RESEARCH
Volume 09 Issue 01
ISSN: 2639-5274, 2096-5168

Effective Deep Ensemble Methods for the Detection of Lung Cancer with SMOTE

Khadija Kanwal[1,*], Muhammad Abubakar[2], Aiza Shabbir[1], Laiba Rehman[1], Hareem Ayesha[1], Humera Batool Gill[1], Afshan Almas[1], Hina Ali[3]

[1]Institute of Computer Science and Information Technology, The Women University, Multan 60000, Pakistan (khadijakanwal@wum.edu.pk,aiza.6322@wum.edu.pk,hareem.ayesha@wum.edu.pk,laiba.rehman@wum.edu.pk humerabatool.gill@wum.edu.pk,afshan.6024@wum.edu.pk)

[2]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan, (aaq.kfu@gmail.com)

[3]Associate Professor Department of Economics, The Women University, Multan 60000, Pakistan. hinaali@wum.edu.pk

Corresponding Author: Khadija Kanwal (khadijakanwal@wum.edu.pk)

Abstract: *Lung cancer is a death-causing disease that indicates the existence of pulmonary nodules in the lung. It is typically caused by increased cancer cells in the lung. Lung nodule detection has a major role in screening and detecting lung cancer images in computed tomography (CT) scans. Lung cancer deserves more attention in human disease investigation methods due to its substantial influence on both males and females, leading to around five million deaths each year. The patient's survival ratio is increased as lung cancer detects at an early stage. This work presents a novel and unique transfer learning model based on deep ensemble techniques for efficient lung cancer detection. A chest CT scan dataset is needed to detect lung cancer. The presented technique diagnoses lung cancer efficiently with high accuracy. The primary objective of this research is to detect lung cancer at an early stage by evaluating the efficacy of deep ensemble classification models. The proposed model achieves outstanding results when compared with state-of-the-art methods to classify lung cancer. The experimental results show significant performance of the presented model.*

Keywords: *Lung Cancer, object detection, computed tomography, SMOTE, medical imaging, Healthcare*

1. **Introduction**

Lung cancer is the most common type of cancer in the United States, as well as the main cause of cancer mortality and morbidity [1]. As a result, numerous countries are formulating ways to facilitate the timely detection of lung cancer. Scientists are developing automated solutions to assist doctors in their work, enhance diagnostic accuracy by eliminating subjectivity, expedite analysis, and reduce expenses related to medical treatment, in anticipation of a rise in preventive and early detection measures [2,3]. Due to the lack of symptoms, early-stage lung cancer often goes undetected, resulting in the majority of newly discovered cases being at an advanced stage. Unfortunately, advanced-stage lung cancer has a low five year survival rate and a poor prognosis. Early detection dramatically enhances the survival rate of lung cancer [4]. Benign and malignant pulmonary nodules are cellular growths in the lung that can be either non-cancerous or cancerous. Early detection significantly impacts the prognosis of malignant pulmonary nodules [5].

Computed tomography (CT) imaging is the optimal application for examining lung problems. However, CT scan tests carry a notable risk of producing false positive findings, and the radiation they emit has the potential to induce cancer. Low-dose CT utilizes significantly reduced radiation exposure compared to standard dose CT scans. The results indicate that there is no noticeable distinction in the sensitivity of image detection [6]. CT scanning is an innovative imaging technique that generates extraordinarily detailed images of the injured body region, including clear views of the internal soft tissues and organs, by utilizing powerful scanners [7]. A variety of diagnostic techniques have been employed in the span of lung cancer development. Chest CT scans were specifically chosen for their robustness in accurately measuring tumor size and their minimal noise levels.

The development of deep learning in medical imaging is that it can do in-depth identification by obtaining the most significant features during training data. The model is resistant to modification since it can extract nodule features from CT scans with different parameter settings. The training set has a lot of variation, which means that invariant features might be able to be automatically learned from malignant nodules. Because there are no artificial qualities, the network must rely on the ground truth that has been supplied to determine the relationship that exists between the lung and features on its own. This study has made the following contributions:

- A novel automated deep ensemble approach is developed, using a transfer learning mechanism, to achieve highly accurate detection of lung cancer.
- The issue of an imbalanced dataset was addressed, and it was demonstrated that the implementation of an oversampling technique such as SMOTE can substantially improve the accuracy of detecting lung cancer through CT scans.
- To evaluate the robustness and effectiveness of the proposed ensemble method, we applied cross-dataset experiments utilizing a 10-fold strategy and utilized a number of individual deep learning models.

2. Literature Review

The study [8] presented a technique for detecting nodules from normal lung anatomical components. The technique retrieves gray-level, statistical, and geometric characteristics. LDA serves as a classifier and was employed for effective thresholding in segmentation. The approach produces outcomes with an accuracy rate of 84%, 97.14% sensitivity, and 53.33% specificity. The model's accuracy remains suboptimal even after successfully identifying the cancer nodule. Instead of utilizing machine learning methods for categorization, conventional segmentation techniques were used. Hence, there is no possibility for enhancement through the integration of any current activities with their novel model. The primary finding of the study [9] was the establishment of a systematic approach to classify and detect various types of lung cancer. Scientists have devised a novel approach to detecting lung cancer by employing machine learning methodologies. The method consists of two components: discrete cosine transform (DCT) and patch basis LBP (local binary pattern) fusion and feature extraction. The classification of various texture types in a collection of chest CT scan images was accomplished using machine learning methods, specifically KNN and SVM. The suggested solution outperforms the latest state-of-the-art methods in terms of accuracy, achieving a 93% accuracy rate for support vector machines and a 91% accuracy rate for K-nearest neighbors.

