A HYBRID FEATURE FUSED DEEP LEARNING APPROACH FOR LUNG CANCER CLASSIFICATION

A. Vettriselvi1*, D. Divyarupakala2, N. Gnanambigai3, P. Dinadayalan4

1Research Scholar, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India;
2Assistant Professor, CSE, St. Joseph's Institute of Technology
3Department of Computer Science, Indira Gandhi College of Arts and Science, Puducherry -605009, India,
4Department of Computer Sri, KanchiMamunivar Centre for Postgraduate Studies, Puducherry - 605008, India

ABSTRACT

In today's world medical images plays a crucial role in treating lung cancer. Lung cancer is the riskiest cancer. It is difficult to cure in the advanced stages. The survival rate can be increased only by effective early detection. The previous work mainly focuses only on the semantic features, which does not hold the texture and edge information of images. In order to overcome this problem. Firstly, this article explores how to identify the relation between feature extraction and classification approach using both semantic and intensity information for classifying lung cancer images as Benign, malignant and normal. Secondly, the selected optimal features reduced using correlated intensity property-based techniques. With this, the ensemble learning based classification was performed on the combined data set Kaggle and Iraq lung cancer dataset using multi-class SVM. Hence, the semantic information is obtained through the Google net and customized CNN and intensity information extracted using local binary pattern (LBP) and Gray level co-occurrence matrix (GLCM). Finally, the proposed work also used canonical correlation fusion feature and ensemble learning for the lung cancer classification analysis which achieves 98% accuracy in just 418 milliseconds for testing set. Here, the combined features will be classified using ensemble learners to further optimize its results.

Keywords: deep learning, Google net, classification, customized CNN, feature extraction, semantic and intensity.

1. INTRODUCTION

Lung cancer is a disease that spread faster to other organs which is caused by the rapid cell growth in a body. There is different type of cancers like skin, breast, blood, bone and lung cancer. It is the second deadliest disease. At earlier stage cell spreads to smaller areas only but in later stages it blocks the passage completely. Therefore, early diagnosis will help to treat the patients more effectively. In today world, in medical field the lung problems can be diagnosed using different types of image modalities namely X-ray, ultrasound, CT, MRI and PET- CT. This article, proposed CT imaging for the classification of lung cancer. It helps to analyse completely the stages of lung cancer and region properties like size, shape and its depth.

This section, summarizes techniques that are used for the CT lung cancer classification and feature extraction.

In[1] this research lung cancer use different steps like enhancement, median filter segmentation for feature extraction. While in [2] the gray level co-occurrence matrix GLCM is proposed and it shows significant improvement. In [3] a 3D texture feature extraction and classification using GLCM and LBP based descriptors are used. In [4] this paper, a robust texture feature based on efficient local binary pattern descriptor issues that gives a significant improvement in texture ability. while in [5] Euclidean distance method is applied to classify the texture pattern by using LBP. Here the various LBP methods are analysed and the future of the LPB method is also pointed out [6].

In [7] author detect lung tumour in CT images using weighted gradient which increases the accuracy. In [8] LBP incorporated with InceptionResnetV2 model to improve diagnostic accuracy for lung and colon cancer with 99.98% accuracy. In [9] used LDNET on Luna 16 and Kaggle dataset for lung cancer classification with accuracy on 98% and 99%. In [10] a novel transfer learning lung-effnet used for lung cancer classification and attained 99% of accuracy on the test set. In [11] detect the location of the cancerous lung nodules using best feature extraction techniques and Fuzzy PSO algorithm also applied for selecting the best feature.

In [12] an optimal deep neural network and LDA used for extracting deep features from CT images and LDR for dimensionality reduction. while in [13] author proposed a lightweight multiple view angles based multi-section CNN architecture. It achieved the 93.18% classification accuracy. In [14] a method based on Google net architecture is used to minimize manual control and Max inference on image. The model highest accuracy estimated as 98%. In [15] the article, a pre-trained conventional neural network like Alex Net, Resnet and Google net are used for training network and CT image classification. In [16] introduces an automatic recognition method and author explained what is better about Google net and Alex net through its feature extraction and performance. Here [13] the pre-processed cancer images are analysed for lung cancer detection using Google net and Alex net. while in [14] transfer learning is used to read just Google net DNN which allows final layer of the DNN to evolve and isolates ROI to outperforms course with 94.38% accuracy.

