Lung Cancer Diagnosis using a light weight deep learning model

Mohit Agarwal[1], Vivek Mehta[1], Rohit Kr Kaliyar[1], and Suneet Kumar Gupta[1]

Bennett University, Greater Noida 201310
26.mohit@gmail.com
vivek.mehta@bennett.edu.in
rohit.kaliyar@bennett.edu.in
suneet.banda@gmail.com

Abstract. Lung cancer is a disease in which lungs get infected by cancerous development of cells. This can be caused due to excessive smoking. However persons who do not smoke may also get the disease in today's polluted environment. The symptoms of lung cancer can be cough which does not cure, blood in cough, pain in chest, loosing weight. etc. The CT scans are used to diagnose type of cancer for their corresponding treatment. Generally lung cancer can be classified into 3 types of cancer: Adenocarcinoma, Squamous cell carcinoma, and Large cell carcinoma. To avoid any mis diagnosis machine learning and deep learning methods are very helpful to classify the exact type of cancer and whether it is present or not. Machine Learning methods such as Decision Trees (DT) and Random Forest (RF) gave very good performance with RF giving 97% accuracy. Similarly Convolution Neural Networks (CNN) such as Mobilenet and VGG19 were tested to give an accuracy of 78.12% and 81.25% respectively. A three layered CNN was also proposed to give an accuracy of 89%. Compressed MobileNet accuracy could be enhanced to 96.5%.

Keywords: CNN · Compression · Lung Cancer · Acceleration.

1 Introduction

Lung cancer being a leading cause of death in all countries worldwide [1] with cancer classified into different types of carcinoma, it is important to diagnose the disease early without physician's difference of opinions.

With new IoT edge devices [2] taking a lead role in computer industry, it is important to make a compressed version of heavy models without loss in performance. This was accomplished in this research using a Differential Evolution (DE) based approach [3].

Several recent research articles show the usage of CNN for image classification in various fields [4–10]. Numerous other research works also show the usage of different type of CNN: UNet, SegNet, FCN, etc. for the segmentation of desired image portions [11–13]. In multiple other research articles usage of CNN for human disease diagnosis is also found [14–16].

ICISML 2024, January 05-06, Pune, India

DOI 10.4108/eai.5-1-2024.2341884

Recently research has started for the compression of deep neural networks using meta-heuristic techniques [17–23].

Chaunzwa et al. [24] showed that using VGG16 based CNN an AUC of 0.71 can be obtained with dataset of 51 patients and two classes: adenocarcinoma (ADC) and Squamous Cell Carcinoma (SCC). Authors also showed that k-NN can also give an AUC value of 0.71.

Wang et al. [25] have proposed a CNN based on VGG16 by replacing last dense layers with their own dense layers of size 1024 and removing padding in the layers. Authors have shown an accuracy of 97.3% on their own dataset and AUC of 0.856 on a public dataset.

Yang et al. [26] have demonstrated that for a six class classification using EfficientNet-B5 a maximum AUC of 0.978 could be obtained using a dataset of six classes: three types of carcinoma, tuberculosis, pneumonia, and normal lung. Authors tested with ResNet50 also and found EfficientNet to give better results.

The article's major contributions are:

- The article helps in effectively diagnosing type of lung cancer from CT scan images.
- The paper also discusses usage of Differential Evolution for compression of heavy CNN models.
- A four objective fitness function was designed for good results of DE.

The rest of paper is laid out as shown in Figure 1.

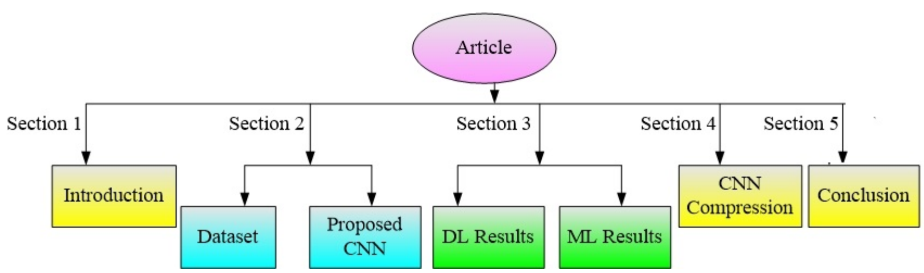

Fig. 1. Layout of the paper.

2 Material and methods

2.1 Lung Cancer CT scan dataset

The lung cancer dataset was obtained from a kaggle challenge posted by Mohamed Hany [27]. Samples of four classes of this dataset are shown in Figure 2. The infected part of lung is clearly indicated by a red arrow.

