IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Lung Cancer Detection using Hyper Parameter Tuned Convolution Neural Network

Shloka Bhattacharyya

Great Valley High School
225 North Phoenixville Pike, Malvern, PA 19355

Abstract - ***Lung cancer is a life-threatening disease; early identification of lung cancer increases the chances of successful treatment for patients. Imaging techniques generally followed for lung cancer detection are X-ray, Computed tomography (CT), and histopathological images. However, CT scan images are more reliable in detecting lung cancer; this paper focused on lung cancer detection as a ternary classification problem using CT scan images. This paper proposes classifying lung images as 'normal,' 'benign,' and 'malignant,' which helps doctors treat patients effectively. This ternary classification is proposed through deep learning models. Deep Neural Network (DNN), Long Short Term Memory (LSTM), and Convolutional Neural Network (CNN) were used. Experimental results showed more promising results on the CNN algorithm than other models. Thus, the CNN algorithm is enhanced using hyper parameter tuning, and HT-CNN is proposed. To improve the novelty of work and get high detection accuracy for the ternary classification of lung cancer, hyper parameter tuning with Grid search cross-validation technique is proposed. The GSCV identifies the best-fit parameter for deep learning models and enhances the algorithm's performance. Experimental results showed that CT images were normal, benign, and malignant for ternary classes of the lung with deep learning models DNN, LSTM, CNN, and HT-CNN. The results are compared, showing that the hyper-tuned CNN model has achieved the highest accuracy, 99.4%.***

Key Words: Artificial Intelligence, Machine Learning, Deep learning, Decision Tree, Convolutional Neural Network (CNN), Grid Search Cross Validation

1.INTRODUCTION

Nowadays, shifts in dietary patterns and lifestyle are contributing to health challenges. According to the World Health Organization (WHO), cancer stands out as a foremost life-threatening ailment globally. In 2020, the world struggled with the Covid-19 pandemic, causing a surge in pneumonia cases. Precisely diagnosing and monitoring respiratory illnesses have become global priorities like never before. Lung cancer is particularly prevalent among cancer cases worldwide. Timely identification and diagnosis of cancer play a crucial role in enhancing the chances of prolonged survival for patients, while a lack of early detection may lead to fatal outcomes. Lung carcinoma, resulting from excessive growth in lung tissue, is the primary factor leading to lung disease. The early identification of the disease can prevent its spread to nearby tissues. According to cancer statistics, lung cancer mortality was the leading cause of death, accounting for over 2.1 million deaths worldwide in 2018. There are two types of carcinoma: small and non-small cell, with non-small cell carcinoma comprising 88% of lung cancer cases. Smoking is the predominant cause of lung cancer, though other factors, such as asthma, can contribute. The availability of experts to check lung images is a challenge as numerous images are to be diagnosed daily, which can even cause medical errors [1]. The use of computer-aided detection is essential for accurate and automatic disease detection. This reduces the waiting time needed for expert image reviews. While traditional methods like X-rays and biopsy tests are available, they often entail a substantial amount of time for diagnosis.

Artificial intelligence (AI) techniques have emerged as a powerful tool for disease diagnosis in the healthcare industry, emphasizing the leverage of machine learning models. Machine learning (ML) and deep learning (DL) are predominantly used in disease diagnosis. However, ML models needed feature extraction techniques to implement more effective learning. Thus, deep learning was found to be more effective for image classification due to the ability to automatically extract important features from images. To leverage deep learning features, this work implemented deep learning models for lung image classification.