The efficacy of all lung cancer detection systems relies on their ability to distinguish between malignant and benign tissues. This technique was utilized during the phase of reducing false positives, which is a crucial component of lung cancer detection. The primary aim of this study was to present a novel method utilizing 3D CNN that exhibits remarkable sensitivity and a low rate of false positives in identifying lung cancer lesions. They were able to achieve a precision rate of 91.23% with just 3.99 instances of false positives per scan by using a cutting-edge fusion method. Through the construction of an innovative fusion technique that leverages the expertise of the classifiers, they achieved a notable improvement in accuracy while simultaneously reducing the false-positive rate [10].

Lung cancer is considered one of the most severe diseases on a global scale. Timely identification and an accurate prognosis can lead to an enhanced survival rate in cases of lung cancer. CT scan images were obtained from multiple hospitals and databases. A novel method is currently being developed and has been suggested to determine the severity of the cancer. The utilization of Gabor filters for image enhancement results in the most advantageous outcomes in the pre-processing phase. The three GLCM attributes obtained from the emerging regions of interest (ROI) are area, perimeter, and eccentricity. The nodules were evaluated based on these three characteristics to determine the magnitude of lung cancer. The size of the nodule determines the various stages of the tumor. Additionally, SVM was recognized as a particularly compelling method for categorization. An approach for controlling the size of the nodule was combined with generalization controls. To ensure generalization control in classification tasks, the weights of vectors in a canonical structure were aligned to maximize the margin [11]. A deep learning system was employed in the Study [12] to investigate the incidence of pneumonia and lung cancer. To evaluate the problem, two distinct deep learning techniques were proposed. (i) The initial deep learning strategy to classify chest X-ray pictures into two unique categories such as normal and pneumonia which were to be implemented using a modified version of AlexNet (MAN). SCM was used to implement classification in the MAN, and Softmax was used as a benchmark to assess their performance. By using a variety of pre-trained deep learning approaches, such as AlexNet, VGG16, VGG19, and ResNet50, the model's performance was further confirmed. (ii) To improve lung cancer classification accuracy, the second deep learning operation incorporates manually created and acquired characteristics within the MAN. To improve the feature vector, this work combines the methods of principal component analysis (PCA) feature selection and serial fusion.

In order to identify and classify nodules based on their level of malignancy, the proposed method makes use of three-dimensional CT scan images of the lungs.

Two encoder-decoder models, especially U-Net with MixNet and deep 3D CNN, were designed to collect knowledge about the properties of the lung nodule and detect it, respectively. The models used the three-dimensional features of the lung CT data. To classify nodules, researchers have developed a new method that combines the gradient boosting machine with a 3D mixed-link network. On a dataset of 1200 images obtained from LIDC-IDRI, the proposed approach was evaluated using statistical metrics and manual contouring by radiologists. There were 3,250 nodules included in this data collection. Lung nodules in LIDC-IDRI are equally likely to be cancerous or benign [13].

The data from CT scans were analyzed computationally, using the training set of over 3,000 CT scans, the algorithm can detect cancer without lung nodule annotations. The system uses sequential segments and multiple features to determine whether or not a nodule is cancerous. In addition, it employs a spatial pyramid for pinpointing nodule identification across zoom levels. The AUC of 0.858 demonstrates that the algorithm's effectiveness in predicting a lung cancer status based on a lung CT scan was outstanding (78.2%). Based on the challenge datasets, their first framework finished 16th out of 72 teams, or in the top one percent [14]. The paper [15 discusses the application of deep learning on a separate test set. Based on the results of this work utilizing deep learning techniques for CT lung cancer screening holds the capacity to substantially diminish the occurrence of false-positive results. An analysis was conducted on a selection of cutting-edge deep learning algorithms and designs that have been suggested as CAD systems with the capability to detect lung cancer. There were two distinct types: (1) nodule identification systems examine the initial CT scan for potential nodules, and (2) false positive reduction techniques classify a group of possible nodules into either benign or malignant tumors. An evaluation of the efficacy of the different strategies was conducted while showcasing their essential attributes. Additionally, the CT lung datasets that apply to research are examined. A comparative analysis of the various strategies is illustrated and discussed.