Description of dataset is shown in Table 1. Since the number of images was small hence these were augmented to create new images from existing images by a random amount of rotation between -10° and 10°. This helped in balancing the images in 4 class so that results will not be biased for any particular class.

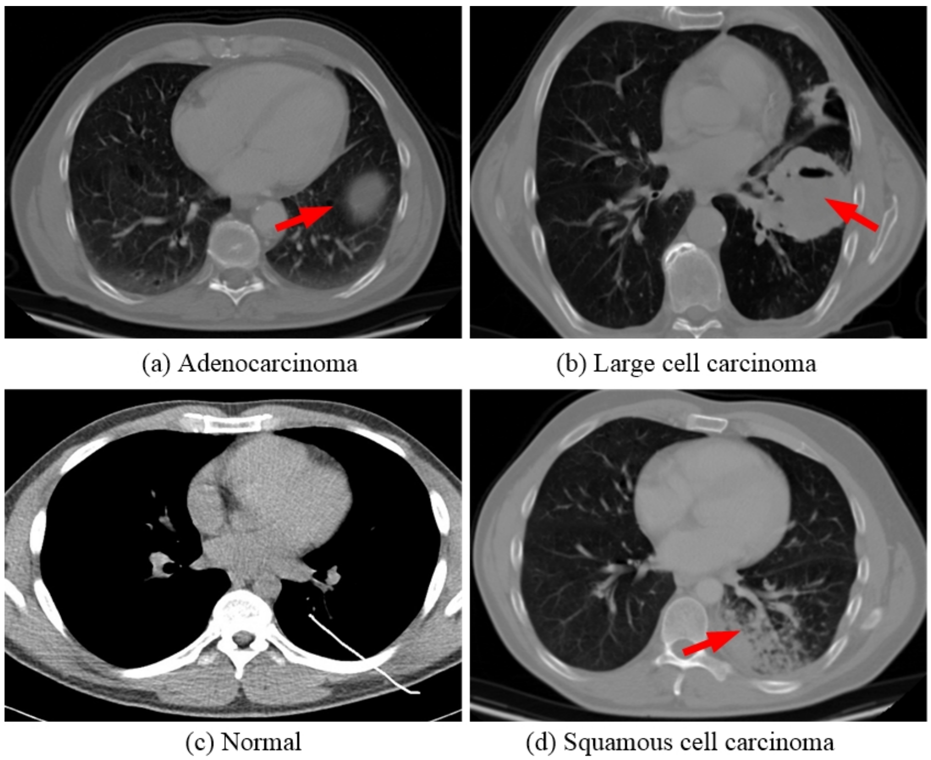

Fig. 2. Sample images for three types of lung cancer and normal CT scan.

2.2 Architecture of CNN

A simple CNN was proposed to diagnose lung cancer. Input layer resized the input images to size 256×256. In the proposed CNN an input convolution layer of 64 filters was followed by a max-pooling layer which was followed by two similar combinations and at the end 2 fully connected layers were present of size 128 and 4. The design of CNN is shown diagramatically in Figure 3.

3 Results of experimentation

Three pre-trained CNN models namely: InceptionV3, MobileNet and VGG19 were trained on training data for 100 epochs after loading imagenet weights.

Table 1. Dataset details.

Class	Number of images	Augmented images
Adenocarcinoma	326	1000
Large cell carcinoma	163	1000
Normal	159	1000
Squamous cell carcinoma	252	1000

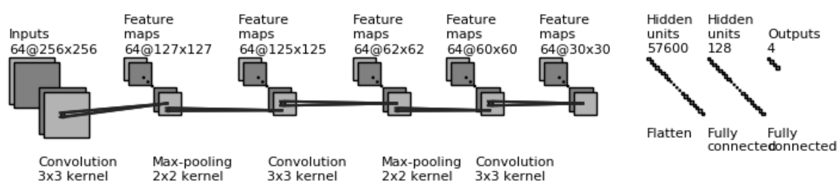

Fig. 3. Proposed CNN architecture.

The dataset was randomly split into 80:10:10 ratio corresponding to train, validation and testing. The proposed CNN was also executed on the same dataset in a similar fashion and accuracy of each model was recorded. The accuracies obtained with different CNN models is shown in Figure 4. It clearly shows best performance was given by proposed CNN. Since pre-trained models were designed for the ILSVRC 1000 class challenge their architecture was huge and complex, hence they were compressed using a meta heursitic based approach of Differential Evolution as explained in next section.

