International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

A NOVEL APPROCH FOR DETECTING LUNG CANCER USING IMAGE PROCESSING

N.V.Chaudhari[1], Dr. A.V.Malviya[2]

[1]ME II Student & Department of Electronics and Telecommunication department, Sipna College of engineering and technology, Amravati, India

[2]Associate Professor& Department of Electronics and Telecommunication department, Sipna College of engineering and technology, Amravati, India

---

Abstract - Rapid identification is a difficult issue for researchers since noise signals are mixed in with original signals during the picture capture process, causing the cancer picture quality to deteriorate and resulting in poor efficiency. So, as to avoid this image processing techniques become popular in a variety of medical fields for image analysis in earlier diagnosis and treatment stages, especially in cancer tumors, where time is essential for detecting abnormality issues in target images. Despite computed tomography is the most widely used imaging technique in medicine; it can be difficult for doctors to precisely identify and diagnose cancer from CT pictures. As a result, computer-aided diagnosis can be beneficial to doctors in accurately identifying cancer cells. The methodology is examined by applying image processing techniques and machine learning classification.

Keywords –CT scan, Image processing, Lung cancer, Feature extraction, Classification, Machine learning, etc.

I.INTRODUCTION

Cancer is the deadliest cancer among all other types of cancer. Lung cancer kills more people each year than any other type of cancer, including breast cancer, since it is difficult to detect and identify. Lung cancer, brain cancer, and prostate cancer are all examples of cancers. According to data from Global Burden Cancer, there were 2,093,876 cancer cases worldwide in 2018, with lung cancer taking first place with a ratio of 11.6%. While 1,761,007 people died from cancer, the leading cause of death was lung cancer, which accounted for 18.4% of all deaths [2].

Cancer Type	No. of cases	No. of deaths
Lung	2,093,876	1,761,007
Breast	2,088,849	626,679
Prostate	1,276,106	358,989
Colon	1,096,601	551,269

Table 1: Number of cancer cases and deaths (2018) [10]

The table shows the incidence and mortality of all cancers by type of cancer. Compared to other cancers, lung cancer mortality is greater. Lung cancer can be divided into two main groups, non-small cell lung cancer, and small cell lung cancer. These assigned lung cancer types depend on their cellular characteristics. As for the stages, in general, there are four stages of lung cancer; I through IV. [11]

The method of early cancer identification is crucial to avoiding cancer cells from growing and spreading. The SVM classification technique has been used in previous studies to analyze lung cancer. [1] The goal of this study is to use an ensemble boosting approach with texture and statistical feature and detection using image morphology to locate the early stage of lung cancer and get a more accurate result. The system contains mainly four stages: Input dataset, Pre-processing, and segmentation, Feature extraction, and Classification.

Initially, input CT scan images are provided as input, pre-processing takes place by use of the median filter. The classification of images will be carried out using a machine learning classification model which finally provides the output classified image into 'cancer' or 'healthy'. Computer-aided diagnosis models will be trained and tested for identifying lung cancer. MATLAB software is used for simulation purposes and classifier operation is carried out using a machine learning toolbox.

II.LITERATURE SURVEY

The developments in machine learning applied to medical imaging, as well as alternative segmentation and classification methods used for lung cancer diagnosis are summarized in this chapter. The automatic detection and classification of distinct stages of lung tumors are critical for lung cancer early diagnosis. Several image processing algorithms allow both automatic and early identification. Several researchers have developed and studied several strategies and methodologies for medical imaging in current history. The purpose of the pre-processing stage is to reduce the amount of noise in these images. Different filtering approaches, such as median filtering [1], wiener filtering [19], Gabor filtering [8][4], and others, have been proposed in the literature to remove these disturbances. The second stage of the image processing system is lung region segmentation. It is the process of dividing a pre-processed CT image into different regions to separate the pixel values about lung tissue in ct Lung image segmentation methods. As a result of the surrounding anatomy. A vast number of image segmentation have been used in various ways. Various segmentation approaches such as region growing[19], marker controlled watershed[8][6], superpixel segmentation [18], local binary pattern were used to extract ROI from the

