An Effective Lung Cancer Diagnosis Model Using the CNN Algorithm

Sonia Kukreja[1], Munish Sabharwal[2], D.S.Gill[3]

{Soniamitkukreja@gmail.com[1], mscheckmail@yahoo.com [2], dsgill@outlook.kr[3]}

Galgotias University, Greater Noida, India [1,2], SRUniversity, Warangal, Telangana[3]

Abstract. Lung cancer is a very complicated disease and can be deadly if not detected on time or at the early phase. A cost effective and more accurate methodology is required to diagnose the lung cancer at an early stage. It results in improved diagnosis, saving the money as well as time and in simplifying the operations. This study focused on the 3Dimensional CNN architecture by categorizing the imaging data or histopathological images in three kinds of cancer: squamous cell carcinoma, adenocarcinoma, and Benign. The current research categorize these 3 types of cancer by implementing the CNN architecture and achieving the better accuracy by comparing with the other methologies used for the diagnosis of lung cancer. While implementing CNN method, accuracy achieved in training stage was 96.11 and accuracy achieved in validation was 97.2%. This proposed method has the potential to improve the detection of lung cancer by grouping them according to the symptoms they have. In this research, along with CNN, random forest technique has been used to to reduce the resources, labour required and time also. CNN model achieved the improved detection accuracy for lung cancer and saving the lives by indulging early disease recognition.

Keywords: Random Forest, Image classification, Deep learning,CT scan, CNN

1 Introduction

Numerous reasons are there for the lung cancer and sometimes it becomes mystery to find out the actual reason. Lung cancer is a serious disease and timely identification of this helps in increasing the chances of saving the life of a patient. Early detection of lung cancer increases the chances of survival and making the methodology more effective and helpful in fighting against the lung cancer [1]. Various parameters play the important role in the diagnosis of the cancer and some important parameters considered are the tumor area, size, exposure to dangerous material like radon, secondhand smoke, asbestos and the speed by which this tumor is spreading in other parts of the body. Among all these parameters speed of spreading the tumor and size are two important factors which helps in categorizing the stage of the cancer [2]. Among various factors, Tobacco and cigarette smoking is the major reason in lung cancer, sometimes exposure to hazardous substances can also be the reason of lung cancer development.

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343145

Identification of lung cancer can be done in different steps, each of which needs a unique detection approach. Medical images can be organized into groups based upon the characteristics, some features are common in one another, so classification becomes an extremely important step [3]. Convolutional neural network methodology has been used to process the JPEG-encoded DICOM images of the lungs and it helps in identifying the abnormalities, if present [4]. Image processing methods like feature extraction, histogram equalization, grayscale conversion and thresholding are useful in identifying the abnormalities in the images [5]. To boost up the speed and accuracy of the lung cancer detection machine learning techniques are used [6].

The primary aim of this study is to examine the classification of 3 unique histopathological images which belongs to lung cancer. These images can be squamous cell carcinomas, adenocarcinoma and benign. To classify these images in accurate manner, CNN has been implemented [7]. The precision of CNN methodology is of the highest understanding because of the high impact it has on the output of therapy [8]. The result of this research highlights that the early detection of the lung cancer is most important but also shows that machine learning algorithms provides a great help in achieving both accuracy and speed in the detection method [9]. Various machine learning algorithms are there which can help in the early detection of cancer.

2 Literature Review

A literature survey on different types of lung cancer diagnosis using CNN algorithms reveals a growing interest in leveraging deep learning techniques for more accurate and classification and detection of lung cancer. A description of a few significant studies in this field is provided below: A 3D dual path net-based deep neural network model for the automated identification and categorization of lung nodules [10]. The model achieves high accuracy in identifying lung nodules from CT scans by using a CNN architecture. In order to identify lung nodules in chest CT scans, the authors [11] propose employing a deep learning algorithm. a combination of an acceptable response and false positive rate, the algorithm's use of a 3D CNN to identify spatial relationships within the data produces positive outcomes. The main objective of this research was to create a new accurate technique for the diagnosis of lung cancer lesions although victoriously impeding false positives. The use of deep learning methods, specifically CNNs, for lung cancer detection is investigated in this paper [12] achieving high accuracy requires feature extraction, which the authors address along with the use of pre-trained models. In this study, deep learning (DL) technology was used as part of an extensive strategy to accomplish the primary goal of early detection of pneumonia and lung cancer. new methodology which contains lung cancer, including SVM, KNN, and CNN. The ultimate goal of this research is to save as many lives as possible by using these approaches. Another method of deep learning approach combined hand-crafted and learned attributes within the MAN framework to improve accuracy specifically for assessing lung cancer. When validated using typical lung cancer CT scans from the LIDC-IDRI dataset, this DL framework had an impressive accuracy rate of 97.27%, as shown in [12]. The literature survey underscores the transformative potential of CNN algorithms in revolutionizing lung cancer diagnosis the precision and effectiveness of lung abnormality diagnosis.
In May of 2021, a number of scholars collaborated on the publication of a paper that used the correlation approach [13]. This research investigates the achievement rates of classification algorithms that are employed in the early diagnosis of lung cancer, including SVM, KNN, and

