eventually leads to an increase in image quality. The K-means approach is then used to segment the images. This segmentation makes it easier to identify the region of interest. Following that, machine learning-based categorization algorithms are used. ANN predicts lung cancer with more accuracy. This research will help to increase the accuracy of lung cancer detection systems that use strong classification and prediction techniques. This study brings cutting-edge images based on machine learning techniques for implementation purposes.

REFERENCES

[1] Q. Abbas and M. M. Q. A. Yasin, ''Lungs cancer detection using convolutional neural network,'' Int. J. Recent Adv. Multidisciplinary Topics, vol. 3,no. 4, pp. 90–92, 2022.

[2] A. Khan, ''Identification of lung cancer using convolutional neural networksbased classification,'' Turkish J. Comput. Math. Educ. (TURCOMAT), vol. 12, no. 10, pp. 192–203, 2021.

[3] C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, ''Earlystage lung cancer prediction using various machine learning techniques,'' in Proc. 4th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Nov. 2020, pp. 1285–1292.

[4] Y. Xie, W.-Y. Meng, R.-Z. Li, Y.-W. Wang, X. Qian, C. Chan, Z.-F. Yu,X.-X. Fan, H.-D. Pan, C. Xie, Q.-B. Wu, P.-Y. Yan, L. Liu, Y.-J. Tang,X.-J. Yao, M.-F.Wang, and E. L.-H. Leung, ''Early lung cancer diagnosticbiomarker discovery by machine learning methods,'' Translational Oncol.,vol. 14, no. 1, Jan. 2021, Art. no. 100907.

[5] X. Liu, K.-W. Li, R. Yang, and L.-S. Geng, ''Review of deep learning basedautomatic segmentation for lung cancer radiotherapy,'' Frontiers Oncol.,vol. 11, p. 2599, Jul. 2021.

[6] R. Mahum, S. U. Rehman, O. D. Okon, A. Alabrah, T. Meraj, andH. T. Rauf, ''A novel hybrid approach based on deep CNN to detectglaucoma using fundus imaging,'' Electronics, vol. 11, no. 1,p. 26,Dec. 2021.
[7] R. Mahum, ''A novel framework for potato leaf disease detection usingan efficient deep learning model,'' Hum. Ecol. Risk Assessment, An Int. J.,vol. 29, no. 2, pp. 303–326, 2022
[8] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.
[9] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.
[10] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.
[11] C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.
[12] Nagarajan C., Neelakrishnan G., Akila P., Fathima U., Sneha S. "Performance Analysis and Implementation of 89C51 Controller Based Solar Tracking System with Boost Converter" Journal of VLSI Design Tools & Technology. 2022; 12(2): 34–41p.
[13] C. Nagarajan, G.Neelakrishnan, R. Janani, S.Maithili, G. Ramya "Investigation on Fault Analysis for Power Transformers Using Adaptive Differential Relay" Asian Journal of Electrical Science, Vol.11 No.1, pp: 1-8, 2022.
[14] G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash "Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756
[15] G.Neelakrishnan, S.N.Pruthika, P.T.Shalini, S.Soniya, "Perfromance Investigation of T-Source Inverter fed with Solar Cell" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:744-749
[16] C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007
[17] M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022
[18] M Suganthi, N Ramesh, CT Sivakumar, K Vidhya, "Physiochemical Analysis of Ground Water used for Domestic needs in the Area of Perundurai in Erode District", International Research Journal of Multidisciplinary Technovation, pp: 630-635, 2019