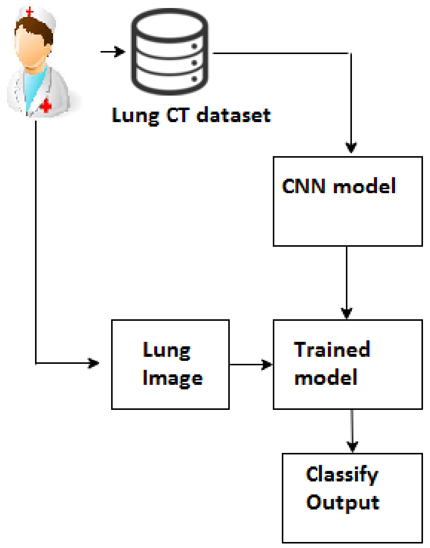

Figure: 1 Overview of Lung Disease Classification

In this paper, Lung cancer detection is proposed as ternary classification, in which lung CT scan images are classified as normal, benign and malignant. This paper studies the implementation of various deep learning models such as Deep Neural Network, Convolutional Neural Network, Recurrent Neural Network (Long short term memory-LSTM) model. The experiments conducted showed that CNN model has achieved better accuracy than other deep learning models. To increase the accuracy of detection, a novel contribution done on CNN, in which hyper parameter tuning of hyper parameters is implemented to increase its performance. Grid search cross validation technique is used to evaluate for five times on the train images to identify the best fit parameter value for the CNN model. Thus, the final CNN model is composed of best fit parameters. Figure 1 shows the overview of the proposed lung disease classification.

- Lung disease prediction is proposed on various deep learning models DNN, LSTM, CNN and performances of these models are compared.
- Identified the best DL model and proposed novel techniques to increase its accuracy of classification.
- The model is improved with a hyper parameter model to give the best performance of lung image classification.
- The proposed work handles the lung CT scan images as ternary classification Normal, benign and malignant.

The organization of the paper includes, literature review on lung disease detection is discussed. In chapter 3, proposed deep learning models are discussed in detail. In chapter 4, the experimental results are discussed. Finally conclusion and further enhancement of this are explored.

2. RELATED WORK

Research on lung cancer prediction has explored various data types, including images, gene expressions, and micro-analysis of protein attributes. Different types of image datasets, such as Computer Tomography (CT), x-ray images, and histopathological images, have been employed in this context. CT scan images, in particular, are widely utilized for lung cancer diagnosis. Existing studies have predominantly employed imaging techniques for lung cancer classification, relying on both machine learning and deep learning approaches. This literature survey on lung tumor detection using artificial intelligence techniques such as machine learning (ML) and deep learning (DL) methods. Traditionally, X-ray images and biopsies were employed for tumor detection, but the advent of AI has facilitated more accurate and rapid detections through ML and DL models. Recent researchers have predominantly focused on the use of CT scan images for lung tumor detection, and some of these studies are discussed in this chapter.

Histopathological images are examined for identifying the specific types: lung adenosquamous carcinoma (ASC), lung squamous cell carcinoma (LUSC), and small cell lung carcinoma (SCLC) in work [2]. Prior to classification, the research employs feature selection and extraction techniques. Histopathological images undergo preprocessing using a Gaussian filtering approach and adaptive histogram equalization technique. Multidimensional features are then extracted from the preprocessed images, employing methods such as wavelet transform and Markov random field. These extracted features undergo further refinement through feature selection techniques, specifically utilizing the relevant features (RELIEF) and LASSO models. The selected features are subsequently trained and evaluated using various classification models such as Support Vector Machine (SVM), Backpropagation (BP), K-nearest neighbor (KNN), Decision Tree (DT), Naïve Bayes (NB), and Convolutional Neural Network (CNN). Experimental findings indicate that the highest accuracy 83.91%, is achieved by employing RELIEF with the SVM model.

Lung cancer detection used machine learning models for classification with image techniques such as segmentation. The work [3] proposed ML models for classifying the disease based on segmentation of lung nodules and features are selected from segmented regions. There are a total 15 features extracted from the lung nodule and surrounding area including Haralick and Gabor features. Classification models applied are SVM and CNN models. The performance SVM is computed through Area Under the Curve (AUC) score, which is 0.967 for SVM model and CNN model has 0.91. However, the machine learning algorithm has performed well on classification of tumor diagnosis, they need feature selection techniques to the best performance.

Context sensitive SVM (csSVM) was proposed in [4], which used the features derived from Region of Interest (ROI). Based on the forward feature selection, a near optimal feature set is selected and given to the classifier. This work aimed to attempt early detection by extracting the features from ROI. These features included texture and shape to detect lung disease. The textual features encompass histogram, gradient, and run-length matrices, as well as gray scale level and shape features such as top-hat transforms. Experimental results demonstrated that the csSVM given an accuracy of approximately 89.88%.

