https://doi.org/10.48047/AFJBS.6.7.2024.2833-2843

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

# DEEP LEARNING-BASED PREDICTION OF LUNG CANCER RISK FACTORS USING AUGMENTED IMAGES AND INTEGRATED FEATURE ELIMINATION

**Dr.K.S.R.Radhika[1], Kallepally Sai Deepthi[2]**

[1]Professor, TKR College of Engineering & Technology, Hyderabad, Telengana, ksrradhika@tkrcet.com

[2]Post Graduate Student, College of Engineering & Technology, Hyderabad, Telengana, 22K91D5803@tkrcet.com

ArticleHistory
Volume:6,Issue7,2024
Received:30May2024
Accepted:26June2024
doi:10.48047/AFJBS.6.7.2024.2833-2843

**ABSTRACT:** Prediction of disease is essential for identifying the risk factor. Lung cancer is a dangerous disease which has to be detected in initial stage. Based on images the risk of the disease is identified but analysing for every patient is tuff. By utilizing DL techniques, the prediction can be analyzed in short period of time. For predicting images CNN is the most efficient methodology for its layer structure. The proposed model to increase the efficiency of the system it has generated augmented images using GAN approach because the performance and generalizability of the trained models may be impacted by biases introduced by GANs into the generated data. Thorough validation and assessment are required to reduce these hazards.Class imbalances are typical in medical datasets, where there may be considerable differences in the quantity of samples for distinct classes. By creating artificial examples for underrepresented classes, GANs can assist in balancing the dataset.The synthetic images are passed as input to the fine-tuned AlexaNet then it has extracted the features using vector integrated feature elimination. Since every layer in AlexNet's feature extraction process captures a distinct level of abstraction in the image, the features are generally interpretable. Understanding which characteristics are crucial for identifying malignant from non-cancerous areas in lung imaging can be aided by this. Finally, the model detects the cancer in lungs using the non-linear SVM. Hence the prediction was performed through validation techniques, ROC, AUC was also performed.

**Keywords: Alexnet, Vector Feature Elimination, Augmentation, Synthetic Images, Deep Learning, Adversial Networks**

**1. INTRODUCTION:**A group of image-processing methods called morphological processes can change the way shapes are put together in an image [9]. When working with binary images, where each of the pixels is either black (0) or white (1), these methods are most appropriate. Noise removal, edge recognition, picture segmentation, and object separation are all popular jobs that use them. Some of the most basic structural processes are weathering,

enlargement, opening and closing, and so on. Erosion can make the edges of items in the centre of a picture smaller or wear away. Moving a structure element over the original picture is what it does. At each structural element location, the output pixel is white if every one of the pixels beneath it are white, otherwise black 10]. This process eliminates small items and makes edges thinner. Erosion is different from dilation. It makes the edges of things in the centre of a picture bigger. If at least one pixel below the structural element is white at any given point, the output pixel turns white; otherwise, it stays black.

In the field of machine learning, generative adversarial networks, or GANs, are a breakthrough technique, especially when it comes to generative modeling. GANs, which were created in 2014 by Ian Goodfellow and his associates, are made up of two neural networks playing a game of competition: the discriminator and the generator. While the discriminator gains the ability to discern between actual and false data, the generator attempts to produce artificial data samples that are identical to real ones [11]. Adversarial training, in which both networks are updated concurrently, enables GANs to produce realistic, high-quality outputs in a variety of domains, such as text, music, and pictures. The working is shown in figure x

