EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

An Effective Lung Cancer Diagnosis Model Using the CNN Algorithm

S. Kukreja [1,*], M. Sabharwal[2]

[1,2] Galgotias University, Greater Noida, India

Abstract

The disease known as lung cancer is a serious condition that may be deadly if it is not diagnosed at an early stage. The diagnosis of lung cancer has to be improved, and there is a need for a cost-effective and user-friendly system that leverages state-of-the-art data science technology. This would help simplify operations, save time and money, and improve diagnosis. This research suggests the use of a convolutional neural network (CNN) architecture for the purpose of categorizing three unique histopathological pictures, namely benign, adenocarcinoma, and squamous cell carcinoma. The purpose of this study is to apply the CNN model to properly classify these three kinds of cancers and to compare the accuracy of the CNN model to the accuracy of other techniques that have been employed in investigations that are comparable to this one. The CNN model was not used in any of the preceding research for the purpose of categorizing these particular histopathological pictures; hence, the relevance of this work cannot be overstated. It is possible to get more positive treatment results by correctly classifying malignant tumors as early as possible. In training, the CNN model obtained an accuracy of 96.11%, and in validation, it earned an accuracy of 97.2%. The suggested method has the potential to improve lung cancer diagnosis in patients by classifying them into subgroups according to the symptoms they exhibit. This approach to machine learning, which makes use of the random forest technique, has the potential to reduce the amount of time, resources, and labor required. Utilizing the CNN model to categorize histopathological pictures may, ultimately, improve the diagnostic accuracy of lung cancer and save lives by allowing early disease identification.

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

Received on 10 03 2024, accepted on 20 06 2024, published on 30 07 2024

doi: 10.4108/eetpht.10.6805

1. Introduction

As a result of our limited knowledge, the root causes of lung cancer, which are numerous and multifaceted, remain a mystery. Lung cancer is a complicated illness. The identification of lung cancer at an earlier stage considerably increases the likelihood that therapy will be successful, making it the most effective tactic in the fight against the disease [1].

* Corresponding author. Email: Soniamitkukreja@gmail.com

The size of the tumor and the pace at which it spreads to other parts of the body are two factors that are used to categorize the stage of cancer. Exposure to dangerous compounds including radon, asbestos, and secondhand smoke may all contribute to the development of lung cancer [2]. Cigarette smoking is the leading cause of lung cancer, but occupational exposure to hazardous substances can also contribute to its development.

The process of detecting lung cancer requires identifying a number of different subtypes, each of which calls for a unique diagnostic approach. When it comes to organizing medical pictures into groups based on the features, they have in common with one another, classification is an extremely important step [3]. In the context of lung cancer, the goal of the proposed system is to perform an analysis of the properties of CT scan pictures in order to locate malignant lesions when they are present. JPEG-encoded DICOM pictures of the lungs are processed by this system using a method known as a convolutional neural network (CNN), which may identify any abnormalities if they are present [4]. After that, the computer computes certain anomalous properties and sends them to a trained system, which uses them to decide if the observed aberration is cancerous or benign. In order to find anomalies within the photos, image processing methods such as thresholding, feature extraction, histogram equalization, and grayscale conversion are used [5]. After then, machine learning algorithms are trained on a dataset so that they can reliably categorize tumors as either benign or malignant. It is impossible to stress the importance of early identification in lung cancer, since it plays a
significant role in determining how well following therapy will work. In this specific instance, techniques of machine learning are used in order to boost the speed of lung cancer diagnosis as well as the accuracy of the diagnosis [6].

The primary objective of this study work is to investigate the categorization of three unique histopathological pictures that are connected with lung cancer. These images include benign, adenocarcinoma, and squamous cell carcinomas. In order to achieve this goal, an architecture known as a convolutional neural network (CNN) is used [7]. The CNN model's accuracy will be compared with the accuracy of other methods that have been used in research projects that are quite comparable to this one. The importance of this study lies in the novel use of the CNN model to classify these particular histopathology pictures, which is a field that has not been properly investigated in any of the prior studies. When identifying malignant tumors at an early stage, the accuracy of the CNN model is of the highest significance because of the enormous influence it has on the results of therapy [8]. The findings of this research not only highlight how important early diagnosis is to the successful care of lung cancer, but they also show how machine learning technologies may improve both the speed and accuracy of the diagnostic process. In conclusion, the results highlight the significance of early diagnosis, provide a unique application of the CNN model, and contribute to the development of more effective ways for recognizing and treating lung cancer [9].