Some of the existing works used chest X-ray (CXR) images for lung disorder detection, the work [5] addressed the enhancement of X-ray images by Fractional order convolution. This was applied in eight directions to improve the features through pattern reconstruction. This technique is more effective in identifying the symptomatic features from X-ray lung images. Multilayer machine learning classifier used k-fold cross validation for lung classification from enhanced X-ray images. The multi-layer vision classifier involves lesion extraction, pattern reconstruction, pooling, and pattern recognition. Gray relation analysis (GRA) is employed for classification with a 10-fold cross-validation. Experimental results showed that this classifier accuracy is 88.88%.

Lung nodule detection based on Region of Interest was proposed in [6] based on classification machine learning algorithms. This work used ROI detection is performed based

on the centroid of each lung nodule. The features are derived from ROI based on Structural Co-occurrence Matrix (SCM), gray-level co-occurrence matrix (GLCM) and local binary pattern (LBP) techniques. Experimental results showed that SCM feature extraction along with Support Vector machine classifier has achieved the highest accuracy of 96.7%.
Existing studies extensively employed deep learning for lung cancer classification, the work [7], proposed Convolutional Neural Network (CNN) for lung cancer detection. The experiment used the interstitial lung diseases (ILD) CT scan dataset. The author proposed CNN architecture with 5 convolution layers with LeakyReLU (Rectified Linear Unit), an average pooling layer, and three fully connected layers with a softmax classifier with an input layer given train images of dimension 32x32. The experimental setup achieved an accuracy of 85.47%.
Existing works studied lung cancer detection with different types of image dataset and with various images techniques such as segmentation feature extraction etc. The work [8] addressed extracting morphological features, Elliptic Fourier descriptors feature and applied Fuzzy entropy for analyzing these features for lung disease detection. Radiomic features for deriving quantitative features from ROI are applied in [9] and then performed classification for lung cancer. Solitary features from chest x-ray images are performed in the work [10] for lung nodule detection, the nodules are extracted based on transform and convergence index filter, then applied ensemble classifier AdaBoost.
From this literature survey, it is inferred that the prior works on lung disease detection used different feature extraction techniques for applying machine learning algorithms. Though Deep learning models were also used in the existing work, the accuracy of lung cancer detection is comparatively less. Some of the prior work showed good performance on classification accuracy, however, they involved the external feature extraction techniques. Thus an effective deep learning model is required, which can extract the features automatically and achieve high accuracy of lung cancer detection. Thus, in this paper, we proposed Hyper tuned Convolutional Neural Network (HT-CNN) to address these issues.

3. PROPOSED WORK

This section discusses the proposed methodology, dataset details and implemented algorithms. The focus is on classifying lung cancer in CT scan images through deep learning models DNN, LSTM and CNN models. Experimental results showed promising results with CNN model, thus CNN is hyper tuned to improve the accuracy. The key parameters such as epochs, batch size, and learning rate are optimized to build the best model. The CNN model is constructed using finely tuned parameters to enhance the accuracy of the classification.

DATASET DETAILS

The dataset comprises Lung CT scan images of both training and testing, categorized into ternary classes: normal, benign, and malignant. The training set includes 217 normal images, 391 malignant images, and 91 benign images. Meanwhile, the test data consists of 175 malignant images, 145 normal images, and 29 benign images. A visual representation of samples from the benign class is presented in the following figure.

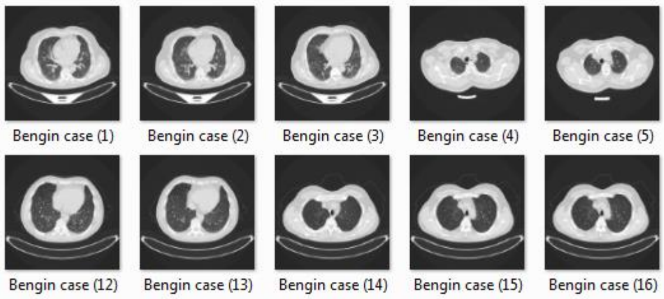

Figure2: Dataset used for Lung disease classification

DATA PRE-PROCESSING

Prior to training, pre-processing is applied to lung CT images. This involves converting the images from BGR to RGB, followed by resizing them to dimensions of 60x60. The images are then transformed into numpy array data and subjected to normalization. The normalized data is utilized for deep learning models DNN, LSTM and CNN during the training and validation processes. Additionally, the labels undergo one-hot encoding to facilitate conversion into categorical representations.