2. RELATED WORKS

Abdollahi (2023) categorises lung cancer using Iraq dataset using Lenet-5 architecture [17]. Lenet-5 is a dense network able to reach an accuracy of 97.18% by joining the convolutional encoder and three fully linked layers as their components. Huang et al., (2024)

in this article, the Kaplan-Meier method and the Cox regression analysis were used to investigate patients with non-small lung cancer (NSCLC) for the study of lung cancer. Zhai et al., (2020) an MT-CNN architecture is obtained for detecting chest CT imaging and show the difference between benign and malignant nodules. The approach shows lowest false positive rate when compared to LUNA-16 and LIDC-IDRI has the greatest AUC and the lowest rate of false positives.

Veasey pre-trained 2-D convolutional feature extractors for identifying lung nodules using CNN and recurrent neural network from the data collected during NLSTX.To achieve high performance compared to that of a 3-D CNN. Lin and Li (2020) proposed Taguchi based CNN for the purpose of determining lung nodules, trachea, and bronchial malignancies are malignant or benign. Orthogonal matrices are determined for determining structural variables. This technique was so effective and attained an accuracy rate of 99.6%.

Mastouri et al., (2020) over the last three years the author gives brief description about deep learning features, nodules, and other related articles .From the findings it is proven that , CNN models have the ability to diagnose and detect lung nodules at early stage, with sensitivity rates increasing from 66% to 100% and false-positive rates increasing from 1to 15%.Paul et al., (2020) the Convolutional Neural Networks (CNNs) identify lung nodules and lung cancer in various examinations. Seven seeds were used to train three distinct CNN architectures, an ensemble model with an accuracy of 90.29% and AUC of 0.96 by yielding.

3. ARCHITECTURE

This paper proposed a hybrid feature fused based rear approach for lung cancer classification model using CT images. The flow of information in the proposed approach is presented in the flow chart in figure 1.

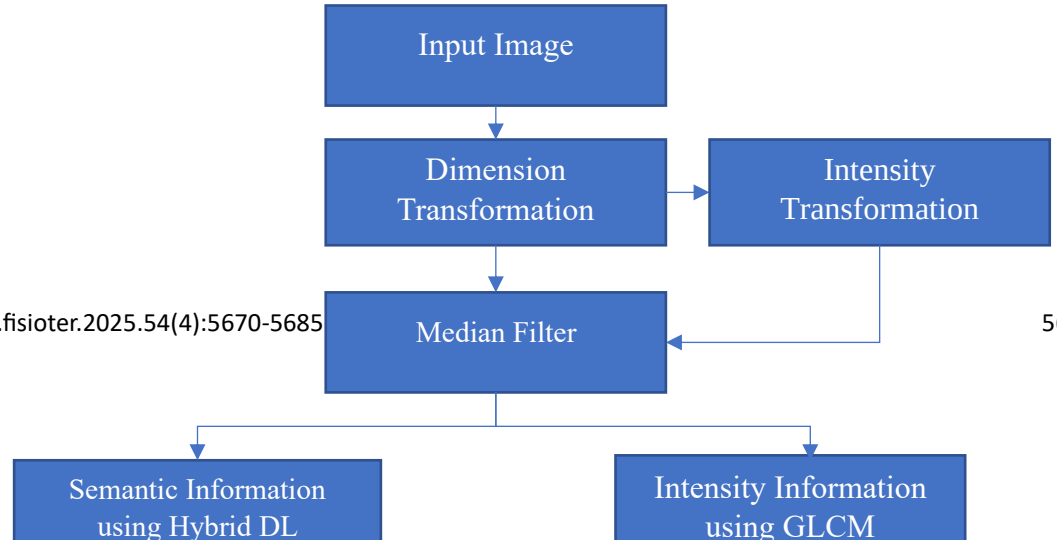

Figure 1. Data Flow Diagram

3.1 Input images

The combined data set Kaggle and Iraq lung cancer data set is used in this work. Here both individual and combined datasets are used for analysis. One of the special features of this dataset is that it has .jpg images from the DICOM images. Therefore, it helps to minimize the pre-processing step and can be used directly for Lung cancer classification process.

3.2 Image preparation

In this research, two different types of image sets done for analysis. Therefore, it has different dimensions and color properties.

In order to normalize the image sets the following Pseudocode is used. The steps in the data flow diagram are as follows:

step1: Load IQ-OTH And Kaggle CT chest dataset.

step 2: Combine both data sets into a single data set.

step 3: Perform pre-process to filter, resize and transformation in the images.

step 4: Split the data set as training. testing and validation in the ratio 0:7. 0:2.0:1.

step 5: Perform semantic based information using GLCM and LBP.

step 6: Perform intensity-based information using Google Net and Customized cnn.

step 7: Combine features for information reduction using canonical correlation fusion.

step 8: Perform final prediction from ensemble learning.

step 9: Evaluate the classifier performance

step 10: Re-train with the combined features and test the set.