images[9] To identify cancer nodules from an extracted lung images, a segmentation technique based on the Sobel edge detection approach is used in [19], however, Tariq et al. employed the gradient mean and variance-based approach to recover lung background because the gradient operator has large values for pixels at the foreground/background boundary.[20] Another stage of image processing is feature extraction. Different features of lung nodules, such as area, perimeter, etc. were extracted using different methodologies by the authors. The classification methodology used by different authors to classify the lung images as cancerous or noncancerous. Unsupervised and supervised classifications are the two types of classification. Supervised classification produces accurate outputs with labels, and this type of training is referred to as supervised classification. Supervised classification techniques include K-nearest neighbors, artificial neural networks, and others. Unsupervised learning is when data is collected automatically without the use of classes or labels. For example, K-means clustering, hierarchical clustering, and so forth. However, SVM outperforms all other classifiers.[7] Support vector machines are supervised learning models that are used to analyze data and recognize various patterns to classify them.[12].

III METHODOLOGY

Figure 1. Depicts the whole working process of the given approach. The proposed model, as indicated in the picture, consists of a sequence of processes that are detailed in the block diagram.

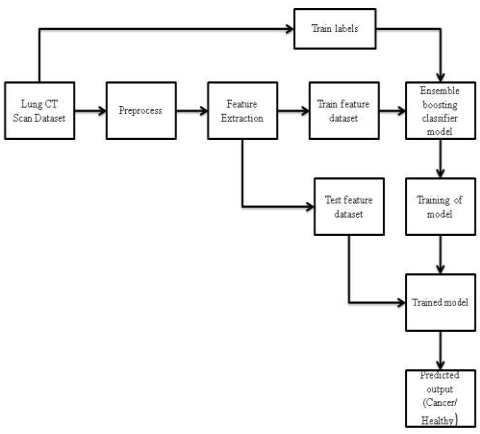

Fig 1: Block diagram

1. Lung CT scans Image Dataset

The proposed methodology is tested on a standard benchmark image database in a variety of ways. The design is based on a collection of lung cancer screening CT scans from the dataset. The Lung Image Database Consortium (LIDC) is being used in this research. A set of images from the Lung Image Database Consortium (LIDC) collection is used to simulate the described model. Following the collection of cancer and healthy lung images, the images are grouped into a train and test image dataset for classification and machine learning application.

Fig2: Lung CT scans

2. Pre-process:

Pre-processing is the process of removing noise from an image. Picture smoothing and image enhancement are two of the primary phases. Using filters to improve the image quality. Image pre-processing is done to improve the interpretability or interpretation of data in an image for display reasons or to provide better input for other computerized image processing processes. The input CT scan lung images are provided as input to the model. The CT scan lung images are fed into the model as input. The goal of the pre-processing step is to reduce noise and improve image quality. The median filter was used to remove noise. The median filter is a non-linear image processing approach for reducing salt and pepper noise. The image enhancement method improves the quality of a digital image, allowing it to be handled more efficiently. To improve the image, a contrast adjustment is applied.

3. Feature extraction:

Feature extraction, which turns input data into required features, is the most important stage of the process. Feature extraction refers to the process of extracting higher-level data from an image, such as color, shape, and texture. Human perception is greatly influenced by texture. Statistical texture techniques analyze the spatial distribution of grey values by producing local features at each location in the image and inferring a set of statistics from the distributions of the local features. This stage extracts image features that will be used to classify CT scan images. The texture features based on wavelets are extracted. The "Haar" wavelet is used to extract the features. A discrete wavelet transform decomposes an image into several sets, each of which has four coefficients: ca, ch, cv, and cd, of which the approximation coefficient is used in subsequent operations. The PCA i.e. Precision component analysis is applied to the approximation coefficient to reduce the dimensionality of the database set. Later, the haralick features, hu moments, and statistical features of the datasets are evaluated. The GLCM algorithm evaluates haralick features. GLCM shows how often each grey level occurs at a pixel positioned at a given geometric position relative to the other pixels as a function of the grey

level. Contrast, correlation, energy, and homogeneity are all taken into consideration. After extracting all features for the training dataset and test dataset images, features are further separated for the train and feature dataset. The train feature dataset with labels are used for training of classifier model and test feature dataset are used for testing purpose.