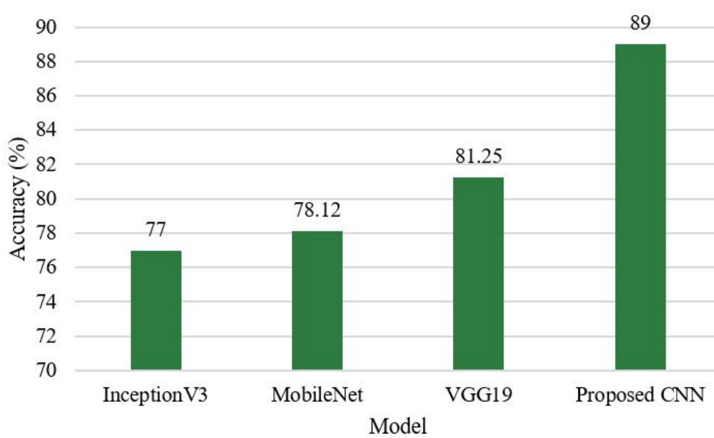

Fig. 4. Bar graph depicting the accuracy of different CNN models.

The activation images of a sample Large cell carcinoma image are shown in Figures 5, 6 and 7. As seen as we go deep in the network more complex features are extracted from the input images to classify them correctly.

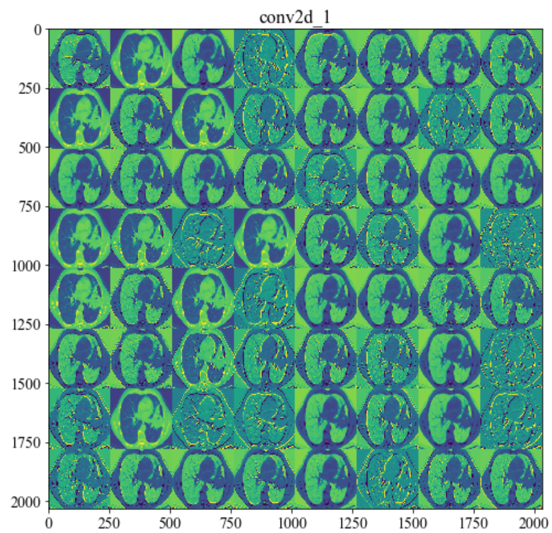

Fig. 5. Activation image for a sample lung cancer image from 1^{st} conv2D layer.

3.1 Machine Learning

Traditional machine learning models such as: SVM, DT, LR, k-NN, LDA, Naive Bayes (NB), RF were utilized to train and test the models. The augmented dataset was split at run time in training and testing in 90:10 ratio. A combination of 3 handcrafted features was used to execute these models. The maximum accuracy was obtained with Random Forest equal to 97% with a combination of all three features. A short description of three features is as given below:

- HSV: These features use the HSV transform of RGB images and then form a histogram of these 3 values.
- Haralick: These features uses GLCM matrix based calculations and help in getting the texture of images.
- Hu-moments: These features are based on moments of objects in images which are not dependent on orientation or size of objects.

The statistics of performance comparison for machine learning models using different features set is presented in Table 2. The performance is also compared in form of Receiver Operating Characteristic Curves (ROC) which shows Area Under Curve (AUC) metrics was best for Random Forest classifier as shown in Figure 8, 9, 10, 11 and 12.

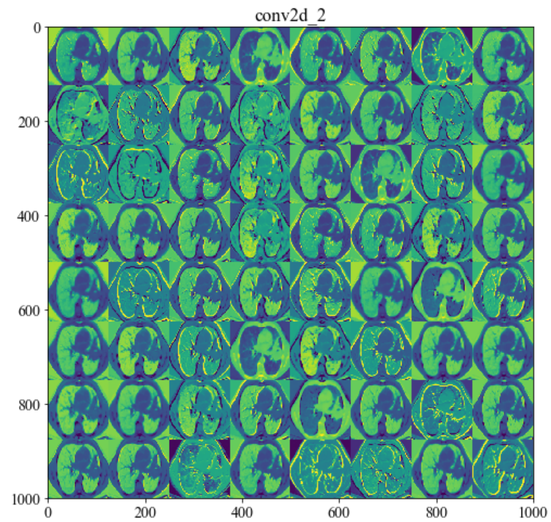

Fig. 6. Activation image for a sample lung cancer image from 2^{nd} conv2D layer.