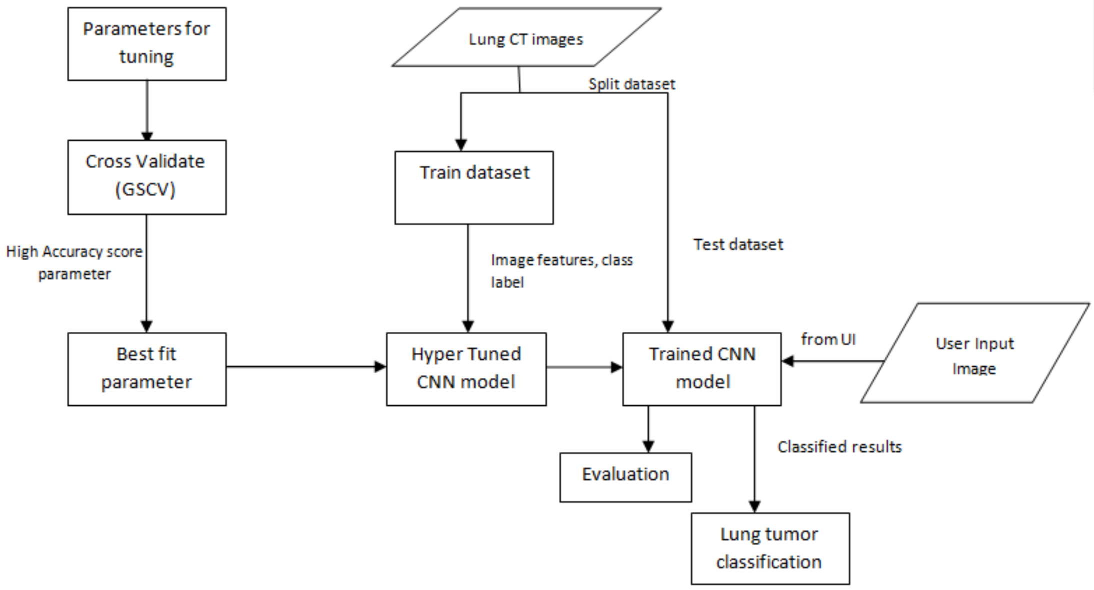

Figure 3: Lung cancer detection using proposed CNN algorithm

Figure 2 shows the architecture of enhanced CNN algorithm for improved accuracy of lung disease detection. The CNN model is built with best tuned parameters using Grid search cross validation. The validation taken five times to evaluate the parameter value based on accuracy score. Classification of lung images is performed with hyper parameter tuned CNN model.

a. Deep Neural Network

Deep neural network (DNN) model implemented for classification of lung disease as normal, benign and malignant. The proposed DNN model has flatten input layer as the first layer given an input dimension of images 60x60 and 3 classes. The second layer is Dense layer with 128 neurons with activation function 'relu'. Followed by dense layer, a dropout is added 50% to avoid over-fitting problem. The third layer is Dense layer with 64 neurons with activation function 'relu' and added a dropout 50%. The third Dense layer is added with 32 neurons with activation function 'relu' and added a dropout 50%. Finally a dense layer with activation 'softmax' is added for ternary classification.