4. Classification

The CT scan images are classified as normal or healthy at the classification step. The Ensemble boosting classifier model will be employed in our methodology to detect lung cancer in CT images. Ensemble classifiers are supervised learning models that consist of a group of classifiers that combine their decisions in almost the same way to categorize incoming images, resulting in a strong classifier from weak classifiers.

Ensemble Classifier Model with Boosting

In this methodology ensemble boosting classifier is used for the classification of lung cancer from CT images EB classifier analyzed the data and classify them according to the patterns. The EB classifier builds a model by using a training dataset and categorizes it into two classes. The EB algorithm then assigns new examples of testing datasets to one of the two classes. EB classifier thus finds the best hyperplane that separates the two groups and thus classifies the lung CT images. Images will be classified using a machine learning classification model, with the outcome being a classified image labeled as either 'Cancer' or 'Healthy.' If a cancer image is later discovered, it will only display and detect the abnormal lung nodule portion. Lung cancer detection and diagnosis models have been trained and tested using computer-aided diagnostics. Boosting is a sequential method in which each subsequent model attempts to correct the errors of the previous model.

5. Detection :

The detecting stage begins once the images have been classified. Morphological operations were utilized to detect cancerous regions in the lungs. The opening procedure smoothes the image's contour, followed by the closing operation, which fills in the holes and tiny gaps between the pixels. Finally, a dilation operation is conducted to increase the lung region's size, and an erosion operation is performed to reduce the lung region's size. Finally, a border clear operation is performed to remove the entire border and produce the lung region where the cancer cells or nodules are presented. The detection of cancer in the lung region was accomplished in this manner.

IV. IMPLEMENTATION

The implementation of system is shown in fig 5.1 in which different tabs are created. The load database tab offers a window where CT images from the train dataset folder are provide as input to the system. After the database has been loaded, the features of the images are analysed, and the model is trained. . Once the model has been properly trained, it is kept for testing purposes. The CT image from the test dataset folder is imported into the load input tab for testing, after which the image is pre-processed, features are retrieved, and the saved trained model is loaded to predict the output where it provides the predicted output as 'Cancer 'or 'Healthy'. . All of the image analysis programs were written in MATLAB, and a graphical user interface was constructed to aid in the research.

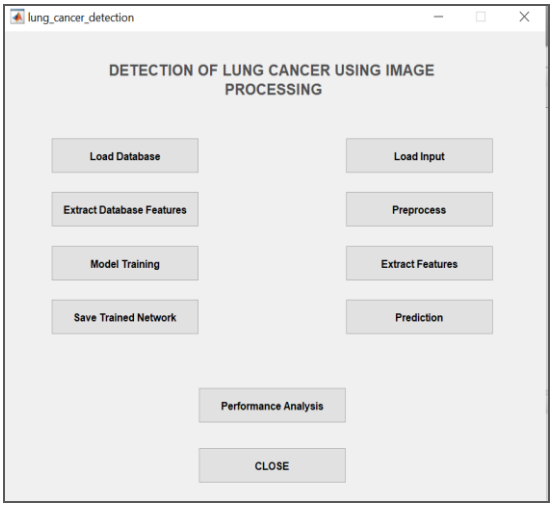

Fig 3: Guide window.

Fig 4 depicts the preprocessed image. Fig 4(b) generated after applying the median filter and contrast adjustment. The accuracy, error rate, sensitivity, specificity, and f-score of ensemble boosting classification are also shown. Figures 5 show feature extraction and testing assessment time, respectively. Fig 8 depicts the lung region where cancer cells are presented. The classification result of the ensemble boosting algorithm model, which classifies the images and provides output as "cancer" or "healthy" is shown in Fig 7. In the example below, the given input is classified as "cancer."