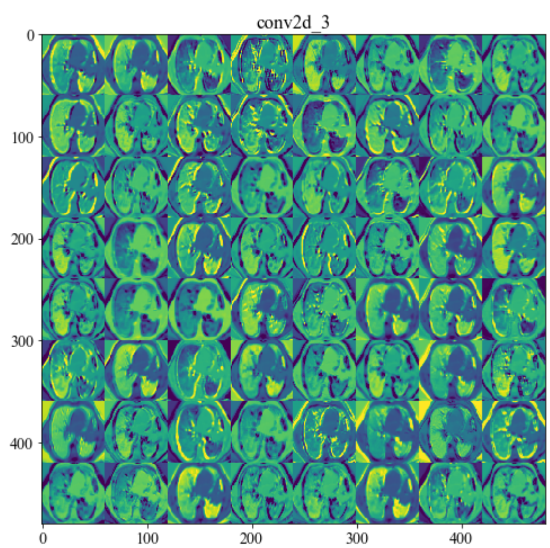

Fig. 7. Activation image for a sample lung cancer image from 3^{rd} conv2D layer.

Table 2. ML models performance statistics with different features.

Features	Model	TP	FP	TN	FN	Accuracy	Precision	Recall	Specificity	F1 score
HSV	SVM	51	249	100	0	0.3775	0.17	1	0.286533	0.290598
	Decision Tree	263	37	92	8	0.8875	0.876667	0.97048	0.713178	0.921191
	Logistic Regression	184	116	22	78	0.515	0.613333	0.70229	0.15942	0.654804
	k-NN	243	57	67	33	0.775	0.81	0.880435	0.540323	0.84375
	LDA	187	113	22	78	0.5225	0.623333	0.70566	0.162963	0.661947
	Naïve Bayes	14	286	100	0	0.285	0.046667	1	0.259067	0.089172
	Random Forest	285	15	93	7	0.945	0.95	0.976027	0.861111	0.962838
Hu-moments	SVM	179	121	12	88	0.4775	0.596667	0.670412	0.090226	0.631393
	Decision Tree	273	27	85	15	0.895	0.91	0.947917	0.758929	0.928571
	Logistic Regression	161	139	15	85	0.44	0.536667	0.654472	0.097403	0.589744
	k-NN	269	31	81	19	0.875	0.896667	0.934028	0.723214	0.914966
	LDA	160	140	19	81	0.4475	0.533333	0.6639	0.119497	0.591497
	Naïve Bayes	164	136	15	85	0.4475	0.546667	0.658635	0.099338	0.59745
	Random Forest	286	14	91	9	0.9425	0.953333	0.969492	0.866667	0.961345
Haralick	SVM	190	110	42	58	0.58	0.633333	0.766129	0.276316	0.693431
	Decision Tree	268	32	80	20	0.87	0.893333	0.930556	0.714286	0.911565
	Logistic Regression	218	82	41	59	0.6475	0.726667	0.787004	0.333333	0.755633
	k-NN	261	39	76	24	0.8425	0.87	0.915789	0.66087	0.892308
	LDA	225	75	34	66	0.6475	0.75	0.773196	0.311927	0.761421
	Naïve Bayes	141	159	79	21	0.55	0.47	0.87037	0.331933	0.61039
	Random Forest	288	12	86	14	0.935	0.96	0.953642	0.877551	0.956811
HSV, Haralick	SVM	85	215	100	0	0.4625	0.283333	1	0.31746	0.441558
	Decision Tree	277	23	86	14	0.9075	0.923333	0.95189	0.788991	0.937394
	Logistic Regression	212	88	43	57	0.6375	0.706667	0.788104	0.328244	0.745167
	k-NN	267	33	76	24	0.8575	0.89	0.917526	0.697248	0.903553
	LDA	229	71	48	52	0.6925	0.763333	0.814947	0.403361	0.788296
	Naïve Bayes	14	286	100	0	0.285	0.046667	1	0.259067	0.089172
	Random Forest	289	11	93	7	0.955	0.963333	0.976351	0.894231	0.969799
HSV, Haralick, Hu-moments	SVM	79	221	100	0	0.4475	0.263333	1	0.311526	0.416887
	Decision Tree	282	18	86	14	0.92	0.94	0.952703	0.826923	0.946309
	Logistic Regression	222	78	44	56	0.665	0.74	0.798561	0.360656	0.768166
	k-NN	278	22	81	19	0.8975	0.926667	0.936027	0.786408	0.931323
	LDA	227	73	56	44	0.7075	0.756667	0.837638	0.434109	0.795096
	Naïve Bayes	14	286	100	0	0.285	0.046667	1	0.259067	0.089172
	Random Forest	293	7	95	5	**0.97**	0.976667	0.983221	0.931373	0.979933