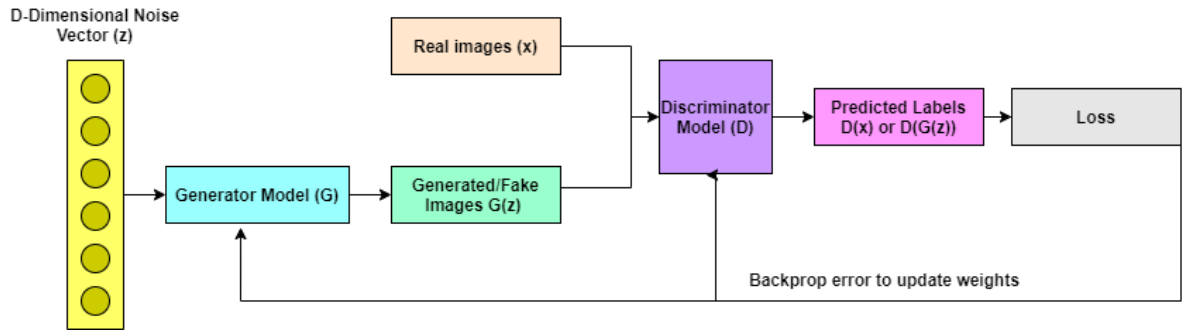


**Figure 1: Working of GAN**

At the heart of GANs lies the generator network, which learns to map random noise from a latent space to the data space, producing synthetic samples. Initially, the generator generates poor-quality outputs, but through iterative training, it refines its parameters to produce increasingly realistic data[12]. Simultaneously, the discriminator network has the ability to discern between authentic and fraudulent input, offering the generator feedback. Training makes the generator more adept at fooling the discriminator, which results generated data that is quite similar to actual samples[13]. Both networks are driven toward a Nash equilibrium by this adversarial process, in which the discriminator is unable to consistently discriminate between the two types of data and the generator produces data that is identical to actual data.

## 2. LITERATURE SURVEY:

Akitoshi Shimazaki et al [1] has considered a chest radiographs images with different age women patients. On the available datasets the DL models are applied then trained & validated. Later the data was tested with independent dataset collected from the University. The DL model was CNN with segmentation. This method also utilized the encoder and decoder path where for the outcome of segmentation. By using the coders decreases the resolution of the mapping of features and increase in robustness and overfitting. Here two

kinds of images are present normal and black-white i.e., Inversion of radiograph. This black-white was considered by the augmentation the spots. This has good advantage i.e., overlapping with blind spots in the image. For every image CNN and ensemble models are applied. For validating five fold of cross validation is applied. For these 100 epochs are utilized from the Adam and it has provided optimized performances.

Abdul Rahaman Wahab Sait et al [2] PET/CT images are derived from the datasets which are rich and diverse info. The images are inputted as initially image pre-processing was applied with SyN function. This function can cross-correlation computation with the generation for the normalization and has high resolution in the smoothness of image. In data augmentation GAN was utilized for the generation of the synthetic images. In extraction of features densenet 121 was utilized from the pre-trained weights. For reducing the dimensionality reduction auto encoders are utilized were the accurate data was detected. The LC model utilizes the MobileNet V3 for the optimization values. The convolution models are used with four layers. Final layer are trained on the weights of the MobileNet and remaining layer interacted with initialized method and followed by softmax function. The model was fine-tuned with all the parameters with Adam optimization techniques.

Stephanie T. Junger et al [3] has utilized the NSCLC clinical data it contains 300 images of different aged patients. 3D CNN method with DeepMedic was utilized on the dataset for efficient performances. The method was divided into two pathways: where one path contains isotropic patches and other similar. CNN contains 11 layers where the size was 3*3 kernels. From 4 to 8 layer's it contains normalization and remaining layer are fully connected and final layer contains of single size 1*1. The data was pre-processed by the DLM method by automatic. This process involves five phases for efficient requirement. Now the images are trained on 5 folds of cross-validation of the overlapping info. The model was trained with 10 set with fixed size of 35 epochs each. Finally, the model is evaluated using the validation techniques.

Marwa Obayya et al [4] has considered the LC25000 dataset contains 5k samples of lungs & colon images. Here the images are trained for the processing into the input and start the prediction. In pre-processing the method utilizes the GFT method for the efficient prediction of the images with accurate prediction. The outcome of pre-processing method now checks for the hyperparameter processing by utilizing the Adaptive Fuzzy AO for the prediction of feature extractions. In extraction of features the GhostNet methodology was utilized. The structure of this method contains four stages features, conv layers, separation, and output. Where the conv layers are directly connected to output with direct identity. Now the weights are goes under hyper parameter processing by using Tuna Swarm Approach. Finally, lungs and colons are identified by the Echo state network methodology and the evaluation of the method was done with five metrics.