CNN. The ultimate goal of this research is to save as many lives as possible by using these approaches. The findings of this study point to a potential method for predicting and determining the stages of lung tumors. As can be seen in figure 1 which depicts the flow diagram of the expected model, the process of the proposed model starts with the preparation of the data and then moves on to the selection of features [14], the categorization of the dataset, and the assessment of the dataset. Weka algorithms were heavily used throughout the process of describing the lung cancer datasets that were included in this research. The correlation attribute approach was used throughout the whole of this book for the purpose of picking features [15]. The accuracy of the SVM is 95.56, whereas the precision of the KNN is 88.40 and the precision of the CNN is 92.11 [16]. Accuracy suffers whenever the datasets in question are run via the KNN classifier that the recommended model employs [17-18].

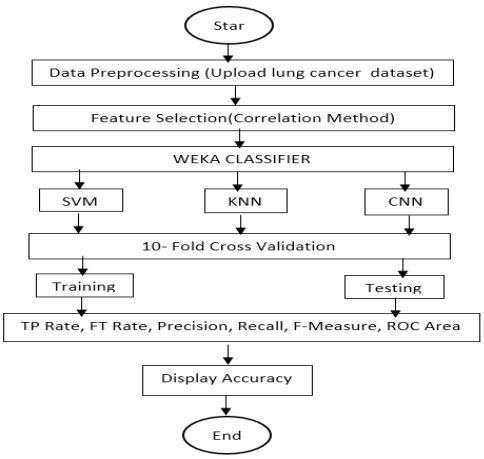

Fig. 1. Lung cancer diagnosis using WEKA Technique and correlation method.

3 Methodology

The following sections offer an extensive overview of the primary steps that comprise our proposed system. These phases are as follows: data collection, data formatting, model training, model testing, and prediction. Our system followed a strong and well-structured strategy, which was ensured by meticulous data collection, formatting, and use to train an accurate prediction model. This method was followed by meticulously carrying out these procedures.

3.1 Data Collection

The data has been collected from LC25000 which is lung and colon histopathological images. Tissue samples have been taken from the patient's Lung and colon and images have been generated from these samples. An extensive set of histopathology images has been carefully selected for this research. Aim is to work on three particular categories of lung tissue: benign tissue, squamous carcinoma cells and adenocarcinoma cells. Training and validation with the help of CNN model, has been performed on this dataset, which results in increasing the accuracy and speed. The availability of such a substantial dataset allowed us to draw reliable conclusions and make informed decisions based on the results obtained from experiments.

3.2 Preprocessing

The dataset is a combination of histopathological photographs in the JPEG file format, and each images contains RGB colors. All the images have been resized to an aspect ratio of one and a pixel size of (180, 180) to ensure the consistency and perfect functioning of CNN. The main objective of this resizing is to maintain the uniformity in the dataset and to simplify the operation of CNN. These pictures have been ranged between 0 and 1 to expedite the process.

It is important to remember that overfitting presents when a model becomes overly worked on the training set, which subsides the caliber to generalize effectively to brand-new, untried data. By implementing data augmentation methods such as zooming and flipping, aimed to create a more robust and diversified training set. These techniques enabled the neural network to learn a wider range of patterns and features, thus reducing the risk of overfitting.

3.3 Deep Learning

Convolutional neural networks are a subclass of feed-forward neural networks. CNNs are able to generate a kind of translational invariance which have the same characteristics to overlapping parts of the layer below them. The capacity of CNNs to identify objects inside their receptive field, regardless of differences in size, position, orientation, and other visual features, is the prime objective offered by these types of cameras. Additionally, in contrast to fully connected neural networks, the training process for CNNs requires less computer power due to the constrained connection of CNNs.

Convolutional layer: This layer receives input pictures that are appropriate for network training and transforms those images into feature maps by applying filters or convolutional kernels to the images. The filters work their way through the input dimensions, pulling out characteristics that are important to the problem.

Pooling layer: This layer takes the feature maps from the convolutional layers and makes them smaller. This helps to reduce the number of parameters. As filters travel through the convolutional layer output, it conducts down sampling by computing the maximum value, which is also referred to as the weighted average.