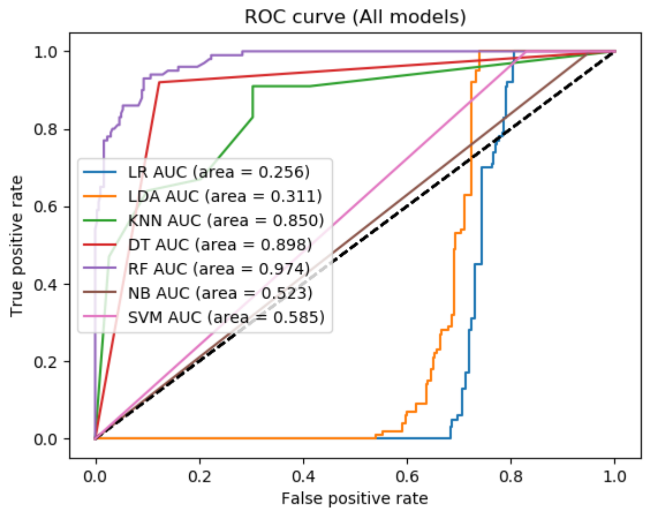

Fig. 8. ROC curve for ML models with HSV features.

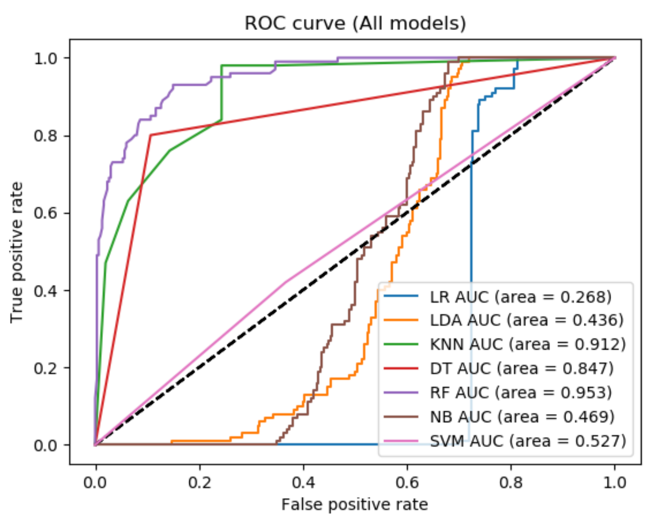

Fig. 9. ROC curve for ML models with Haralick features.

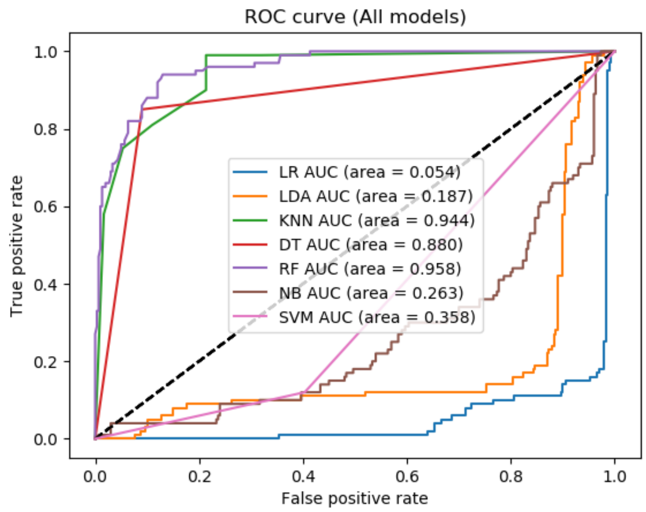

Fig. 10. ROC curve for ML models with Hu moments features.