Mohammad Alamgeer et al [5] Lungdb was the dataset considered where it contains four class where 197 samples are present. The trained images are inputted to the fusion process of the prediction. The method was processed to the fusion features with three methodologies. Densely connected network model, mainly for the integrations of features in various distinct places. The Inception-ResNetV2 are utilized both the method involves 1D conv layers, BN,

AF with the kernel size of 7 and padding. From hyper-parameter are retrieve by Dung Beetle optimization approach. To predict the accurate image LSTM is utilized. Based on the time the HN method initiates to zero this is for the gradient diminishing which leads to RNN for long time sequencing models. Now this calculates by edge set of weight nodes. This has allowed for stored huge data in a period and vanishing issues are removed by NN. Finally, the model is evaluated with five techniques.

Lulu Wang [6] The dataset contains lung and clinical data which has little and huge data where in this 16 datasets are derived. Now the scanning image are collected and removed all the unnecessary data from the dataset. The pre-processing results are segmented with related models. In segmentation the model has scaled the images in different stages and high parameters are considered for further process. Now all segmented images are transferred to the DL model to the detection of lung cancer images. CNN was utilized for the prediction it contains 5 layers where conv layer, ReLu, Pooling layer, FC layer, and Softmax layers. The conv layer contains four layers where the ReLU model for the couple of data. Those are divided into several pooling layers. Now the FC connects with the several phases of the image and then the final layer contains the softmax layer which has intense of images. Now the testing phase was evaluated with the two divisions i.e., malignant and benign.

Imran Shafi et al [7] LUNA16 is utilized which contains CT scanned data this are in DICOM format. The implementation was very simple with five stages. In acquisition the images are collected and send to the pre-processing phase. In this stage the raw data was converted into the required pixels and they are segmented using capsNet segmentation then again the data is segmented previous pixel sizes then the data is extracted as ROI and finally the pixel sizes are reduced with ROI patch. Now this data was transformed to classification path. In segmentation the capsule NN method was utilized for accurate and efficient prediction. For extraction of feature the CNN methodology was applies then for the final stage of four ML methods are considered. This intake was considered as the hybrid classification. Among all the methodologies SVM has high performances and efficiency.

A. Gopinath et al [8] the National Cancer Institute provided 1018 lung CT images with proteomic and genetic clinical data. The suggested DL-based diagnostic and classification paradigm has three parts. The pre-processing and augmentation phase Histogram algorithms improve picture quality. Data augmentation using affine transformations addresses overfitting. The third step uses pixel-based processing to produce saliency maps based on colour and spatial variations. The planned training uses CNN. A Cat optimisation technique inspired by cat sleep and hunting behaviours addresses SGD's limits in CNN training. Cat optimisation uses seeking & tracing modes. In searching modes, CDC, SMP, and SRD matter. Tracing modes assign random velocity values to all cat position dimensions. The suggested approach optimises fully connected layer hyperparameters using Cat. Key hyperparameters are learning rate, number of epochs, input weights, and hidden layers. Table 1 discusses about the limitations of the existing approaches

**Table 1: Limitations of Existing Approaches**

| **Author** | **Algorithm** | **Merits** | **Demerits** | **Accuracy** |
|---|---|---|---|---|
| Akitoshi Shimazaki et al | DL | With minimum folds the DL is performed. | In segmentations the accuracy is low. | 89.10% |
| Abdul Rahaman Wahab Sait | LC | At each stage the model is evaluated. | Compared methodologies does not have much differences. | 98.6% |
| Stephanie T. Junger et al | DLM | 3D segmentation was accurate. | Validation was not defined properly. | 96% |
| Marwa Obayya et al | BICLCD-TSADL | Multiple techniques are applied for perfect prediction. | Computational complexity is high. | 99.3% |
| Mohammad Alamgeer et al | DBOMDFF-LCC, TL | Fusion and parameter tuning is perfect. | It accurately identifies only normal persons compared to remaining attributes. | 99.17% |
| Lulu Wang | CNN | Each and every part was accurately defined. | Based on the clinical variations the validations has huge change. | |
| Imran Shafi et al | DL, SVM | The implementation was accurate and efficient. | Performances has to be increased. | 94% |
| A. Gopinath et al | CNN, CAT Optimization | Based on the decisions the prediction was accurately. | Only particular disease can be identified. | 99.9% |