3.3 Dimension transformation

The lung cancer classification was subjected to dimension transformation and filtering process. It is mainly used to normalize the input size and speed up information extraction of all images by using optimal size. Here the image size is chosen as 224 in both rows and columns. This value is commonly used input size for many deep learning techniques. Therefore, it requires lesser memory and smaller computational time for processing compared to larger size image.

After reducing the image dimension from 500 to 224, the corresponding dimension transformed image is shown in the figure 2.

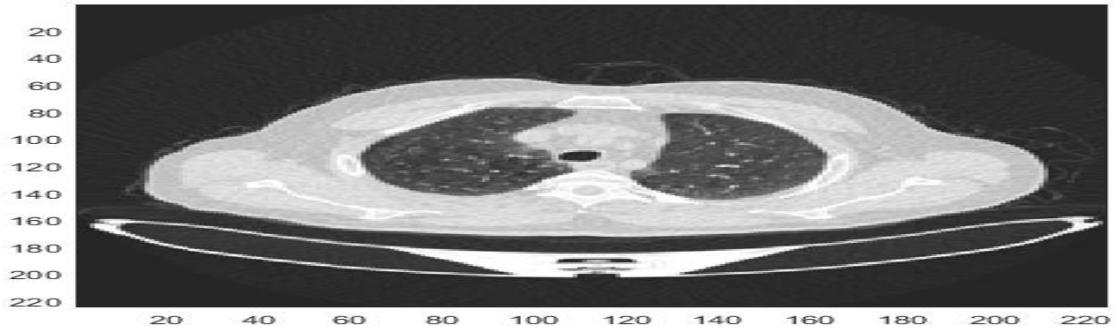

Figure 2. Dimension Transformation

3.4 Intensity transformation

In this phase both semantic and intensity information used for analysis. Input images have three color channel values red, green and blue. which represent the primary color values of the images.

These primary colors transformed into a single-color value using following equation

$$LCIS_{P1} = 0.299 * LCIS_{P1}(:,:,1) + 0.587 * LCIS_{P1}(:,:,2) + 0.114 LCIS_{P1}(:,:,3) \quad (3.4)$$

3.5 Filtering

After dimension transformation the image is subjected to blurring. The blurring in the image removed using median filtering. It is a nonlinear filter and preserve the edge information of image by using median filter.

Blurriness in the image was removed using the following steps:

step1: Median filter is a type of block filter process image block by blocks. Here the kernel sizes 3*3.

step 2: First take 3*3 values.

step 3: Replace centre pixel with the median of its neighbouring pixel.

step 4: Replace original value with sorted values.

step 5: Repeat 2 and 3 for all other blocks of the images.

3.6 Information extraction

This is can be easily diagnosed by seeing the image in an advanced stage of cancer. But in early-stage additional learning is required to define the uniqueness in each image. The uniqueness is extracted from the image information.

4. INTENSITY INFORMATION

The intensity features extracted from the images using their intensity values. The value ranges from 0 and 255. The intensity value 0 and 17 were repeated more than once. Likewise different intensity values can also be repeated. Different information can be extracted from these repeated values and therefore it is called intensity information. Intensity count is a graph which is required before intensity information. It draws between intensity value and repeated count in an image. Based on this count information it can be calculated for image as follows.

4.1 Intensity Properties

The mathematical formula for intensity count calculation for a lung image LI is as follows:

$$Intensity\ Count\ (IC) = \sum_{i=1}^{224}\sum_{j=1}^{224} L_I(i,j) \quad (4.1)$$

In above equation, the value 224 denotes the number of rows and columns of an image. Intensity Count (IC) is calculated using the following six properties of an image.

4.2 Mean

Identify the average value of the intensity levels in an image using the equation

$$Mean\ (\mu) = \frac{1}{224 * 224} \sum_{IL=0}^{256-1} IL * IC \quad (4.2)$$

4.3 Variance

Variance (0^2) is used to describe the difference between the intensity levels

$$\sigma^2 = \frac{1}{224 * 224} \sum_{IL=0}^{256-1} (IL - \mu)^2 * IC \quad (4.3)$$

4.4 Skewness

The intensity level distribution is observed by skewness. It also identifies the distribution side. 0 means equally distributed; positive means highly distributed on right side and negative means highly distributed on left side.

$$S = \frac{1}{224 * 224 * (\sigma^2)^3} \sum_{IL=0}^{255} (IL - \mu)^3 * IC \quad (4.4)$$

4.5 Kurtosis

Identify the peak values in the intensity level of an image

$$K = \frac{1}{224 * 224 * (\sigma^2)^4} \sum_{IL=0}^{255} (IL - \mu)^4 * IC \quad (4.5)$$