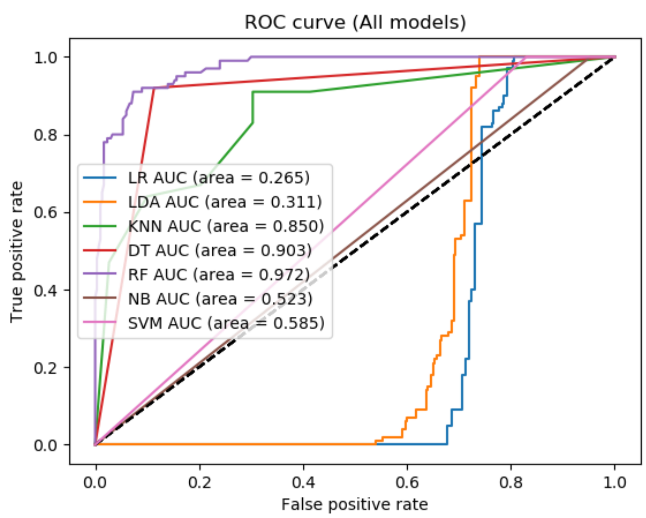

Fig. 11. ROC curve for ML models with HSV and Haralick features.

4 Compression of CNN

As explained in previous section the pre-trained models possess a big and complex architecture, hence they were compressed using DE algorithm. The original

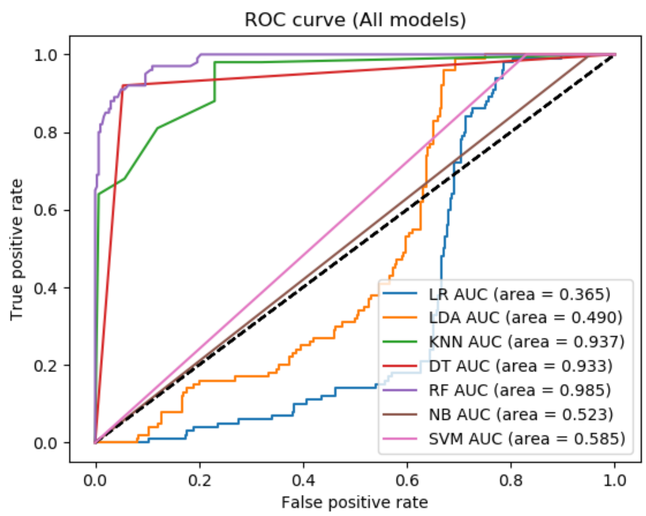

Fig. 12. ROC curve for ML models with HSV, Haralick and Hu-moments features.

size for pre-trained models was 163,710 KB for InceptionV3, 217,642 KB for mobilenet, and 211,421 KB for VGG19. The proposed CNN was only 838 KB. The compression of these models can help to deploy these models on tiny edge devices for developing robotic devices as these devices have limited memory.

4.1 Differential Evolution

Differential Evolution was proposed by Storn and Rice [3] in 1997 based on human evolution over a period of several ages. A pool of vectors is created initially equal to the number of hidden units in any CNN model with random 0 and 1 as the vector elements. The population pool is chosen as consisting of 100 random vectors. In next step iteratively for each target vector 3 different vectors are randomly chosen and a new vector is created using the equation 1. The value of F is chosen as 0.5.

$$v_{new} = v_1 + F \times (v_2 - v_3) \tag{1}$$

The new vector and target vector is then recombined by generating a random value between 0 and 1 for each vector position. If this value is more than a recombination factor chosen as 0.7 then value is picked from new vector else it is chosen from target vector. Thus a entirely new trial vector is created. Then based on fitness criteria a decision is made to retain the old target vector or newly created trial vector in the population pool for next iteration. After either the difference in best population pool vector fitness value does not change by more than 0.00001 or maximum 100 iterations have not been performed the iterations are continued. After exiting from this loop the best vector with best fitness value is returned from the algorithm which helps to decide which nodes to retain and which to remove in compressed model. The fitness function is based on a sum of 4 performance metrics namely precision, recall, F1-score and accuracy as shown in equation 2. The objective is to maximize this value in the compressed model so that compression will not reduce the performance.

$$Maximize(X) = Accuracy + Precision + Recall + F1{-}score \tag{2}$$

The equation for Accuracy is given by:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

The equation for Precision is:

$$P = \frac{TP}{TP + FP} \tag{4}$$

The equation for Recall is:

$$R = \frac{TP}{TP + FN} \tag{5}$$

The formula for F1-score is:
The equation for Precision is:

$$F1{-}score = \frac{2 * P * R}{P + R} \tag{6}$$

4.2 Compression statistics

The compressed model size and accuracy is compared in Table 3.

Table 3. Differential Evolution based compression statistics.