3. Proposed Methodology

The proposed methodology comprises dataset information, preprocessing, the oversampling technique SMOTE, and the proposed ensemble model. The flowchart of the presented methodology is described in Figure 1. The detail of each step is discussed in the following:

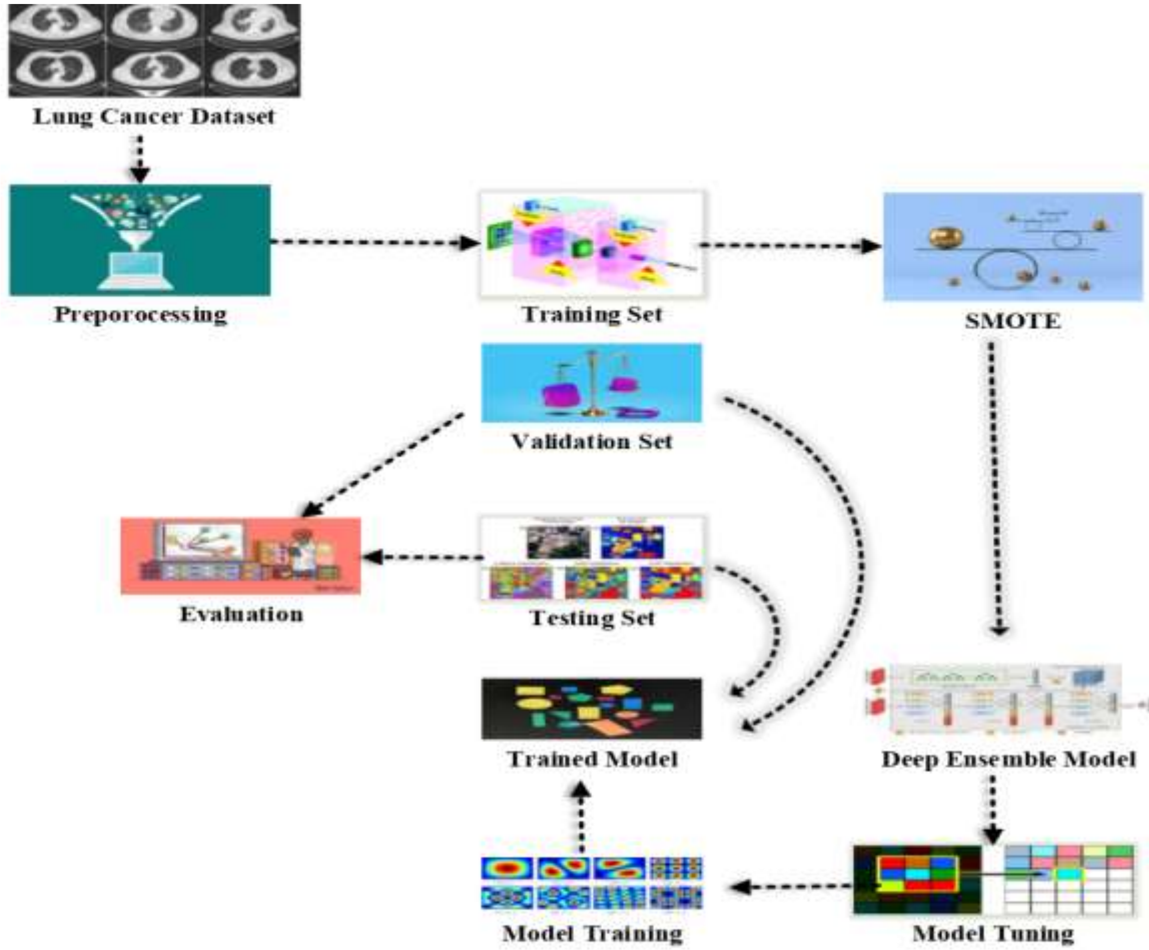


Figure 1: Flowchart of the presented methodology

3.1 Dataset information and preprocessing

The CT scan data is divided into three distinct forms of chest cancer namely squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. There is also one subfolder with normal cells. The middle region of the lung, specifically the major airway branches or the point where the larger bronchi and trachea converge, is where squamous cell lung cancer is most frequently observed. About 30\% of non-small cell lung cancers are squamous cell lung cancers, which are frequently associated with tobacco use. Aggressive splenic large-cell undifferentiated carcinoma is a fast-growing malignancy that can appear anywhere in the lung. Roughly 10–15\% of cases of non-small-cell lung cancer are caused by this sub-type. The most prevalent subtype of lung cancer, accounting for around 30\% of all cases that are reported and 40\% of non-small cell lung cancer cases, is lung adenocarcinoma. In addition to other places, the colon, prostate, and breast are frequently affected by adenocarcinoma [16].

Preprocessing is crucial in medical imaging since it enhances the accuracy and simplifies the structure of a model at the same time. CT scans are large images that need to undergo various preprocessing to extract significant information for the classification of lung nodules. Due to its significant memory requirements, the system cannot immediately process CT images for analysis. The images were uniformly reduced to meet the requirements of the deep learning models [17].

3.2 SMOTE

The existence of imbalanced data has a substantial influence on the quality of the model that is developed. Deep or machine learning models trained on imbalanced data would struggle to accurately learn the decision boundary when there is a lack of instances from the minority class. The SMOTE is a commonly employed approach to tackle this issue. The objective of this technique is to provide novel examples for the minority class rather than copying the current ones. The SMOTE algorithm produces a distinct instance along the boundary between the samples in the feature space by choosing samples that are nearest to the feature. A random sample is first chosen from the minority example. The closest neighbors of the provided sample are subsequently identified. The position of this synthetic sample is likewise randomly decided, and it falls between the positions of the two original samples. Continuing the process until the minority class reaches the same proportion as the majority class is enough [18].