4.6 Energy

Define the magnitude of intensity levels and its value being squared.

$$E = \sum_{IL=0}^{255} \frac{IC}{224 * 224} \quad (4.6)$$

4.7 Entropy

Intensity or image information is calculated in depth

$$H = \sum_{IL=0}^{255} \frac{IC}{224 * 224} * \log_2 \left(\frac{IC}{224 * 224} \right) \quad (4.7)$$

The first information set is formed with this intensity properties

$I1 = [\mu_i \ \sigma_i^2 \ S_i \ K_i \ E_i \ H_i]$	(4.8)

5. CORRELATED INTENSITY PROPERTIES

The intensity values were repeated in an image. In order to minimize the repeated values Grey Level Co-occurrence matrix (GLCM) is used. The sample GLCM of an intensity value is shown in figure 3.

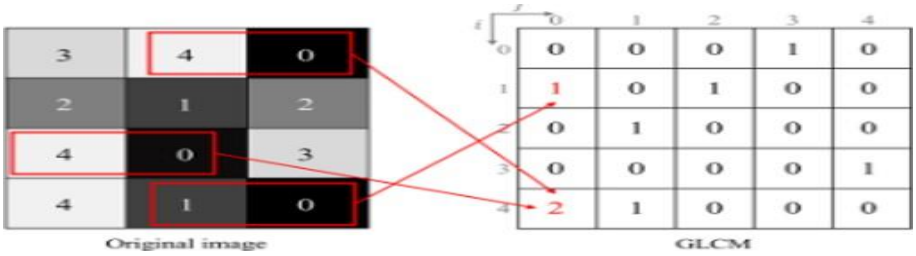

Figure 3. GLCM

5.2 Local Binary Pattern

LBP analyses texture properties of the image texture. Feature vectors are calculated using intensity level and intensity count feature vectors is formed by analysing the image block by block.

LBP feature vector is calculated as:

step 1: The image split block by block.

step 2: Block centre value is calculated by finding the difference with its neighbouring pixel value.

step 3: Monochromatic block is formed by threshold the positive differences as 1 and negative differences as 0.

step 4: Monochromatic block is multiplied with LI block to form LBP feature.

step 5: Repeat step 2 and 4 for all blocks in the image LI.

The overall intensity information IF is calculated as follows:

$I3 = \sum LBP_i$	(4.13)

$I_F = [I1 \ I2 \ I3]$	(4.14)

5.3 SEMANTIC INFORMATION

It gives deeper information through the Google Net and user defined deep learning network.

5.3.1 Google network

It extracts semantic information through its deeper architecture. These deeper architecture does not increase the computational time like other classifier model. It starts with image input layer with size [224, 224, 3]. Computational time is minimized due to parallel operation and also increase the semantic information at different scales.

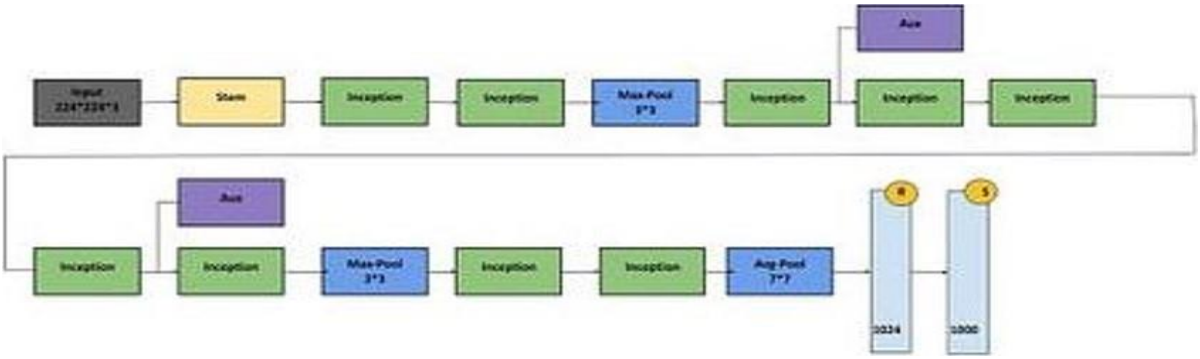

Figure 4. Googlenet

In this proposed research, Googlenet was used only for extracting semantic information. Therefore, the final freely fully connected layers and softmax classification layer was removed.