## 3. PROPOSED METHODOLOGY:

GANs have found applications in numerous fields, including image generation, image-to-image translation, super-resolution, and style transfer. In image generation tasks, GANs have demonstrated remarkable capabilities in producing photorealistic images of faces, landscapes, and even abstract art. Moreover, GANs have facilitated the creation of deepfake technology, which generates highly realistic fake videos by manipulating existing footage. Even with their success, GANs have drawbacks include training instability and mode collapse, in which the generator generates a restricted range of samples. Researchers continue to explore novel architectures and training techniques to address these issues and unlock the full potential of GANs in various domains, driving advancements in generative modeling and artificial

intelligence as a whole. We'll consider a scenario where we want to generate grayscale images of size 28x28.

**1. Generator Network (G):**

- Input: Random noise vector z of size 100 (arbitrary choice)

- Output: Image G(z) of size 28x28

**2. Discriminator Network (D):**

- Input: Image x of size 28x28

- Output: Probability D(x) that x is a real image

**3. Example:**

Let's consider a specific iteration of the GAN training process:

**- Step 1: Generator Generates Fake Images**

- Generate a batch of 32 noise vectors $z^{(1)}$, $z^{(2)}$, ..., $z^{(32)}$ randomly sampled from a normal distribution.

- Pass each noise vector through the Generator to generate fake images: $G(z^{(1)})$, $G(z^{(2)})$, ..., $G(z^{(32)})$.

**- Step 2: Discriminator Evaluates Real and Fake Images**

- Simultaneously, sample a batch of 32 real images $x^{(1)}$, $x^{(2)}$, ..., $x^{(32)}$ from the dataset.

- Compute the probabilities assigned by the Discriminator to real images: $D(z^{(1)})$, $D(z^{(2)})$, ..., $D(z^{(32)})$.

- Compute the probabilities assigned by the Discriminator to fake images generated by the Generator: $D(G(z^{(1)}))$, $D(G(z^{(2)}))$, ..., $D(G(z^{(32)}))$.

**- Step 3: Update Discriminator**

- Calculate the Discriminator loss using binary cross-entropy for both real and fake images:

$$J^{(D)} = -\frac{1}{64}\left[\sum_{i=1}^{32} \log log\, D\left(x^{(i)}\right) + \sum_{i=1}^{32} \log log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right]$$

- Update the Discriminator's weights based on the gradient of $J^{(D)}$ with respect to its parameters.

**- Step 4: Update Generator**

- Generate another batch of 32 noise vectors $z^{(1)}$, $z^{(2)}$, ..., $z^{(32)}$.

- Pass these noise vectors through the Generator to generate fake images: $G(z^{(1)})$, $G(z^{(2)})$, ..., $G(z^{(32)})$.

- Calculate the Generator loss using binary cross-entropy, aiming to fool the Discriminator:

$$J^{(G)} = -\frac{1}{32}\left[\sum_{i=1}^{32} \log log\ D\left(G(z^{(i)})\right)\right]$$

- Update the Generator's weights based on the gradient of $J^{(G)}$ with respect to its parameters.

**- Step 5: Iteration**

- Repeat the process for multiple iterations, alternating between updating the Discriminator and the Generator.

**4. Evaluation:**

- After training, the Generator should be able to produce realistic-looking handwritten digits that resemble those in the dataset.

- The Discriminator should have learned to effectively distinguish between real and fake images.

When extracting features from a dataset, a ranking process is used to determine which characteristics are most important or significant. The aforementioned procedure has significant importance in the domains of ML and data analysis, as it facilitates the reduction of dataset dimensionality, enhances model efficacy, and provides valuable insights into the fundamental patterns within the data. This concept of feature importance pertains to the significance or impact of each feature in forecasting the target variables or capturing the fundamental patterns within the data. The estimation of feature significance may be accomplished by the use of several methodologies, such as statistical testing, analysis of correlation, and ML techniques. Following the determination of feature significance, a ranking method is used to arrange the features in order of their respective importance ratings

The RFE method is a technique for feature selection that employs an iterative process to rank and delete variables according to their significance or impact on the performance of the model. It aids in determining a subset of characteristics that optimise predicted accuracy while minimising the negative effects of overfitting and computing complexity. RFE is a feature selection method that gets rid of the least important features in a dataset over and over again based on how well a certain machine learning strategy works. To get to the desired number of features, it trains the model over and over again on the leftover group of features and rates how important each one is. Let's use mathematical language to go into more information about the RFE algorithm. Let's say having a dataset with m samples as well as n features, which is shown by X (an m×n matrix), and a goal variable called y that goes with it.