b. Convolutional Neural Network

This paper focuses on the implementation of a Convolutional Neural Network (CNN) algorithm for the purpose of classifying lung diseases into three categories: normal, benign, and malignant. The proposed CNN architecture comprises three Convolutional layers and four Dense layers. The first Convolutional layer features an input dimension of 60x60, 32 input neurons, a filter size of 3x3, and employs the Rectified Linear Unit (ReLU) activation function. Subsequently, a Maxpooling layer with a pool size of (2,2) is incorporated, along with a 20% dropout. The second Convolutional layer introduces 64 neurons, a filter size of 3x3, and utilizes the ReLU activation function. Following this, a Maxpooling layer and a 20% dropout are applied. The third Convolutional layer incorporates 128 neurons, a filter size of (3,3), and the ReLU activation function. Post the third convolutional layer, a Maxpooling layer and a 40% dropout are introduced. Subsequently, a Flatten layer is added to the network. Four Dense layers follow with neurons set at 32, 64, and 128, respectively. The fourth Dense layer serves as the output layer with the 'softmax' activation function. The training parameters include 25 epochs, a batch size of 125, and a validation split of 20%.

c. Long Short Term memory (LSTM) model

Long Short Term Memory (LSTM) is a type of Recurrent Neural Networks model. The LSTM model classifies the lung images into normal, benign and malignant. The LSTM model is given input train_X and train_y, where train_X is the image features extracted from Lung CT images. train_y is the label of three classes.

The training parameters include 25 epochs, batch size of 128 and a validation split of 20%. This model comprises three LSTM layers and a Dense layer. The first LSTM layer is configured with 128 neurons, using the Rectified Linear Unit (Relu) activation function. Followed by this layer, a dropout of 50% added to overcome over-fitting. The second layer is added with 64 neurons, followed by a dropout of 50%. The third LSTM is added with 32 neurons followed by a dropout 50%. Flatten layer is added. Finally a dense layer with activation function 'sofmax' is added. The model is compiled with optimizer 'adam', accuracy metric and loss is 'categorical cross entropy'.

d. Hyper parameter Tuned CNN model

Tuning hyperparameters, namely batch size, learning rate, and epochs, is conducted using the Grid Search Cross Validation (GSCV) technique. Cross-validation is executed within the specified search space for each parameter, calculating the mean and standard deviation of errors. The parameter value with the lowest error is selected. In the following equations, 'b' represents the number of folds, 'm' denotes the mean, and 'se' signifies the standard deviation. Equations (1) and (2) below outline the calculations for mean and standard deviation.

$$m = \frac{\sum_{i=1}^{b} err_i}{b} \quad \text{---(1)}$$

$$se = \sqrt{\frac{var(err)}{b}} \quad \text{---(2)}$$

Algorithm: Enhanced CNN algorithm

Input: Search space of hyper parameters Batchsize, epochs, learning rate

Output: Best fit value of hyper parameters Batchsize, epochs, learning rate

Step 1: Input the hyperparameter search space for learning rate (L), batch size (B), and epochs (E).

Step 2:Compute error= $\sum_{i=1}^{b} err_i \, /b$

Step 3: Compute standard deviation error= √(var(err)/b)

Step 4: Sort Error value to select the best fit parameter value

Step 5: Build tuned CNN model with parameters to get best_estimator

Table 1: Parameters and its values for Cross validation and tuning

Parameter for tuning	Hyper parameter values
Batch Size	4,8,16,32
Epochs	5,10,15,20
Learning Rate	0.1,0.01,0.001

The table presented above illustrates the search space employed for cross-validating hyperparameters. The optimal parameters, determined using the Grid Search Cross-Validation (GSCV) technique, were then utilized in constructing the CNN model. Hyper tuned CNN model improved the accuracy of lung disease classification. The architecture of CNN Hyper tuned model is given in figure 4.