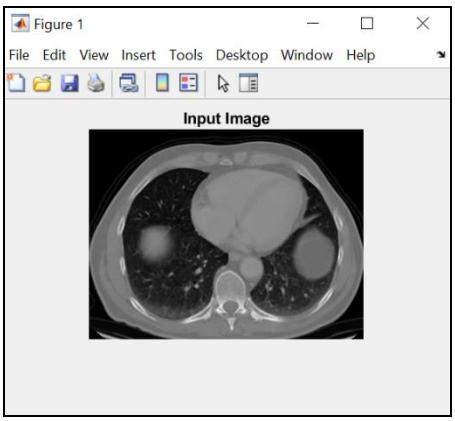

Fig 4 (a): Input image

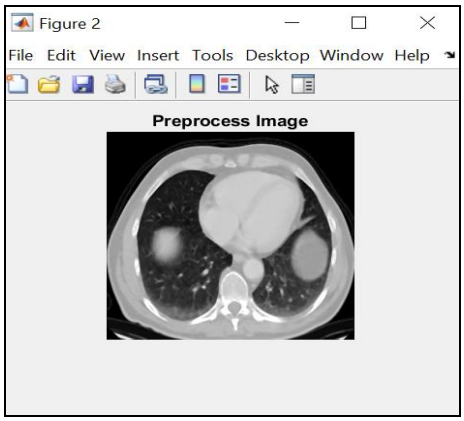

Fig 4(b): Pre-process

Fig 5: Feature extraction & Evaluation time

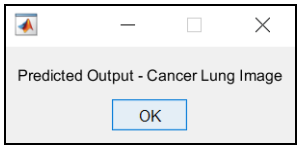

Fig 6: Predicted output

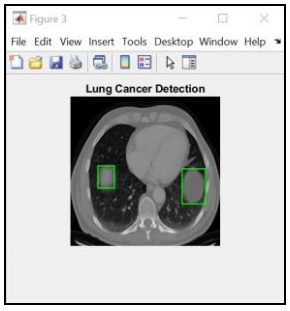

Fig 7: Detection of lung cancer

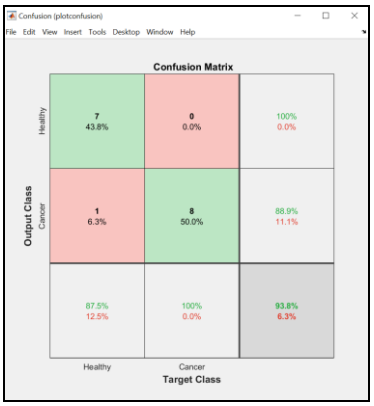

Fig 8: Confusion matrix of system analysis

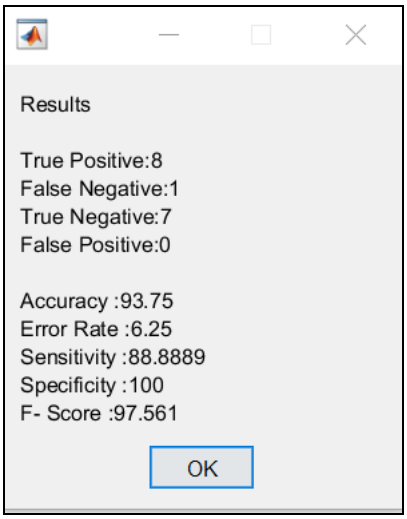

Fig 9: Result Panel

The confusion matrix above depicts the system's performance. A total of 16 samples are analyzed. The matrix shows that 15 samples are accurately classified and 1 sample is incorrectly classified out of 16 samples. The result panel above depicts the system parameter such as accuracy, specificity, sensitivity and F- score. In this way, the system is implemented.

V. RESULT

In this section, the experimental results and their implications are discussed. The binary classification is used to create a confusion matrix. The results of classification are summarized using a confusion matrix. A confusion matrix shows the actual and expected results of the classifier, as well as the classifier's performance. ROC, or Receiver Operating Curve, is used to evaluate the classifier's performance. Based on this matrix, the parameters such as accuracy, error rate, true positive rate (TPR), false-positive rate (FPR), truth negative rate (TNR), and false-negative rate (FNR) can be determined. The definitions of terminology are depicted in table 3. The system's purpose is to achieve a level of accuracy in detecting lung cancer. Although there are a variety of strategies for appropriately presenting information to aid diagnosis, the findings piqued the researchers' interest. This project's purpose was to divide lung scans into two groups: healthy and cancerous. Cross-validation is a technique for evaluating performance.

known	Predicted Label	
Label	**Positive**	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Table 2: Confusion Matrix