CNN Model	**Initial Size (KB)**	**Final Size (KB)**	**Initial Accuracy (%)**	**Final Accuracy (%)**
InceptionV3	163,710	94,637	77	78.3
VGG19	211,421	107,767	81.25	80.37
MobileNet	217,642	51,543	78.12	96.50
Proposed Model	29,123	17,257	89	87.8

As seen the MobileNet compression increased the accuracy by 17.38%. This is due to the fact the fitness function used performance metrics for checking the fitness of compressed model design. For other models also the performance was not impacted much due to compression.

4.3 Discussion

Inspired by good results on MobileNet the compressed model was compressed one more time with space reducing to 11,134 KB and accuracy still remaining high equal to 95.5%. In yet another effort to compress the model the size was reduced to 2,743 KB but the performance of model dropped below original MobileNet accuracy. The model performance was enhanced as in each iteration the fitness function was trying to find the nodes which were hindering in accuracy and dropping them which produced good performance of the compressed model.

5 Conclusion

Lung cancer is a deadly disease and the diagnosis of the type of lung cancer is very important for the corrective treatment in a timely fashion. This research focuses on classification of type of lung cancer using transfer learning with three pre-trained CNN models and learning from scratch on a proposed CNN. The proposed CNN gave best performance of 89%. However mobilenet accuracy could be boosted even to 96.5% by compressing its size using Differential Evolution. The machine learning methods were also tested to find the best model for classication of lung cancer CT scan images. Random Forest classifier was found to give the best accuracy of 97% using Hu-moments, Haralick and HSV histogram as features. The study also explores usage of Differential Evolution for the compression of heavy CNN models. This study can be extended in future for the usage of other pre-trained models for other human diseases.

References

1. Lauren G Collins, Christopher Haines, Robert Perkel, and Robert E Enck. Lung cancer: diagnosis and management. *American family physician*, 75(1):56–63, 2007.
2. Neeraj Gupta, Mahdi Khosravy, Nilesh Patel, Nilanjan Dey, Saurabh Gupta, Hemant Darbari, and Rubén González Crespo. Economic data analytic ai technique on iot edge devices for health monitoring of agriculture machines. *Applied Intelligence*, 50(11):3990–4016, 2020.
3. Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.
4. Mohit Agarwal, Abhishek Singh, Siddhartha Arjaria, Amit Sinha, and Suneet Gupta. Toled: Tomato leaf disease detection using convolution neural network. *Procedia Computer Science*, 167:293–301, 2020.
5. Mohit Agarwal, Suneet Kr Gupta, and KK Biswas. Grape disease identification using convolution neural network. In *2019 23rd International Computer Science and Engineering Conference (ICSEC)*, pages 224–229. IEEE, 2019.
6. Mohit Agarwal, Rohit Kumar Kaliyar, Gaurav Singal, and Suneet Kr Gupta. Fcnn-lda: A faster convolution neural network model for leaf disease identification on apple's leaf dataset. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, pages 246–251. IEEE, 2019.