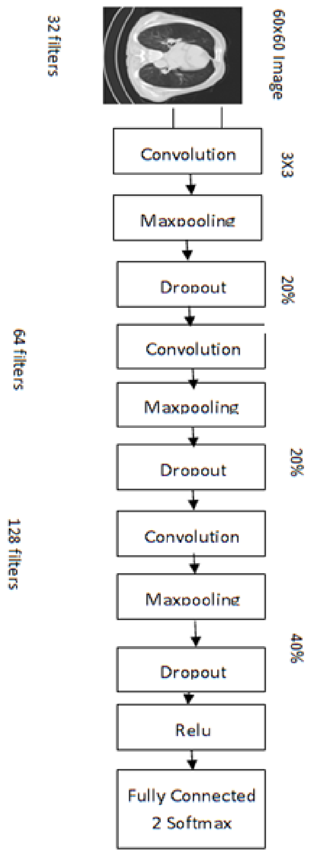

Figure 4: CNN Architecture for lung disease detection

4. RESULTS AND DISCUSSIONS

There are three DL models implemented are DNN, LSTM and CNN and their performances are compared. CNN model has given promising results in classification of lung cancer and further improved performance. Thus the proposed lung cancer detection system involves ternary classification, distinguishing between normal, benign and malignant cases. This is implemented using hyperparameter-tuned Convolutional Neural Network (CNN) architecture. Hyperparameters optimized are batch size, epochs, and learning rate. Mean and standard deviation errors were calculated through cross-validation, and parameter values associated with low error were chosen to establish the best-fit parameters to build the CNN model.

The accuracy of lung cancer classification for the LSTM model is around 57.4%. The model is not converged and thus the accuracy has variations for every epoch. The accuracy plot for the LSTM model is provided in Figure 5.

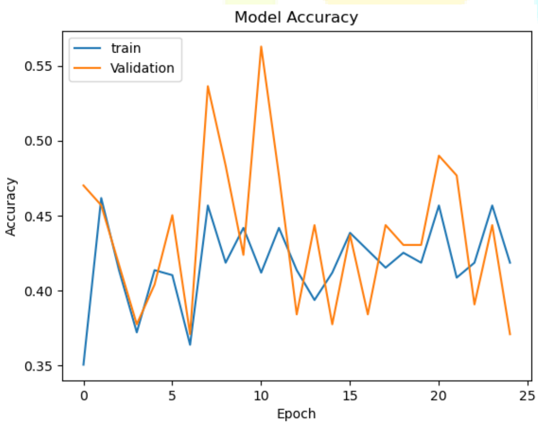

Figure 5: Accuracy of LSTM model

The accuracy of lung cancer classification using the DNN algorithm is 80.86%. Though the model has achieved higher accuracy than the LSTM model, the DNN model is not much converged and thus the variation of accuracy values seen in every epoch. The accuracy plot for the DNN model is given in Figure 6.

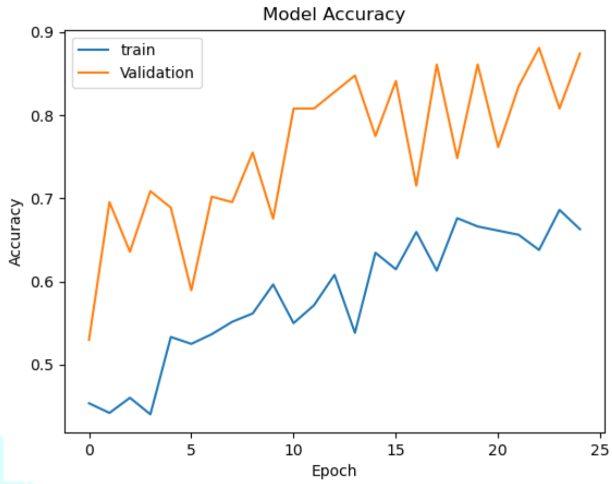

Figure 6: Accuracy of DNN model

The accuracy of CNN model for lung cancer as ternary classification has achieved the highest around 92.32%. The model is more converged with increase in epochs as shown in Figure 7.