Sr.No	Terminology	Definitions
1	TP (True positive)	Outcome in which the model detects the positive class accurately.
2	FN (False Negative)	Outcome in which model detects negative class accurately.
3	TN (True Negative)	Outcome in which model detects positive class accurately.
4	FP (False Positive)	Outcome in which model detects negative class accurately.
5	Accuracy	Number of correct predictions overall
6	Error Rate	It is the misclassification rate i.e. the number of incorrect outcomes.
7	Sensitivity	The total number of correct positives outcomes divide by the number of positives(TP/P)
8	Specificity	Total number of incorrect outcomes divided by the number of negatives (TN/N)
9	F-score	Precision and recall harmonic mean

Table 3: Basic terminologies of confusion matrix

Table 4 shows the comparison of different methods in terms of parameter such as accuracy specificity and sensitivity.[6] With a score of 77, the linear regression model has been shown to be less accurate than the other models. The DNN model, on the other hand, achieves an accuracy of 87.65, which is deemed moderate. The KNN, RBF, ANN, and MPL, on the other hand, yield 91.00, 84.00, 86.00, and 82.00, respectively. The sensitivity of MPL models is lower than that of other models.

The specificity of DNN was 89.67, which was higher than the specificities of the other models. The linear model has the lowest specificity of 36.00 when compared to the other models, and the RBF model has the second lowest specificity of 36.00. On the other hand, ANN models predicted moderate specificity. The proposed model has a higher sensitivity and accuracy than the prior methodologies, as demonstrated in table 4.

[1] Methods	[2] Accuracy	[3] Sensitivity	[4] Specificity
[5] Proposed	[6] 93.75	[7] 88.88	[8] 100
[9] MLP	[10] 82.00	[11] 77.00	[12] 72.00
[13] RBF	[14] 84.00	[15] 86.00	[16] 54.00
[17] LINEAR	[18] 77.00	[19] 89.00	[20] 36.00
[21] ANN	[22] 86.00	[23] 87.00	[24] 79.00
[25] KNN	[26] 91.00	[27] 90.00	[28] 83.00
[29] DNN	[30] 87.65	[31] 82.43	[32] 89.67

Table4: analysis of various methods. [6]

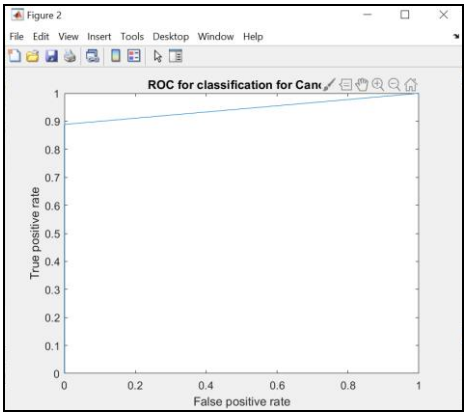

Fig 10: ROC of Classification

The system parameters that are examined to determine the system performance are displayed in the result pane depicts in fi.9. The classification accuracy is 93.75, error rate 6.5, specificity 100, sensitivity 88.88, and f- score 97.56. Also, the detection of lung cancer region and the ROC curve of the classification model which projected 93.75 accuracy depicts in fig7 and fig 10.respctvely.

VI. CONCLUSION

Lung cancer is one of the most common cancers that lead to mortality. It's tough to spot because symptoms don't develop until late in the disease's progression. Early discovery and treatment of the condition, on the other hand, can reduce the death rate and probability. CT images were taken for the implementation of the project. These images are less noise as compared to X-ray and MRI images. An image improvement technique is developed. The CT images from the LIDC dataset pre-processed by using median filter and contrast adjustment and segmentation of the images carried out by morphological operations. The median filtering technique was effective in eliminating salt and pepper noise without blurring the image. The wavelet-based feature extraction approach followed by PCA was used to extract important features. The haralick features are evaluated by the GLCM algorithm also the hu moments and statistical features of the dataset are evaluated. and the classification carried out with ensemble boosting algorithm correctly classifies cancer and healthy images and the output is predicted. MATLAB was used to write all programs for image analysis and a graphical user interface was created to help assist with analysis. The results obtained are satisfactory for identifying lung cancer in the manner indicated above. By employing this technique on more images in the procedure, the accuracy can be improved.