SVM seeks the optimum hyperplane to partition the feature space across categories with the greatest space. Check how well the SVM model works on a reference or test dataset with the chosen group of features. Standard ways to judge sorting jobs. Check how well the SVM model works with the chosen features compared to how well it works when trained on all the features. To get even better results from the SVM model, you can tune its hyperparameters. The regularisation parameter as well as the kernel type are two

hyperparameters that can have a big effect on how well the SVM works. For example, grid search or randomised search can help you find the best hyperparameters. Use cross-validation to check the SVM model's efficacy and see how well it can generalise to data it hasn't seen before. Cross-validation helps stop overfitting and gives a more accurate picture of how well the model works.

## 4. RESULTS & DISCUSSION:

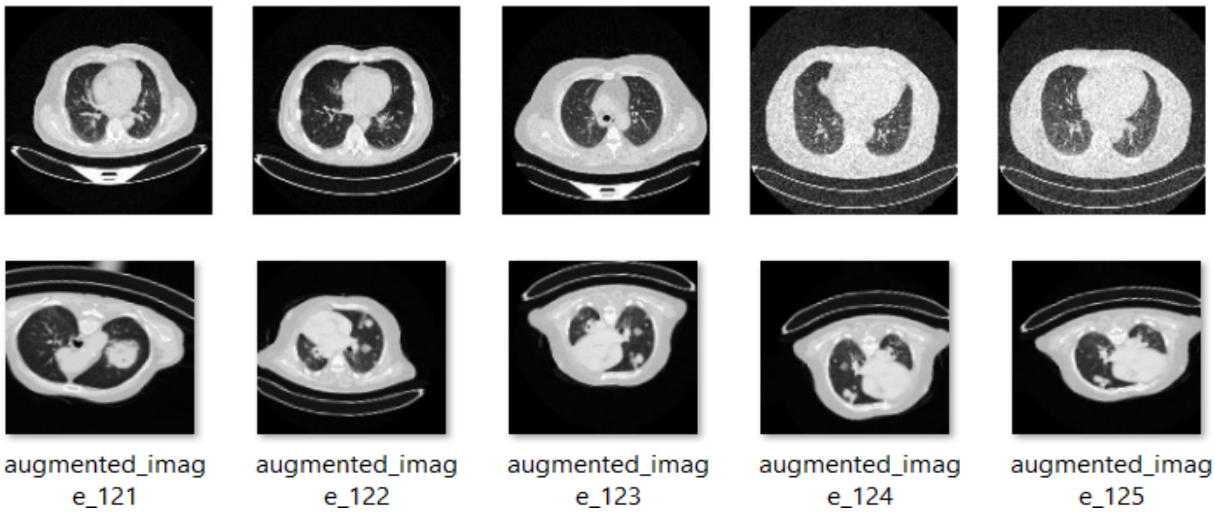


**Figure 2: Augmented Images**

Figure 2 describes about the synthetic images created by the GAN. For training purposes, the produced images must closely mimic real lung scans. Ensuring the quality and realism of generated images is critical to the effectiveness of GAN-based augmentation.

Figure 3 displays the augmented images of the original images. These images help the model to solve the real time scenarios. GANs are powerful tools to handle the uncontrolled conditions with ease and efficient accuracy. The number of images in the dataset is increased by 3 times approximately

```
Number of original images: 1097
Number of augmented images: 3291
Total number of images after augmentation: 4388
Total number of labels after augmentation: 4388
```

**Figure 3: Dataset Size After Augmentation**

Figure 4 represents the layered summary of AlexNet. AlexNet, a convolutional neural network (CNN), is a powerful feature extractor for lung cancer detection applications. AlexNet provides various advantages by leveraging pre-trained features from large-scale picture categorization tasks like ImageNet. It introduces transfer learning capabilities, allowing the model to inherit knowledge from a variety of picture datasets and apply it to lung cancer detection, minimizing the requirement for substantial labeled data and improving generalization.