3.3 Proposed model

The proposed model was developed by integrating the VGG-19 and CNN layers. In the beginning, we built the VGG-16 model and declared that we did not intend to include the top layer. Additionally, we ensure that the input size for both models remains consistent at $224 \times 224 \times 3$. Following the VGG-19 model, the ensemble model includes two Conv2D layers, which come before pooling, batch normalization, and dropout layers. Consequently, three Conv2D layers, two dropout layers, three pooling layers, and one batch normalization layer are incorporated. During the final stage, we compress the layers and incorporate two dense layers, utilizing the "Relu function" as the activation function. Additionally, we include a final dense layer employing the "softmax" function for classification purposes. The subsequent paragraphs explore the execution of each layer and provide a comprehensive analysis of its intricacies. The architecture of the proposed model for lung cancer detection is depicted in Figure 2.

Convolutional neural networks (CNNs), which are a type of advanced neural network, possess the capability to recognize and classify specific components present within images. They are of significant importance in the visual image analysis process. The functions executed by these applications are diverse and encompass the analysis of medical images, the comprehension of human language, and identification. To facilitate analysis, a convolution technique is employed to separate and isolate the image's numerous features. The feature extraction network is

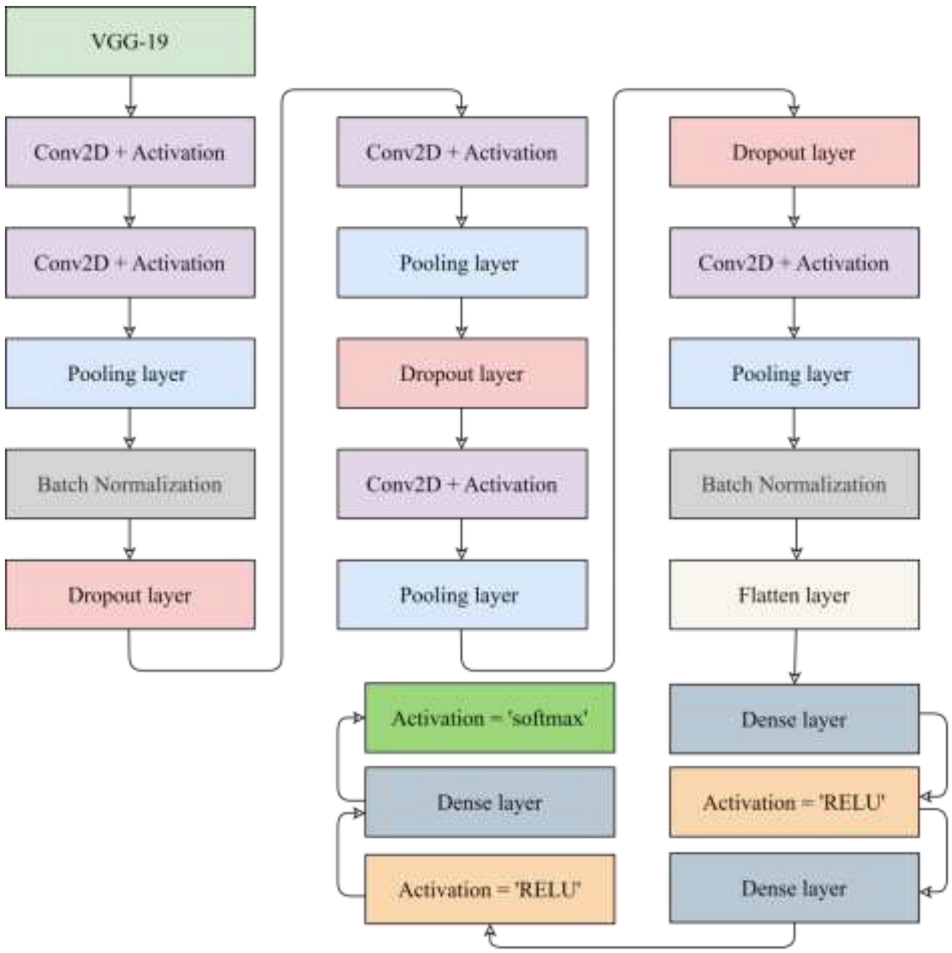


Figure 2: Architecture of proposed model for lung cancer detection

Comprised of numerous layer pairings between convolutional and pooling methods. A densely connected layer that infers the category to which an image belongs using the data collected in previous stages and the output of the convolution process A reduction in the quantity of features present in a dataset will ensue from the utilization of this CNN model for feature extraction. New features are generated through the process of condensing the existing qualities that were included in the initial set of features into a reduced number of more significant categories. Comprising the CNN are three discrete layers, and stacking these layers in the specified sequence, CNN architecture can be generated. In addition to the three layers that were previously mentioned, there are two additional critical components known as the dropout layer and the activation function.