$SI_1 = G_i(1024)$	(5.3.1)

5.4 User Defined DL

In this article, the Deep learning architecture were modelled using customized CNN as follows:

- Convolution layer of size 7* 7 with filter size 64,128, 256 and 512.
- Max pooling layer of size 3 *3 followed by each convolutional layer.
- Followed by flatten layer.
- Finally, a dense layer with 1024 is used for final feature extraction.

Like google net, it also generates 1024 feature vectors from the image. Second set of semantic information is denoted as follows:

$SI_2 = DL_i(1024)$	(4.16)

The overall semantic information is as follows:

$SI = SI_1 + SI_2$		(4.17)

5.5 Information reduction

Here the total images obtained 2059 feature vectors. Extracting and analysing all information for lung cancer classification results in higher computational time and memory. This 2059 feature vectors were reduced to 8 feature vectors through canonical correlation fusion process.

5.5 Classification

Classification using multi-class support vector machine is used for individual feature vector and reduced feature vectors. Finally, the results were combined to form an ensemble learning and the weights for the individual classifier is based on the individual.

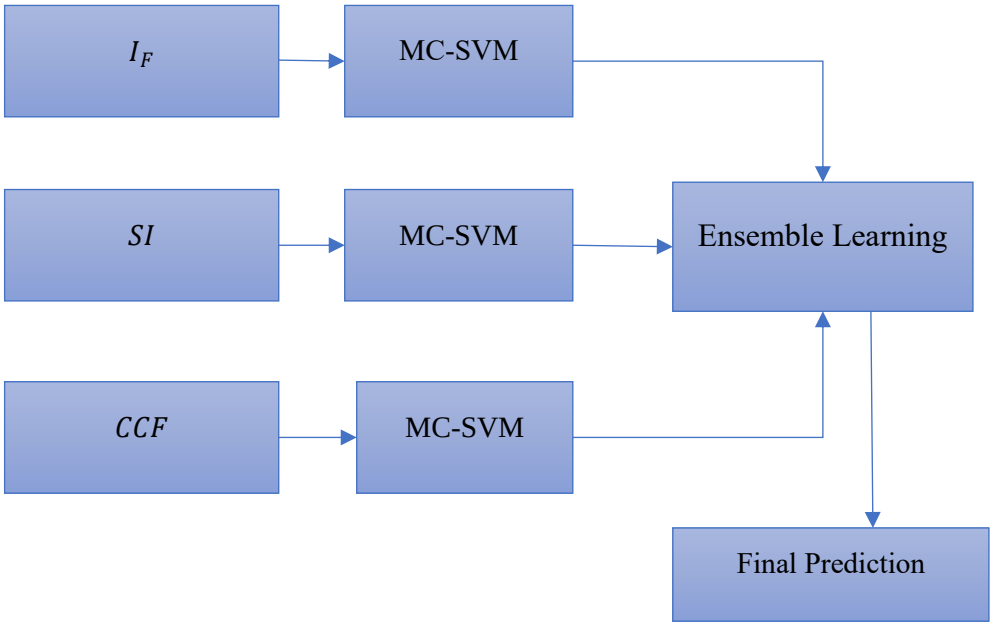

Figure 5. Ensemble Learning

The final prediction of ensemble learning is given in equation 5.5

$Final\ Prediction = 0.25 * MC_{svm}(I_F) + 0.25 * MC_{svm(\ CCF)} + 0.50 * MC_{svm}(SI)$	(5.5)

5.6 Performance evaluation

The performance evaluation is performed on the final prediction using the Multi-Class Confusion Matrix.

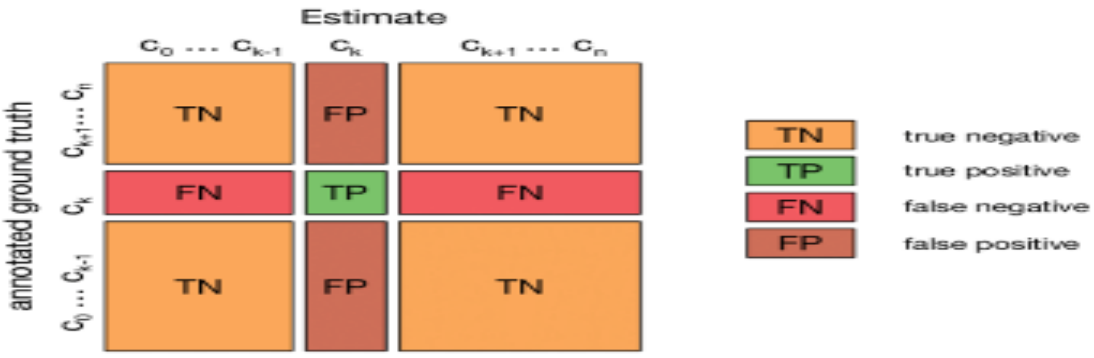

Figure 6: Confusion matrix

Using Multi-Class Confusion Matrix, the evaluation metrics Accuracy, precision, recall and F1-score were calculated.