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_5 (Conv2D) | (None, 57, 57, 96) | 34944 |
| max_pooling2d_3 (MaxPooling2D) | (None, 28, 28, 96) | 0 |
| conv2d_6 (Conv2D) | (None, 28, 28, 256) | 614656 |
| max_pooling2d_4 (MaxPooling2D) | (None, 13, 13, 256) | 0 |
| conv2d_7 (Conv2D) | (None, 13, 13, 384) | 885120 |
| conv2d_8 (Conv2D) | (None, 13, 13, 384) | 1327488 |
| conv2d_9 (Conv2D) | (None, 13, 13, 256) | 884992 |
| max_pooling2d_5 (MaxPooling2D) | (None, 6, 6, 256) | 0 |
| flatten_1 (Flatten) | (None, 9216) | 0 |

...
Total params: 58293635 (222.37 MB)
Trainable params: 58293635 (222.37 MB)
Non-trainable params: 0 (0.00 Byte)

**Figure 4: Layers in Neural Network**

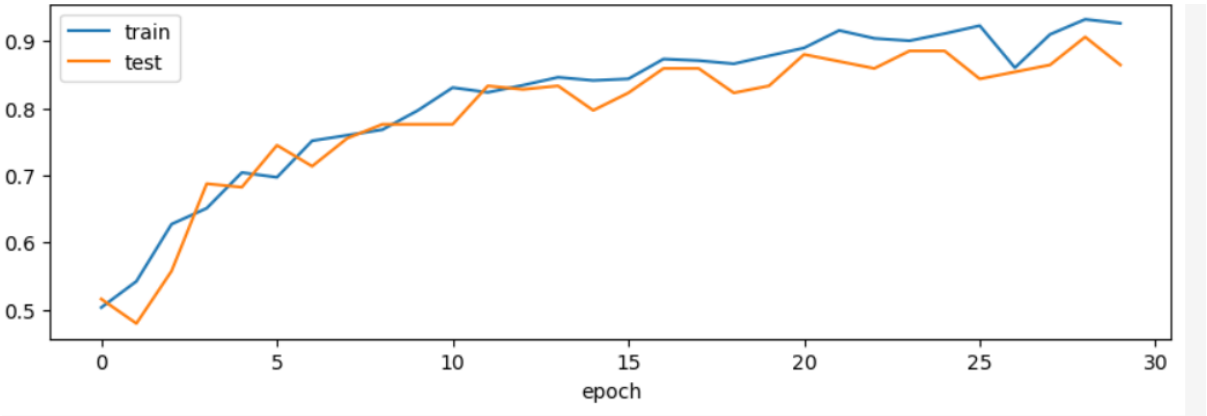


**Figure 6: Accuracy Presentation**

Figure 6 represents the evaluation of the model using non-linear SVM and AlexNet. The model has trained a non-linear SVM classifier using the extracted features from the training set and has performed cross-validation to obtain more robust estimates of the model's performance.

**5. CONCLUSION**:Lung cancer was most effective disease in the time of COVID and differentiating the disease with other lung disease is been complicated. To overcome this phase CNN was the best prediction through X-rays and reports of the patient. The dataset was considered from the Kaggle, an opensource. Pre-processing and augmentation of data has retrieved best parameters for transmission of data to train and test phase. Now the core of the process CNN can be customized according to the clinical data requirement. The model has extracted the important features using the ranking mechanism. In the proposed CNN it is designed with four stages of prediction conv layers and pooling are used to segment pots on the images. The outcome was assessed to flatten layer the parameters are arranged in a sequences order. The optimized parameters are transformed in fully connected layers and those are predicted according to the requirement. Finally, the process has evaluated using

validation techniques and curve direction. Hence the proposed method was compared to previous approaches for identifying the efficiency and performances.