- Convolutional layer: The primary function of the neural network is convolution. The network is primarily responsible for the computational demands. This layer computes the dot product between the adjustable parameters of the kernel and one of the matrices that constrain the receptive field. While the image is relatively smaller, the kernel exhibits a greater level of complexity [19].
- Pooling layer: The pooling layer replaces the network output with a summary statistic calculated from nearby outputs at fixed positions. This allows for a smaller spatial range of the representation, which in turn reduces the computational load and the necessary weight. The pooling operation treats each segment of the representation separately.
- Fully connected layer: In FC, each input is coupled to every neuron, allowing the FC to process a condensed input. In CNN architecture, FC layers are commonly placed at the last stage to optimize objectives such as class ratings.
- Dropout layer: CNN relies heavily on dropout layers during training because they greatly reduce the likelihood of overfitting the training dataset. When a significant percentage of a learner's initial training data is lost, the learning process suffers. On the hidden layer, a dropout layer is utilized to disable some of the neurons and eliminate features from the input vector.
- Activation function: The study used two functions, RelU and Softmax. RelU is a piecewise linear function that outputs 0 when the input is negative or zero and outputs 1 when the input is positive. The Softmax

activation function is utilized to compute the probabilities. Softmax enables Convolutional Neural Networks (CNNs) to generate a precise probability distribution over the different classes. This is crucial as it amplifies CNN's predictive capabilities [20].

Transfer learning gives the ability to apply previously developed and trained architectures to the solution of a new problem domain in a field that is connected to it. Transfer learning and data augmentation are two methods that can be used to overcome the challenge of training a CNN model with only a small number of medical images. These methods allow the problem to be solved. Transfer learning is essential in the field of medicine because there are insufficient datasets that are open to the public and because collecting and interpreting these datasets requires a large amount of effort and expertise from qualified radiologists. It comes at a high cost as well. Training a model using a technique called deep learning demands a significant amount of memory as well as computer resources. Transfer learning (TL) is a method that improves the performance of the suggested architecture by removing particular features from the data and then applying those features to a more condensed dataset that is associated with the problem [21].

4. Results and Discussion

The research is carried out on computers running the Windows operating system, including a Core i7 processor of the 10th generation and 32 gigabytes of random access memory. After obtaining the dataset from Kaggle, it is then subjected to preprocessing and oversampling before two types of deep transfer learning models: individual and ensemble. We use a batch size of 16, the cross entropy-loss function, and 20 epochs in the implementation of the models that are utilized in this study.

4.1 Evaluation metrics

We utilized four evaluation measures for the performance assessment of deep learning. The evaluation measures encompass precision, F1 score, accuracy, and recall [22]. The performance measurements may be calculated using the data from the confusion matrix. The confusion matrix is a widely used assessment matrix in binary or multi classification scenarios. The confusion matrix consists of four values. True positive (TP): instances that are correctly identified as positive by the estimation. A true negative (TN) refers to instances where the negative class is correctly identified as negative. False Positive (FP): instances that are classified as negative but were predicted to be positive. A false negative (FN) refers to situations that are expected to be positive but are incorrectly identified as negative.

$$\text{Accuracy} = TP + TN / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$AUC = 0.5 * (TPR + TNR)$$

4.2 Performance of deep transfer learning

Table 1 presents the results of the study conducted on the testing data. The VGG-16 achieved 92.09% accuracy, a recall of 92.76%, and an AUC score of 94.66%. The ResNet-50 model achieved comparable performance to the VGG-16 model, with an AUC score of 93.86% and an accuracy of 90.75%. The MobileNet-V2 model achieved a f1 score of 88.56% and an accuracy of 87.65%. The VGG-19 model attained a recall rate of 94.34% and an accuracy rate of 94.39%. The CNN model attained an accuracy of 95.57%, a recall of 95.60%, and an AUC score of 98.06%. All deep models yielded comparable results. The model we presented achieved exceptional results. The accuracy of our model reached 96.67%, with a f1 score of 96.89% and an AUC score of 98.72%.

Table 1: Performance of deep transfer learning using Testing data

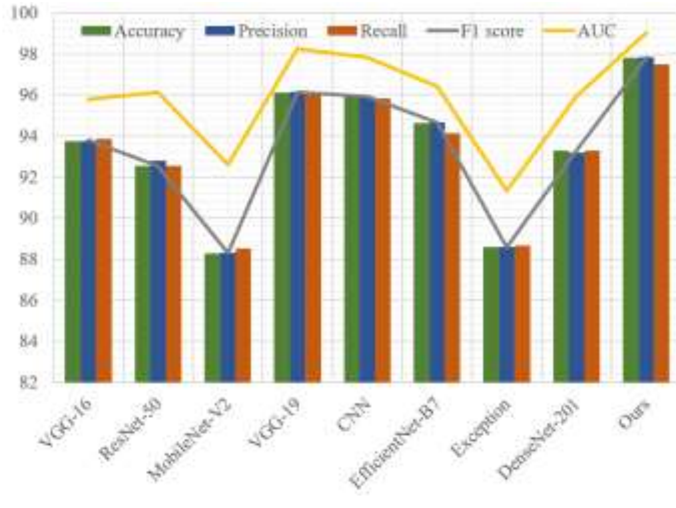
Models	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
VGG-16	92.09	91.76	92.76	91.12	94.66
ResNet-50	90.75	90.75	91.65	90.57	93.86
MobileNet-V2	87.65	87.47	87.35	88.56	90.64
VGG-19	94.34	94.34	94.39	94.32	96.17
CNN	95.57	95.58	95.46	95.60	98.06
EfficientNet-B7	93.03	93.07	93.07	93.03	96.23
Exception	85.80	85.68	85.92	85.80	89.37