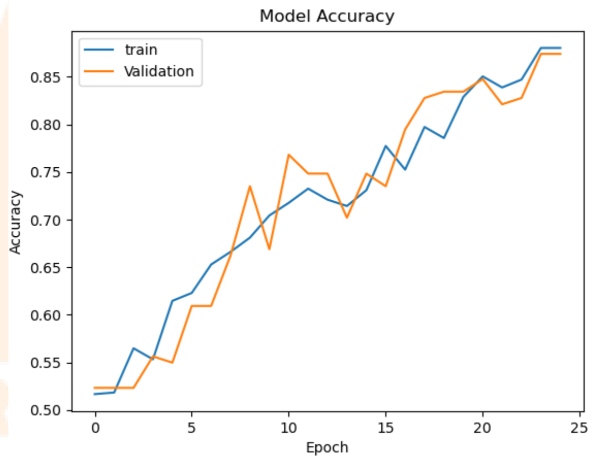

Figure 7: Accuracy of CNN model

The hyper-tuned CNN model achieved the highest accuracy of 99.4%, surpassing the 94.6% accuracy achieved by the standard CNN model. The model underwent training for 15 epochs, with the validation data given the highest accuracy in the experimental results. Figure 8 illustrates the accuracy of both training and validation data.

In contrast to the standard CNN model, the hyper-tuned CNN model demonstrates a notably lower loss metric. The loss consistently decreases with the increased epochs, as depicted in Figure 9. Cross-entropy loss is computed for both the training and validation datasets, resulting in metrics of 0.14 and 0.12, respectively.

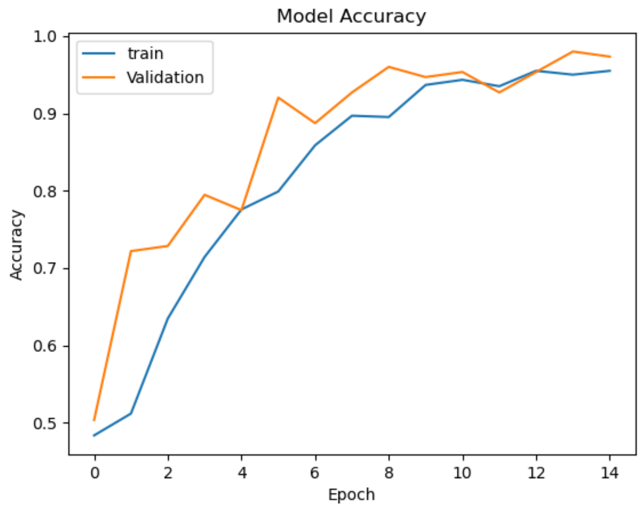

Figure 8: Training and validation accuracy of proposed CNN model

The loss metrics computed by the hyper tuned CNN model is comparatively lower than normal CNN model. The loss is decreased with increasing epochs as shown in figure 7. The cross entropy loss is computed with training and validation dataset. The loss metric is 0.12 and 0.14 for validation and training data respectively.

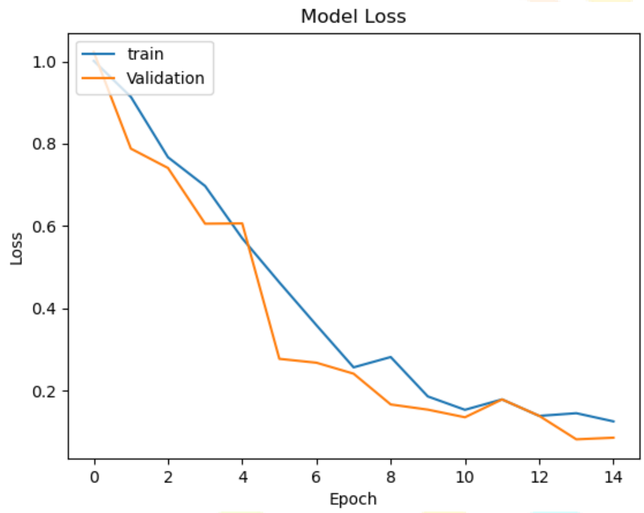

Figure 9: Loss metric for Hypertuned CNN model

The tabulated results of the experimental results of deep learning models are presented in the preceding table. The results indicate the superior performance of the hyperparameter-tuned Convolutional Neural Network model in terms of classification accuracy when compared to alternative models. The high precision in predicting lung cancer facilitates early-stage detection, thereby enabling timely interventions and appropriate medical treatments for patients.