**REFERENCES:**

[1] Shimazaki, A., Ueda, D., Choppin, A., Yamamoto, A., Honjo, T., Shimahara, Y., & Miki, Y. (2022). Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-021-04667-w

[2] Wahab Sait, A. R. (2023). Lung Cancer Detection Model Using Deep Learning Technique. Applied Sciences, 13(22), 12510. https://doi.org/10.3390/app132212510

[3] Jünger, S. T., Hoyer, U. C. I., Schaufler, D., Laukamp, K. R., Goertz, L., Thiele, F., Grunz, J. P., Schlamann, M., Perkuhn, M., Kabbasch, C., Persigehl, T., Grau, S., Borggrefe, J., Scheffler, M., Shahzad, R., & Pennig, L. (2021). Fully Automated MR Detection and Segmentation of Brain Metastases in Non-small Cell Lung Cancer Using Deep Learning. Journal of Magnetic Resonance Imaging, 54(5), 1608–1622. https://doi.org/10.1002/jmri.27741

[4] Obayya, M., Arasi, M. A., Alruwais, N., Alsini, R., Mohamed, A., & Yaseen, I. (2023). Biomedical Image Analysis for Colon and Lung Cancer Detection Using Tuna Swarm Algorithm with Deep Learning Model. IEEE Access, 11, 94705–94712. https://doi.org/10.1109/ACCESS.2023.3309711

[5] Alamgeer, M., Alruwais, N., Alshahrani, H. M., Mohamed, A., & Assiri, M. (2023). Dung Beetle Optimization with Deep Feature Fusion Model for Lung Cancer Detection and Classification. Cancers, 15(15). https://doi.org/10.3390/cancers15153982

[6] Wang, L. (2022). Deep Learning Techniques to Diagnose Lung Cancer. In Cancers (Vol. 14, Issue 22). MDPI. https://doi.org/10.3390/cancers14225569

[7] Shafi, I., Din, S., Khan, A., Díez, I. D. L. T., Casanova, R. del J. P., Pifarre, K. T., & Ashraf, I. (2022). An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network. Cancers, 14(21). https://doi.org/10.3390/cancers14215457

[8] Gopinath, A., Gowthaman, P., Venkatachalam, M., & Saroja, M. (2023). Computer aided model for lung cancer classification using cat optimized convolutional neural networks. Measurement: Sensors, 30. https://doi.org/10.1016/j.measen.2023.100932

[9] A. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha, “Efficient Face Recognition by Compact Symmetric Elliptical Texture Matrix (CSETM)”, Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 4-Regular Issue, 2018.

[10] SRI SILPAPADMANABHUNI SRINIVASA REDDY K, P. VENKATESWARA RAO, A.MALLIKARJUNA REDDY, K SUDHEER REDDY, DR. J. LAKSHMI NARAYANA "NEURAL NETWORK AIDED OPTIMIZED AUTO ENCODER AND DECODER FOR DETECTION OF COVID-19 AND PNEUMONIA USING CT-SCAN" Journal of

Theoretical and Applied Information Technology,15th November 2022. Vol.100. No 21,pp.no.6346-6360, 2022.

[11] Mallikarjuna A. Reddy, Sudheer K. Reddy, Santhosh C.N. Kumar, Srinivasa K. Reddy, “Leveraging bio-maximum inverse rank method for iris and palm recognition”, International Journal of Biometrics, 2022 Vol.14 No.3/4, pp.421 - 438, DOI: 10.1504/IJBM.2022.10048978.

[12] Ozdemir, O., Russell, R. L., & Berlin, A. A. (2019). A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1902.03233

[13] Z. Li et al., "A Novel Deep Learning Framework Based Mask-Guided Attention Mechanism for Distant Metastasis Prediction of Lung Cancer," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 7, no. 2, pp. 330-341, April 2023, doi: 10.1109/TETCI.2022.3171311

[14] H. Jiang, H. Ma, W. Qian, M. Gao and Y. Li, "An Automatic Detection System of Lung Nodule Based on Multigroup Patch-Based Deep Learning Network," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 4, pp. 1227-1237, July 2018, doi: 10.1109/JBHI.2017.2725903

[15] Y. Xie et al., "Knowledge-based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT," in IEEE Transactions on Medical Imaging, vol. 38, no. 4, pp. 991-1004, April 2019, doi: 10.1109/TMI.2018.2876510