DenseNet-201	91.12	91.64	91.12	91.19	94.50
Proposed	96.67	96.68	96.62	96.89	98.72

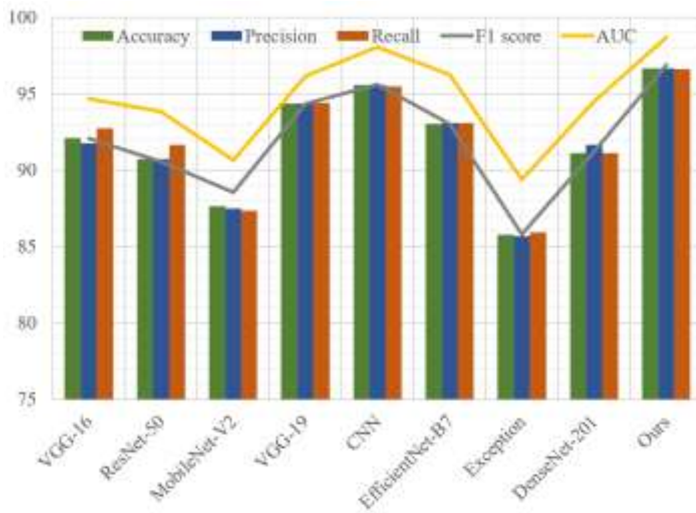
Table 2 displays the results obtained from the study that was carried out, making use of the validation data. The VGG-16 model performed well, having an accuracy of 93.76%, a recall of 93.86%, and an AUC score of 95.76%. When compared to the VGG-16 model, the ResNet-50 model obtained a greater level of accuracy, with a score of 92.54% as well as a superior AUC score of 96.14%. The MobileNet-V2 model performed admirably, with an f1 score of 88.28% and an accuracy of 88.29%. Both the recall rate and the accuracy rate for the VGG-19 model were measured at 96.13%. AUC score of 97.83%, recall rate of 95.83%, and accuracy of 95.92% were all achieved using the CNN model. The findings that each deep model produced were comparable to one another. The novel method that we used produced quite impressive results. The accuracy of the model came out to 97.82% after it earned an AUC score of 99.03% and an f1 score of 97.84%, respectively. Figure 3 shows the performance of the proposed model on test and validation data.

Table 2: Performance of deep transfer learning using Validation data

Models	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
VGG-16	93.76	93.76	93.85	93.79	95.76
ResNet-50	92.54	92.78	92.55	92.54	96.14
MobileNet-V2	88.29	88.30	88.53	88.28	92.61
VGG-19	96.13	96.16	96.13	96.13	98.25
CNN	95.92	95.90	95.83	95.92	97.83
EfficientNet-B7	94.63	94.67	94.12	94.64	96.40
Exception	88.58	88.58	88.65	88.59	91.34
DenseNet-201	93.29	93.18	93.30	93.26	95.91
Proposed	97.82	97.84	97.49	97.84	99.03



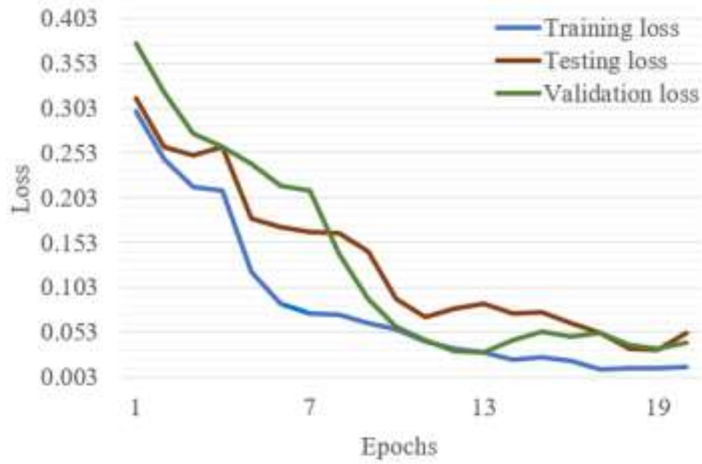
(a)



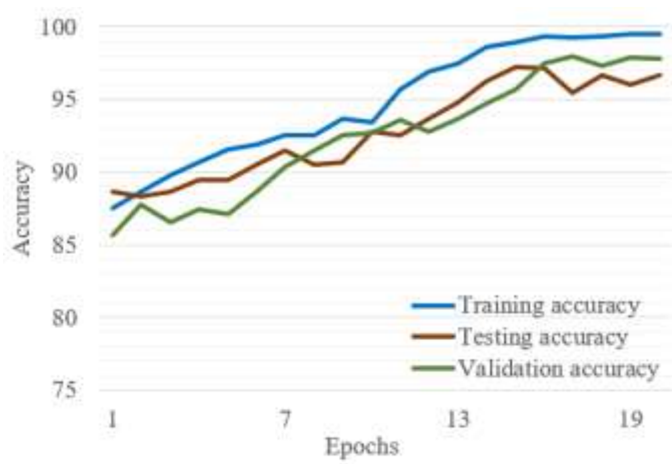
(b)

Figure 3: Performance of proposed model on test and validation data.