6. EXPERIMENTAL ANALYSIS

The experimental work is done using Google collab in CPU environment using python 3. The analysis was performed as follows:

- Semantic Information based Lung cancer analysis
- Intensity Information based Lung cancer analysis
- Information based Lung cancer analysis

The above evaluate featurer extraction process on Lung cancer analysis. While the below evaluate the overall performance.

- SF & IF on Iraq set
- SF & IF on kaggle set
- CCF on combined set.

6.1 SF & IF on Iraq Set

In this, the proposed SF and IF classification performance was analyzed on the Iraq dataset and its values were tabulated in table 1.

Table 1. SF & IF on Iraq set

Method	Image Size	Accuracy	Precision	Recall	F I-score	Computational Time(ms)

SCA-CNN [2]	448	99.0	93.0	-	92.4	-
Efficient Net Bl [31]	227	99.1	98.63	98.64	98.63	-
Alex Net [41]	227	98.45	97.10	-	96.4	-
Ebola -CNN [51]	240	93.21	100	90.71	92.72	-
VGG-16	224	97.67	93	98	95	320
Sl (Work 1)	224	98.3	98.7	98.8	98.6	418
IF (Work1)	224	97.5	93.4	98.2	95.2	425

Figure 7, shows minimum image size in the proposed when compared to the existing approach.

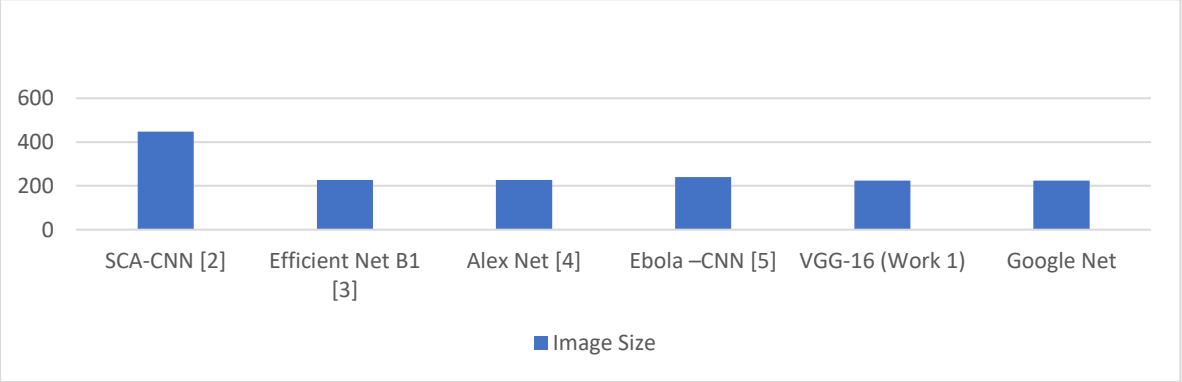

Figure 7. Image Size

6.2 SF & IF on Kaggle Set

In this, the proposed SF and IF classification performance was analyzed on the Iraq dataset and its values were tabulated in table 2.

Table 2. SF & IF on Kaggle set

Method	**Accuracy**	**Precision**	**Recall**	**F I-score**	**Computational time (ms)**
Dense Net (201) [6]	85.8	88.9	86.2	86.3	-
VGG 16 Decision Tree [7]	91	91	93	91	-

VGG-16 [8]	85.8	77	85.75	80.25	-
RESNET	97	97	97	97	488
SI (Work 1)	98.8	98.7	98.8	98.6	418
IF (Work 1)	98.6	97.8	98.3	97.6	428