7. Mohit Agarwal, Amit Sinha, Suneet Kr Gupta, Diganta Mishra, and Rahul Mishra. Potato crop disease classification using convolutional neural network. In *Smart Systems and IoT: Innovations in Computing*, pages 391–400. Springer, 2020.
8. Mohit Agarwal, Vijay Kumar Bohat, Mohd Dilshad Ansari, Amit Sinha, Suneet Kr Gupta, and Deepak Garg. A convolution neural network based approach to detect the disease in corn crop. In *2019 IEEE 9th international conference on advanced computing (IACC)*, pages 176–181. IEEE, 2019.
9. Mohit Agarwal, Suneet Kr Gupta, and KK Biswas. Development of efficient cnn model for tomato crop disease identification. *Sustainable Computing: Informatics and Systems*, 28:100407, 2020.
10. Mohit Agarwal, Suneet Gupta, and KK Biswas. A new conv2d model with modified relu activation function for identification of disease type and severity in cucumber plant. *Sustainable Computing: Informatics and Systems*, 30:100473, 2021.
11. Mohit Agarwal, Suneet Kr Gupta, and KK Biswas. A compressed and accelerated segnet for plant leaf disease segmentation: A differential evolution based approach. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 272–284. Springer, 2021.
12. Mohit Agarwal, Suneet Kr Gupta, and KK Biswas. Plant leaf disease segmentation using compressed unet architecture. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 9–14. Springer, 2021.
13. Mohit Agarwal, Suneet K Gupta, and KK Biswas. Development of a compressed fcn architecture for semantic segmentation using particle swarm optimization. *Neural Computing and Applications*, pages 1–14, 2023.
14. Mohit Agarwal, Luca Saba, Suneet K Gupta, Amer M Johri, Narendra N Khanna, Sophie Mavrogeni, John R Laird, Gyan Pareek, Martin Miner, Petros P Sfikakis, et al. Wilson disease tissue classification and characterization using seven artificial intelligence models embedded with 3d optimization paradigm on a weak training brain magnetic resonance imaging datasets: a supercomputer application. *Medical & Biological Engineering & Computing*, 59(3):511–533, 2021.
15. Mohit Agarwal, Luca Saba, Suneet K Gupta, Alessandro Carriero, Zeno Falaschi, Alessio Paschè, Pietro Danna, Ayman El-Baz, Subbaram Naidu, and Jasjit S Suri. A novel block imaging technique using nine artificial intelligence models for covid-19 disease classification, characterization and severity measurement in lung computed tomography scans on an italian cohort. *Journal of Medical Systems*, 45(3):1–30, 2021.
16. Luca Saba, Mohit Agarwal, Anubhav Patrick, Anudeep Puvvula, Suneet K Gupta, Alessandro Carriero, John R Laird, George D Kitas, Amer M Johri, Antonella Balestrieri, et al. Six artificial intelligence paradigms for tissue characterisation and classification of non-covid-19 pneumonia against covid-19 pneumonia in computed tomography lungs. *International journal of computer assisted radiology and surgery*, 16(3):423–434, 2021.
17. Mohit Agarwal, Sushant Agarwal, Luca Saba, Gian Luca Chabert, Suneet Gupta, Alessandro Carriero, Alessio Pasche, Pietro Danna, Armin Mehmedovic, Gavino Faa, et al. Eight pruning deep learning models for low storage and high-speed covid-19 computed tomography lung segmentation and heatmap-based lesion localization: A multicenter study using covlias 2.0. *Computers in Biology and Medicine*, page 105571, 2022.
18. Mohit Agarwal, Suneet K Gupta, Mainak Biswas, and Deepak Garg. Compression and acceleration of convolution neural network: a genetic algorithm based approach. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–11, 2022.

19. Mohit Agarwal, Suneet Kumar Gupta, Deepak Garg, and Dilbag Singh. A novel compressed and accelerated convolution neural network for covid-19 disease classification: A genetic algorithm based approach. In *International Advanced Computing Conference*, pages 99–111. Springer, 2021.
20. Mohit Agarwal, Suneet Kumar Gupta, Deepak Garg, and Mohammad Monirujjaman Khan. A partcle swarm optimization based approach for filter pruning in convolution neural network for tomato leaf disease classification. In *International Advanced Computing Conference*, pages 646–659. Springer, 2021.
21. Hikmat Yar, Tanveer Hussain, Mohit Agarwal, Zulfiqar Ahmad Khan, Suneet Kumar Gupta, and Sung Wook Baik. Optimized dual fire attention network and medium-scale fire classification benchmark. *IEEE Transactions on Image Processing*, 2022.
22. Sanagala S Skandha, Mohit Agarwal, Kumar Utkarsh, Suneet K Gupta, Vijaya K Koppula, and Jasjit S Suri. A novel genetic algorithm-based approach for compression and acceleration of deep learning convolution neural network: an application in computer tomography lung cancer data. *Neural Computing and Applications*, pages 1–23, 2022.
23. Mohit Agarwal, Rohit Kr Kaliyar, and Suneet Kr Gupta. Differential evolution based compression of cnn for apple fruit disease classification. In *2022 International Conference on Inventive Computation Technologies (ICICT)*, pages 76–82. IEEE, 2022.
24. Tafadzwa L Chaunzwa, Ahmed Hosny, Yiwen Xu, Andrea Shafer, Nancy Diao, Michael Lanuti, David C Christiani, Raymond H Mak, and Hugo JWL Aerts. Deep learning classification of lung cancer histology using ct images. *Scientific reports*, 11(1):5471, 2021.
25. Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE transactions on cybernetics*, 50(9):3950–3962, 2019.
26. Huan Yang, Lili Chen, Zhiqiang Cheng, Minglei Yang, Jianbo Wang, Chenghao Lin, Yuefeng Wang, Leilei Huang, Yangshan Chen, Sui Peng, et al. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study. *BMC medicine*, 19:1–14, 2021.
27. Mohamed Hany. Lung cancer dataset. *https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images*, 2020.