Figure 4 displays the training, validation, and test accuracy of the proposed ensemble model. Figure 4(a) illustrates the train, test-and-validation accuracy of the proposed model, which combines the EfficientNet and VGG-16 architectures. The training accuracy was observed to be 99.54%. Fluctuations are noticed starting from epoch 5 and continuing afterward. The ensemble of deep learning models was trained for 20 epochs, resulting in exceptional performance. As shown in Figure 4(b), the training loss reduced from 0.298 to 0.015, and the testing loss fell from 0.314 to 0.053 during 20 epochs. The Ensemble's deep learning model utilizes transfer learning by utilizing pre-trained weights and fine tuning to accurately identify lung cancer from CT scans.



(a)



(b)

Figure 4: Learning performance for the proposed model

K-fold cross-validation is a prevalent technique for assessing the efficiency of machine learning models on a specific dataset. The k-fold cross-validation procedure divides a finite dataset into k separate and non-overlapping folds. Each set is individually chosen to serve as the test set. Next, one of the remaining sets is designated as the validation set, while the other two sets are assigned as the training sets. This process is repeated until all possible combinations have been evaluated. The training set is utilized to train the model, while the validation set is employed to evaluate the model's performance for each set of hyperparameters. Table 3 shows the 10 Fold cross dataset evaluation results.

Table 3: 10 Fold cross dataset evaluation

Models	**Testing accuracy**	**Validation accuracy**
VGG-16	91.32 ± 0.055	92.87 ± 0.059
ResNet-50	89.56 ± 0.076	90.34 ± 0.064
MobileNet-V2	87.18 ± 0.067	85.46 ± 0.074
VGG-19	93.87 ± 0.057	95.03 ± 0.046
CNN	95.26 ± 0.076	94.67 ± 0.032
EfficientNet-B7	92.10 ± 0.054	92.53 ± 0.065
Exception	85.64 ± 0.062	87.75 ± 0.070
DenseNet-201	90.56 ± 0.052	92.21 ± 0.061
Proposed	95.93 ± 0.043	96.73 ± 0.024

4.3 Comparative Analysis

The authors assess the effectiveness and robustness of the proposed model by comparing its results with those of other cutting-edge studies. Aggarwal et al.[8] utilized LDA as a classifier and applied it for efficient thresholding in segmentation. The study [9] utilized DCT and employed KNN and SVM models for classification, achieving an accuracy rate of 93%. Moradi et al. [10] used a 3D CNN model to classify lung cancer lesions and achieved an accuracy rate of 91.23%. They accomplished this by employing a state-of-the-art fusion method, resulting in only 3.99 occurrences of false positives per scan. Sang et al. [13] employ deep learning convolutional neural networks and U-Net models to identify early-stage lung cancer. They achieved a level of accuracy of 94%. In Research [14] the authors employed a deep convolutional deepscreener model that achieved an accuracy of 78%. Analyzed computationally, the data from CT scans were examined. The program can accurately identify cancer in CT scans without relying on lung nodule annotations, using a training set consisting of more than 3,000 scans. The preceding literature and Table 4 indicate that previous studies mostly employed CNN and machine learning techniques with lung data for lung cancer detection, although they yielded lower results. Additionally, previous research has not addressed the issue of imbalanced class problems. The study aims to address the issue of imbalance class by employing the SMOTE technique. The proposed model achieves a remarkable accuracy of 97.82% in detecting lung cancer.

Table 4: Comparative results with other studies

Reference	Methods	Data	Accuracy
[1]	LDA	Lung cancer	84
[2]	SVM, KNN, DCT	Lung cancer	93
[3]	3D CNN	Lung cancer	91
[6]	CNN, U-Net	Lung cancer	94
[7]	DeepScreener	Lung cancer	78
[8]	3D CNN	Lung cancer	87
Proposed	VGG-19+CNN	Lung cancer	98

5. Conclusions

The major aim of this study is to propose a robust ensemble model for accurately detecting lung cancer from CT scans using a balanced dataset. We employed a dataset of chest CT scans to detect cases of lung cancer. The dataset is preprocessed because deep learning depends on a fixed-size input for both training and testing applications. The experiments were conducted on individual deep learning models, resulting in 96.13% accuracy and a 96.16% precision score using validation data. The testing data yielded 95.57% accuracy and a 98.06% AUC score while utilizing the CNN model. The deep exception model performed poorly under all constraints. We proposed an ensemble model, as ensemble models improve accuracy and strengthen the framework. The proposed model achieved a testing accuracy of 96.67% and a validation accuracy of 97.82%. In addition to improving the study, the following components should be implemented in the future:

- It would be worthwhile to expand our research on the chest dataset by conducting experiments using chest X-ray imaging data and subsequently comparing it with CT scans.

- Feature selection approaches and advanced augmentation techniques can be employed.
- Machine learning models can be utilized in conjunction with deep learning to optimize performance.