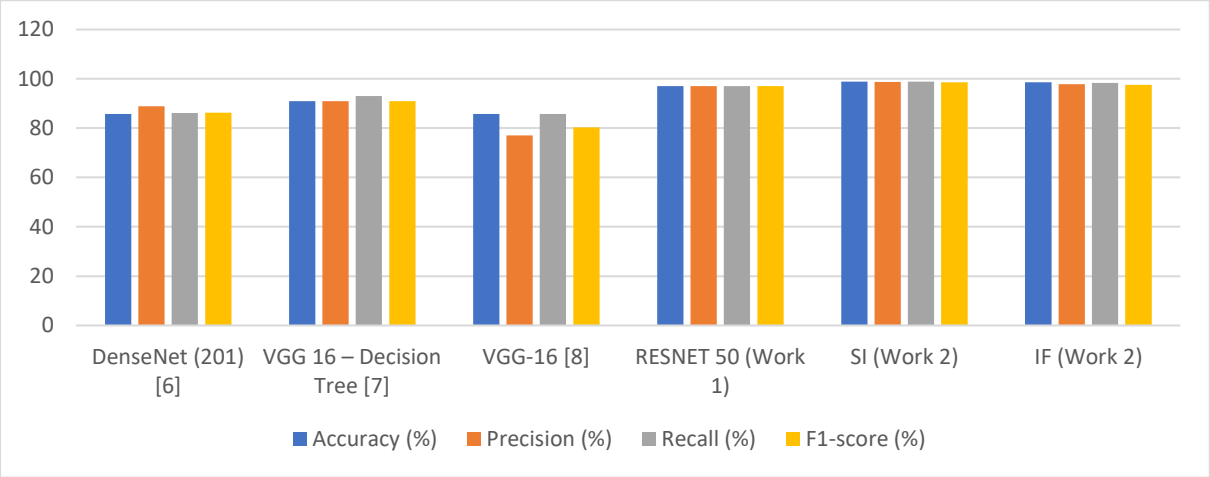

Figure 8. SI & IF performance analysis on Iraq Set

Figure 8, shows that Iraq dataset produced best results for semantic information. whereas proposed hybrid google net and user defined CNN is best for the kaggle dataset.

6.3 CCF ON COMBINED DATASET

In this, the proposed CCF feature set and ensemble learning performance on combined set was analysed and its values are presented in table 4.

Table 3 CCF performance on combined set

Method	**Accuracy**	**Precision**	**Recall (%)**	**F I-score**	**Computational time (ms)**
Work 1	98.9	98	98	98	422
CCF (Work2)	99.2	98.7	98.8	98.6	418

Table 3. shows that the proposed canonical feature and ensemble learning is best for the Lung cancer analysis. For a visual analysis, the results were presented in figure format in fig.7 and 8.

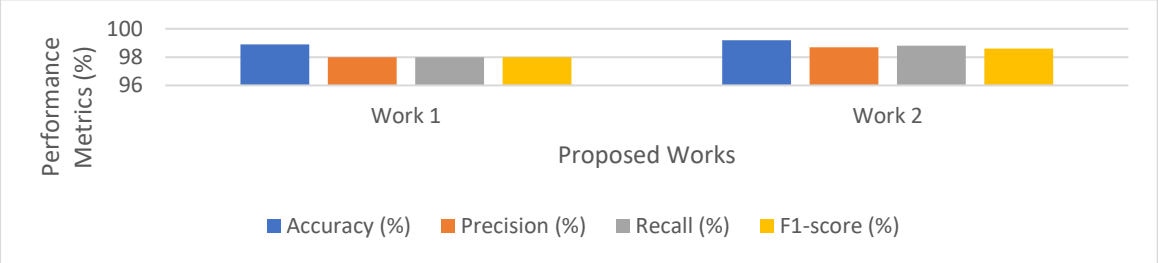

Figure 7. Work 2 performance analysis

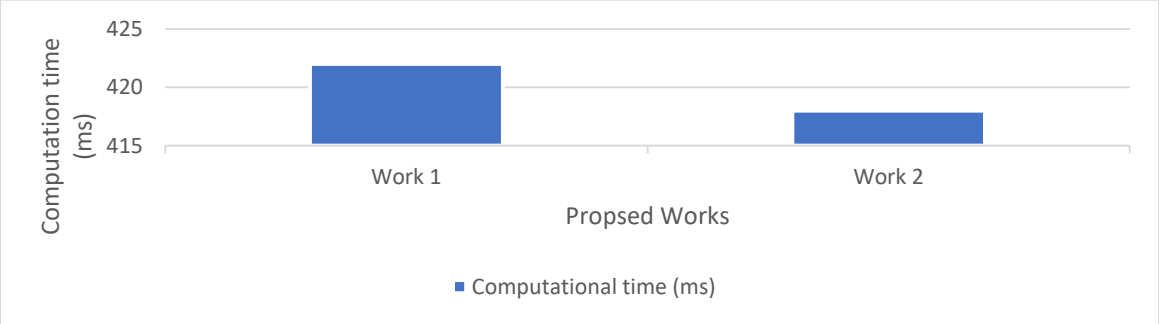

Figure 8. Work 2 computational time

Figures showed that the proposed work 2 performed well on the combined sets with minimal computational time for testing. From, this it can be concluded that the proposed CCF feature set and ensemble learning is best for lung cancer classification.