Fully connected layer: This layer is responsible for assigning precise labels to the pictures that were produced by the layers that came before it. It does this by using the SoftMax layer in order to calculate the probability of values ranging from 0 to 1. Batch normalization is used in order to increase the pace of training and decrease the likelihood of overfitting.

Here, a deep CNN determines whether a particular nodule is benign or cancerous. Finally, a SoftMax layer is used at the end of the architecture. The persuasiveness of the deep CNN architecture is calculated through experimental studies, and the outputs are shown in Figure. 2 and 3.

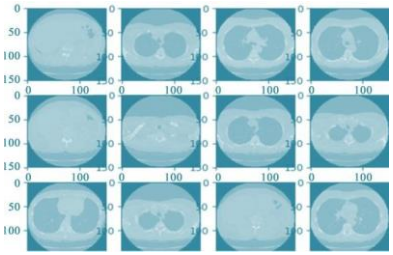

Fig. 2. Benign Image

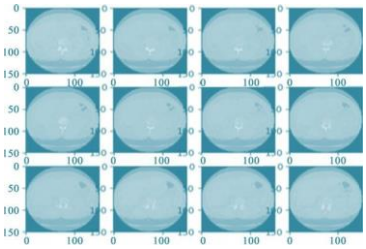

Fig. 3. Malignant Images

3.4 Model Training, Evaluation, and Forecasting

Image classification can be easily handled by CNNs. To process the input image and extract the features, ConvNet-layered convolutional layers are commonly used. Sigmoid function has been used to calculate the probabilities and a dropout value of 0.1 was used to remove overfitting. The Adam optimizer was used to adjust the learning rates of the model's parameters. The loss function used for optimization was categorical cross-entropy (CE), which computes the difference between the predicted class probabilities and the true labels for a given input.

$$C.E = -\log\left(\frac{e^{Sp}}{\Sigma_j^c e^{Sj}}\right) \quad (1)$$

The categorical cross-entropy loss function is described in the equation (1), where C defines the number of output classes, Sp is the CNN score of the positive class, and Sj represents the scores inferred by the network for each class C. This loss function helps in guiding the learning process of the CNN by penalizing incorrect predictions and encouraging convergence towards accurate class probabilities.

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} \quad (2)$$

$$Precision = \frac{\text{TP}}{(\text{TP} + FP)} \quad (3)$$

$$Recall = \frac{\text{TP}}{(\text{TP}+FN)} \quad (4)$$

$$F1 - Score = \frac{2*(Recall*precision)}{(Recall+Precision)} \quad (5)$$

To assess the effectiveness of the newly created CNN model, a confusion matrix plot was generated. Equation (2) calculates accuracy as the proportion of correctly classified instances among all instances. Precision which is defined in (equation (3)) evaluates the model's ability to predict positive instances correctly, yet recall which is defined in (equation (4)) produces the share of true positive instances that were correctly anticipated. The f1-score defined in (equation (5)). True positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) values are used in these evaluation metrics. The predicted and true labels for both the training and validation images are compared to obtain these values. The accuracy, f1-score, recall, precision and provide a snapshot of the CNN technique's performance as well as capacity to appropriately categorize instances from different classes.

4 Result and Discussion

Training was carried out in this study to train the CNN architecture using a meticulously selected dataset of 64 samples per batch. The training process included 211 steps per epoch, for a total of 20 epochs. The training phase achieved an impressive 96.11% accuracy, while the validation phase achieved an even better accuracy of 97.20%. Figure 4 shows the accuracy of the model plotted against the number of epochs for the images used for training to visualize its performance throughout the training process. Figure 5 additionally illustrates the model's loss across epochs during the validation phase. It's worth noting that both figures were created with the same set of images.The achieved accuracy, as well as the visualization of the training and validation progress, contributes to a better understanding of the model's performance and potential clinical applications.

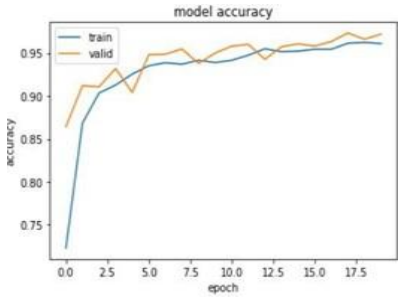

Fig. 4. The plot shows the model accuracy as a function a function of time for both training and validation images.

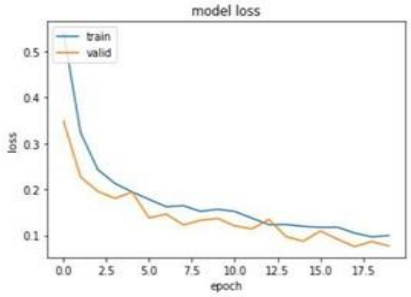

Fig. 5. The plot shows the model Loss as for training and validation images.