References

1. Sousa, Joana, Tania Pereira, Francisco Silva, Miguel C. Silva, Ana T. Vilares, António Cunha, and Hélder P. Oliveira. "Lung Segmentation in CT Images: A Residual U-Net Approach on a Cross-Cohort Dataset." Applied Sciences 12, no. 4 (2022): 1959.
2. Siegel, Rebecca, Jiemin Ma, Zhaohui Zou, and Ahmedin Jemal. "Cancer statistics, 2014." CA: A Cancer Journal for Clinicians 64, no. 1 (2014): 9-29.
3. Chiang, Tai-An, Ping-Ho Chen, Pei-Fen Wu, Tsu-Nai Wang, Po-Ya Chang, Albert Min-Shan Ko, Ming-Shyan Huang, and Ying-Chin Ko. "Important prognostic factors for the long-term survival of lung cancer subjects in Taiwan." BMC Cancer 8, no. 1 (2008): 1-8.
4. Siegel, Rebecca L., Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. "Cancer statistics, 2021." Ca Cancer J Clin 71, no. 1 (2021): 7-33.
5. Hervier, Baptiste, Jules Russick, Isabelle Cremer, and Vincent Vieillard. "NK cells in the human lungs." Frontiers in Immunology 10 (2019): 1263.
6. Duma, Narjust, Rafael Santana-Davila, and Julian R. Molina. "Non–small cell lung cancer: epidemiology, screening, diagnosis, and treatment." In Mayo Clinic Proceedings, vol. 94, no. 8, pp. 1623-1640. Elsevier, 2019.
7. Makaju, Suren, P. W. C. Prasad, Abeer Alsadoon, A. K. Singh, and A. Elchouemi. "Lung cancer detection using CT scan images." Procedia Computer Science 125 (2018): 107-114.
8. Aggarwal, T., Furqan, A., \& Kalra, K. (2015) “Feature extraction and LDA based classification of lung nodules in chest CT scan images.” 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI), DOI: 10.1109/ICACCI.2015.7275773.
9. Rehman, Amjad, Muhammad Kashif, Ibrahim Abunadi, and Noor Ayesha. "Lung cancer detection and classification from chest CT scans using machine learning techniques." In 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), pp. 101-104. IEEE, 2021.
10. P. Moradi and M. Jamzad, "Detecting lung cancer lesions in CT images using 3D convolutional neural networks", Proc. 4th Int. Conf. Pattern Recognit. Image Anal. (IPRIA), pp. 114-118, Mar. 2019.
11. M. H. Jony, F. Tuj Johora, P. Khatun and H. K. Rana, "Detection of Lung Cancer from CT Scan Images using GLCM and SVM", 2019 1st International Conference on Advances in Science Engineering and Robotics Technology (ICASERT), pp. 1-6, 2019.
12. Bhandary, A., Prabhu, G.A., Rajinikanth, V., Thanaraj, K.P., Satapathy, S.C., Robbins, D.E., Shasky, C., Zhang, Y.D., Tavares, J.M.R. and Raja, N.S.M., 2020. Deep-learning framework to detect lung abnormality–A study with chest X-ray and lung CT scan images. Pattern Recognition Letters, 129, pp.271-278.

13. Sang, Jun, Mohammad S. Alam, and Hong Xiang. "Automated detection and classification for early stage lung cancer on CT images using deep learning." In Pattern recognition and tracking XXX, vol. 10995, pp. 200-207. SPIE, 2019.
14. Causey, Jason L., Yuanfang Guan, Wei Dong, Karl Walker, Jake A. Qualls, Fred Prior, and Xiuzhen Huang. "Lung cancer screening with low-dose CT scans using a deep learning approach." arXiv preprint arXiv:1906.00240 (2019).
15. Riquelme, Diego, and Moulay A. Akhloufi. "Deep learning for lung cancer nodules detection and classification in CT scans." Ai 1, no. 1 (2020): 28-67.
16. ChestCTScanimagesDatasetkaggle,https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images
17. Mujahid, M., Rustam, F., Álvarez, R., Luis Vidal Mazón, J., Díez, I.D.L.T. and Ashraf, I., 2022. Pneumonia classification from X-ray images with inception-V3 and convolutional neural network. Diagnostics, 12(5), p.1280.
18. Özdemir, A., Polat, K. and Alhudhaif, A., 2021. Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods. Expert Systems with Applications, 178, p.114986.
19. Kido, S., Hirano, Y. and Hashimoto, N., 2018, January. Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN). In 2018 International Workshop on Advanced Image Technology (IWAIT) (pp. 1-4). IEEE.
20. Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J. N. V. R., Jawarneh, M., \& Asenso, E. (2022). Lung cancer classification and prediction using machine learning and image processing. BioMed Research International, 2022.
21. Humayun, Mamoona, R. Sujatha, Saleh Naif Almuayqil, and N. Z. Jhanjhi. "A transfer learning approach with a convolutional neural network for the classification of lung carcinoma." In Healthcare, vol. 10, no. 6, p. 1058. MDPI, 2022.
22. Baranwal, N., Doravari, P. and Kachhoria, R., 2022. Classification of histopathology images of lung cancer using convolutional neural network (CNN). In Disruptive Developments in Biomedical Applications (pp. 75-89). CRC Press.