CONCLUSION

This paper utilized the clubbed data set of Iraq and Kaggle data set. This club data set was analyzed to propose a lung cancer classification approach in terms of both performance and computational manner. Both the semantic and intensity information is extracted for classifying the lung cancer images as Benign, Malignant and normal. From the analysis it was observed that semantic information performed well compared to intensity information. Hence, semantic information used the 50% weightage for ensemble classification and 25% was given to the intensity information and CCF based results. Following, this the ensemble learning based classification performed on combined dataset and the results showed that performance is better in both classification and computational wise then the existing techniques. In future computational time can be further reduced by using swam intelligence approach which helps to select minimal features then the correlated features from CCF.

REFERNCES

1. Lakshmana Prabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems, 92, 374-382.

2. Khan, T., Usman, Y., Abdo, T., Chaudry, F., Keddissi, J. 1., & Youness, H. A. (2019). Diagnosis and management of peripheral lung nodule. Annals of Translational Medicine, 7(15).

3. Asuntha, A., & Srinivasan, A. (2020). Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications, 79(11), 7731-7762.

4.Ofiara, L. M., Navasakulpong, A., Beaudoin, S., & Gonzalez, A. V. (2014). Optimizing tissue sampling for the diagnosis, subtyping, and molecular analysis of lung cancer. Frontiers in oncology, 4, 253.

5.Bak, S. H., Kim, C., Kim, C. H., Ohno, Y., & Lee, H. Y. (2022). Magnetic resonance imaging for lung cancer: a state-of-the-art review. Precision and Future
Medicine, 6(1), 49-77.

6. Mahersia, H., Zaroug, M., & Gabralla, L. (2015). Lung cancer detection on CT scan images: a review on the analysis techniques. Lung Cancer, 4(4), 10-14569.Jalali, V., & Kaur.

7.D. (2020). A study of classification and feature extraction techniques for brain tumor detection. International Journal of Multimedia Information Retrieval, 9(4), 271-290.

8.Serin, G., Sener, B., Ozbayoglu, A. M., & Unver, H. O. (2020). Review of tool condition monitoring in machining and opportunities for deep learning. The International Journal of Advanced Manufacturing Technology, 109(3), 953-974.

9.Luca, A. R., Ursuleanu, T. F., Gheorghe, L., Grigorovici, R., lancu, S., Hlusneac, M., & Grigorovici, A. (2022). Impact of quality, type and volume of data used by deep learning models in the analysis of medical images. Informatics in Medicine Unlocked, 29, 100911.

10.Da Nöbrega, R. V. M., Peixoto, S. A., da Silva, S. P. P., & Rebouqas Filho, P. P. (2018, June). Lung nodule classification via deep transfer learning in CT lung images. In 2018 IEEE 31st international symposium on computer-based medical systems (CBMS) (pp. 244-249). IEEE.

11.Sathyan, H., & Panicker, J. V. (2018, July). Lung nodule classification using deep ConvNets on CT images. In 2018 9th International conference on computing, communication and networking technologies (ICCCNT) (pp. 1-5). IEEE.

12.Nishio, M., Sugiyama, 0., 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.

13.Zhang, G., Jiang, S., Yang, Z., Gong, L., Ma, X., Zhou, Z., & Liu, Q. (2018). Automatic nodule detection for lung cancer in CT images: A review. Computers in biology and medicine, 103, 287-300.

14.Dey, R., Lu, Z., & Hong, Y. (2018, April). Diagnostic classification of lung nodules using 3D neural networks. In 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) (pp. 774-778). IEEE.

15.Rodrigues, M. B., Da Nobrega, R. V. M., Alves, S. S. A., Reboucas Filho, P. P., Duarte, J. B. F., Sangaiah, A. K., & De Albuquerque, V. H. C. (2018). Health of things algorithms for malignancy level classification of lung nodules. IEEE Access, 6, 18592-18601.

16.Chen, C. H., Chang, C. K., Tu, C. Y., Liao, W. C., wu, B. R., Chou, K. T., ... & Huang, T. C. (2018). Radiomic features analysis in computed tomography images of lung nodule classification. Plos one, 13(2), e0192002.

17.Liu, X., Hou, F., Qin, H., & Hao, A. (2018). Multi-view multi-scale CNNs for lung nodule type classification from CT images. Pattern Recognition, 77, 262-275.

18.Malik, H., & Anees, T. (2022). BDCNet: Multi-classification convolutional neural network model for classification of COVID-19, pneumonia, and lung cancer from chest radiographs. Multimedia Systems, 28(3), 815-829.

19. Asuntha, A., & Srinivasan, A. (2020). Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications, 79(11), 7731-7762.

https://www.kaggle.com/datasets/nih-chest-xrays/data