5 Conclusion

In the present investigation, a carefully chosen dataset consisting of 64 samples per batch, total 211 steps as per 20 epochs, was used for the training process. With an accuracy of 96.11%, the training process' results accuracy increases at 97.20% during the validation phase. These high accuracies demonstrate the Convolutional Neural Network (CNN) architecture's strong learning capabilities on the provided dataset. In order to give an overview of the model's performance during training, in Figure 4 shows the training model for image's epoch count and Figure 5 show the model's loss across epochs during the validation process, it insights into the model to generalize the unseen data. Figure 4 and 5, uses the same set of images, ensuring consistency in the evaluation process. The results show that the trained CNN architecture works to improve in the area of medical imaging and healthcare by categorization of lung cancer. Since the same collection of photos was used to create both figures, the evaluation was consistent.

References

[1] Moradi, P., & Jamzad, M. (2019, March). Detecting lung cancer lesions in CT images using 3D convolutional neural networks. In *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)* (pp. 114-118). IEEE.

[2] Jothilakshmi, R., and SV, R. G. “Early Lung Cancer Detection Using Machine 6 Learning And Image Processing” In Journal of Engineering Sciences, in 2020.

[3] Bhandary, A., Prabhu, G. A., Rajinikanth, V., Thanaraj, K. P., Satapathy, S. C., Robbins, D. E., ... & 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*, 271-278.

[4] Khan, A. (2021). Identification of lung cancer using convolutional neural networks based classification. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(10), 192-203..

[5] Pandian, R., Vedanarayanan, V., Kumar, D. R., & Rajakumar, R. (2022). Detection and classification of lung cancer using CNN and Google net. *Measurement: Sensors*, *24*, 100588..

[6] Kumar, R. R., Polepaka, S., Likithasree, D., & Keerthika, S. (2023, January). An Investigation on CNN-based Lung Cancer Prediction Method. In *2023 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-5). IEEE.

[7] Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung cancer prediction and classification based on correlation selection method using machine learning techniques. *Qubahan Academic Journal*, *1*(2), 141-149.

[8] Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B. P., Mohanavel, V., ... & Asfaw, A. K. (2022). Lung cancer prediction from text datasets using machine learning. *BioMed Research International*, *2022*.

[9] Sasikala, S., Bharathi, M., & Sowmiya, B. R. (2018). Lung cancer detection and classification using deep CNN. *international journal of innovative technology and exploring engineering*, *8*(25), 259-262.

[10] Hatuwal, B. K., & Thapa, H. C. (2020). Lung cancer detection using convolutional neural network on histopathological images. *Int. J. Comput. Trends Technol*, *68*(10), 21-24.

[11] D. Jain, P. Singh, A. K. Pandey, M. Singh, H. Singh and A. Singh, "Lung Cancer Detection Using Convolutional Neural Network," *2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, Ghaziabad, India, 2022, pp. 1-4, doi: 10.1109/ICICT55121.2022.10064513.

[12] Shah, A. A., Malik, H. A. M., Muhammad, A., Alourani, A., & Butt, Z. A. (2023). Deep learning ensemble 2D CNN approach towards the detection of lung cancer. *Scientific Reports*, *13*(1), 2987.

[13] Onozato, Y., Iwata, T., Uematsu, Y., Shimizu, D., Yamamoto, T., Matsui, Y., ... & Yoshino, I. (2023). Predicting pathological highly invasive lung cancer from preoperative [18F] FDG PET/CT with multiple machine learning models. *European Journal of Nuclear Medicine and Molecular Imaging*, *50*(3), 715-726.

[14] Huang, T., Le, D., Yuan, L., Xu, S., & Peng, X. (2023). Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit. *Plos one*, *18*(1), e0280606.

[15] Chandran, U., Reps, J., Yang, R., Vachani, A., Maldonado, F., & Kalsekar, I. (2023). Machine learning and real-world data to predict lung cancer risk in routine care. Cancer Epidemiology, Biomarkers & Prevention, 32(3), 337-343.

[16] Adams, S. J., Mikhael, P., Wohlwend, J., Barzilay, R., Sequist, L. V., & Fintelmann, F. J. (2023). Artificial Intelligence and Machine Learning in Lung Cancer Screening. Thoracic Surgery Clinics.

[17] Khouadja, O., & Naceur, M. S. (2023, April). Lung Cancer Detection with Machine Learning and Deep Learning: A Narrative Review. In 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET) (pp. 1-8). IEEE.

[18] Kong, C., Lai, L., Jin, X., Chen, W., Ding, J., Zheng, L., ... & Ji, J. (2023). Machine Learning Classifier for Preoperative Prediction of Early Recurrence After Bronchial Arterial Chemoembolization Treatment in Lung Cancer Patients. Academic Radiology