REFERENCES

1] Nadkarni, N.S. and Borkar, S., 2019, April. Detection of lung cancer in CT Images using image processing. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 863-866). IEEE.

2]GLOBOCAN 2012, International Agency for Research on Cancer, World Health Organization. http://globocan.iarc.fr/

3]Makaju, S., Prasad, P.W.C., Alsadoon, A., Singh, A.K. and Elchouemi, A., 2018. Lung cancer detection using CT scan images. Procedia Computer Science, 125, pp.107-114.

4]Abdillah, B., Bustamam, A. and Sarwinda, D., 2017, October. Image processing based detection of lung cancer on CT scan images. In Journal of Physics: Conference Series (Vol. 893, No. 1, p. 012063). IOP Publishing.

5]Santosh Singh, Ritu Vijay & Yogesh Singh. (2017). Segmentation of Lung Nodule Using Image Processing Techniques from CT Images. Special Conference Issue: National Conference on Cloud Computing & Big Data.

6] Jayaraj, D. and Sathiamoorthy, S., 2019, November. Random Forest-based Classification Model for Lung Cancer Prediction on Computer Tomography Images. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 100-104). IEEE.

7]Sivakumar, S. and Chandrasekar, C., 2013. Lung nodule detection using fuzzy clustering and support vector machines. International Journal of Engineering and Technology, 5(1), pp.179-185.

8]]Al-Tarawneh, M.S., 2012. Lung cancer detection using image processing techniques. Leonardo Electronic Journal of Practices and Technologies, 11(21), pp.147-58.

9]Miah, M.B.A. and Yousuf, M.A., 2015, May. Detection of lung cancer from CT image using image processing and neural network. In 2015 International conference on electrical engineering and information communication technology (ICEEICT) (pp. 1-6). ieee.

10]Manju, B.R., Athira, V. and Rajendran, A., 2021, January. Efficient multi-level lung cancer prediction model using support vector machine classifier. In IOP Conference Series: Materials Science and Engineering (Vol. 1012, No. 1, p. 012034). IOP Publishing

11]Neelima, S. and Asuntha, 2016. Image Processing Used for Lung Cancer Detection in Medical Imaging. Journal of Chemical and Pharmaceutical Research, 4(8), pp.1044-1049.

12] Sangeetha, M., and Mythili, S., 2021, January. Detection and Characterization of Lung Tumor by Using Convolution Neural Network. In Journal of Physics: Conference Series (Vol. 1717, No. 1, p. 012004). IOP Publishing

13]Gajdhane, V.A. and Deshpande, L.M., 2014. Detection of lung cancer stages on CT scan images by using various image processing techniques. IOSR Journal of Computer Engineering (IOSR-JCE), 16(5), pp.28-35.

14]M.Priya, A.Nagarajan, "Automatic Detection and Classification of LungCarcinoma using Image Processing Techniques in Lab VIEW." International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878,Volume-8 Issue-5, January 2020

15]Sevani, A., Modi, H., Patel, S. and Patel, H., 2018. Implementation of image processing techniques for

identifying different stages of lung cancer. Int J Appl Eng Res, 13(8), pp.6493-6499.

16]Khehrah, N., Farid, M.S., Bilal, S. and Khan, M.H., 2020. Lung Nodule Detection in CT Images Using Statistical and Shape-Based Features. Journal of Imaging, 6(2), p.6.

17]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texturalfeatures for image classification," IEEE Transactions on Systems,Man and Cybernetics, vol. 3, no. 6, pp. 610–621, 1973.

18]Alam, J., Alam, S. and Hossan, A., 2018, February. Multi-stage lung cancer detection and prediction using multi-class SVM classifier. In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

19]Sharma, D. and Jindal, G., 2011. Identifying lung cancer using image processing techniques. In International Conference on Computational Techniques and Artificial Intelligence (ICCTAI) (Vol. 17, pp. 872-880).

20]Taher, F. and Sammouda, R., 2011, February. Lung cancer detection by using artificial neural networks and fuzzy clustering methods. In 2011 IEEE GCC conference and exhibition (GCC) (pp. 295-298). IEEE.