Table 2: Comparison of Performance in the existing and proposed model

Models	Accuracy	MAE Loss
DNN	80.86%	0.19
LSTM	57.4%	0.52
CNN	92.32%	0.13
CNN - tuned	96.34%	0.11

5. CONCLUSIONS

Lung cancer, being a life-threatening disease, necessitates early detection. To facilitate automated diagnosis and precise identification, this paper proposed the utilization of deep learning models for classifying lung cancer as normal, bening and malignant. This work is the implementation of four distinct deep learning models: DNN, LSTM, CNN, and HT-CNN. Hyperparameter tuning is performed using Grid Search Cross Validation, focusing on parameters such as the number of epochs, learning rate, and batch size. The empirical findings showed that the hyperparameter-tuned CNN model attains the highest accuracy in classifying lung diseases.

Future investigations could extend this work by incorporating image augmentation techniques and integrating pre-trained models like DenseNet and VGG16 architectures. The exploration of hyperparameter tuning for these pre-trained models could further enhance the depth of study in this domain.

REFERENCES

[1] Y. Chen, Y. Wang, F. Hu, L. Feng, T. Zhou and C. Zheng, "LDNNET: Towards Robust Classification of Lung Nodule and Cancer Using Lung Dense Neural Network," in IEEE Access, vol. 9, pp. 50301-50320, 2021, doi: 10.1109/ACCESS.2021.3068896.

[2] M. Li et al., "Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathological Images," in IEEE Access, vol. 9, pp. 53687-53707, 2021, doi: 10.1109/ACCESS.2021.3071057.

[3] Kadir, Timor & Gleeson, Fergus. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. Translational Lung Cancer Research. 7. 304-312. 10.21037/tlcr.2018.05.15.

[4] Lim, Jonghyuck & Kim, Namkug & Seo, Joon Beom & Lee, Young & Lee, Youngjoo & Kang, Suk-Ho. (2011). Regional Context-Sensitive Support Vector Machine Classifier to Improve Automated Identification of Regional Patterns of Diffuse Interstitial Lung Disease. Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology. 24. 1133-40. 10.1007/s10278-011-9367-0.

[5] C. -H. Lin, J. -X. Wu, C. -M. Li, P. -Y. Chen, N. -S. Pai and Y. -C. Kuo, "Enhancement of Chest X-Ray Images to Improve Screening Accuracy Rate Using Iterated Function System and Multilayer Fractional-Order Machine Learning Classifier," in IEEE Photonics Journal, vol. 12, no. 4, pp. 1-18, Aug. 2020, Art no. 4100218, doi: 10.1109/JPHOT.2020.3013193.

[6] M. B. Rodrigues et al., "Health of Things Algorithms for Malignancy Level Classification of Lung Nodules," in IEEE Access, vol. 6, pp. 18592-18601, 2018, doi: 10.1109/ACCESS.2018.2817614.

[7] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe and S. Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network," in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1207-1216, May 2016, doi: 10.1109/TMI.2016.2535865.

[8] L. Hussain, W. Aziz, A. A. Alshdadi, M. S. Ahmed Nadeem, I. R. Khan and Q. -U. -A. Chaudhry, "Analyzing the Dynamics of Lung Cancer Imaging Data Using Refined Fuzzy Entropy Methods by Extracting Different Features," in IEEE Access, vol. 7, pp. 64704-64721, 2019, doi: 10.1109/ACCESS.2019.2917303.

[9] Alahmari, Saeed & Cherezov, Dmitry & Goldgof, Dmitry & Hall, lawrence & Gillies, Robert & Schabath, Matthew. (2018). Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2884126.

[10] X. Li, L. Shen and S. Luo, "A Solitary Feature-Based Lung Nodule Detection Approach for Chest X-Ray Radiographs," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 516-524, March 2018, doi: 10.1109/JBHI.2017.2661805.

[11] M. Šarić, M. Russo, M. Stella and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 2019, pp. 1-4, doi: 10.23919/SpliTech.2019.8783041.

[12] Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC), Budapest, Hungary, 2017, pp. 88-91, doi: 10.1109/ISCSIC.2017.22.

[13] Z. Tariq, S. K. Shah and Y. Lee, "Lung Disease Classification using Deep Convolutional Neural Network," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 732-735, doi: 10.1109/BIBM47256.2019.8983071.