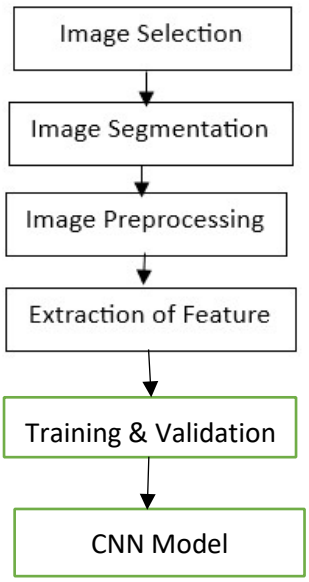

Figure 1. Flow Chart of ML approach for lung cancer diagnosis.

2. Literature Review

This study primarily focuses on utilizing a convolutional neural network (CNN) architecture to classify three types of lung cancer: benign, adenocarcinoma, and squamous cell carcinomas. What makes this work significant is its exploration of the accuracy of the CNN model specifically in identifying and categorizing these three distinct histopathological images, which has received limited attention in previous research studies. Although deep learning and ensemble learning algorithms have been widely used in classification tasks, the current computer-aided diagnostic (CAD) applications for lung cancer classification of lung nodules closely align with the objective of this study. Thus, the research team conducted a comprehensive analysis of recently developed and state-of-the-art lung nodule categorization systems [10].

In 2019, a comprehensive study was conducted to explore the effectiveness of convolutional neural networks (CNNs) in the

diagnosis of lung cancer. The researchers constructed an innovative architecture which incorporates three-dimensional CNNs with a meta-classifier, enabling the network to identify nodules of different dimensions. The study's significant findings included the construction of multiple layers 3D CNNs and a novel approach integrating classifiers to mitigate faults. The study's major goal was to develop a new and accurate method for detecting lung cancer lesions while successfully preventing false positives [11].

The proposed fusion method achieved an impressive accuracy rate of 91.23%, accompanied by only 3.99 false positives per scan. The researchers achieved this by leveraging the knowledge acquired from the classifiers, leading to improved accuracy and a reduction in false positive rates. Overall, the study demonstrated the significant potential of CNNs in enhancing the precision and effectiveness of lung cancer diagnosis, representing a valuable contribution to the field. Another group of researchers presented a study that employed deep learning techniques in the diagnosis of lung abnormalities. The study focused on analyzing chest X-ray and lung CT scan images to identify various anomalies, as illustrated in Fig. 2. The objective of this method was to assist in the early diagnosis and diagnosis of lung cancer.

Deep learning (DL) technology was employed in this research as a component of an all-encompassing strategy to accomplish the main goal of early identification of lung cancer and pneumonia. The findings of the investigation proposed two distinct deep learning techniques, one of which involved an initial deep learning approach called Modified Alex Net (MAN) that classified chest X-ray images into normal and pneumonia categories.

The second DL 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 study demonstrated a significant advancement in the use of deep learning methods for early diagnosis and diagnosis of lung cancer, demonstrating the potential of deep learning technology in improving 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 study investigates the performance 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

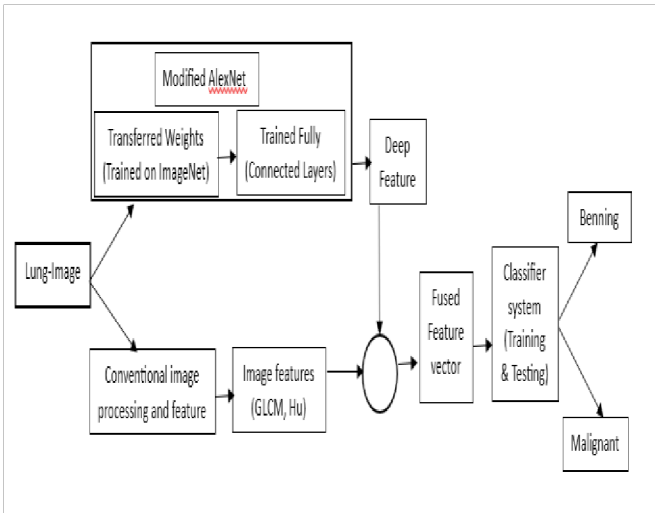

Figure 2. Lung cancer diagnosis using Deep learning

potential method for predicting and determining the stages of lung tumors. As can be seen in fig. 3 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 Waikato University of Technology in New Zealand was home to the research team that was responsible for the development of WEKA. The correlation attribute approach was used throughout the whole of this book for the purpose of picking features [15].

The attribute is basically a feature extraction approach. SVM, KNN, and CNN are the three main types of machine learning classifiers that were used in this study. Each has a distinct level of accuracy. 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].

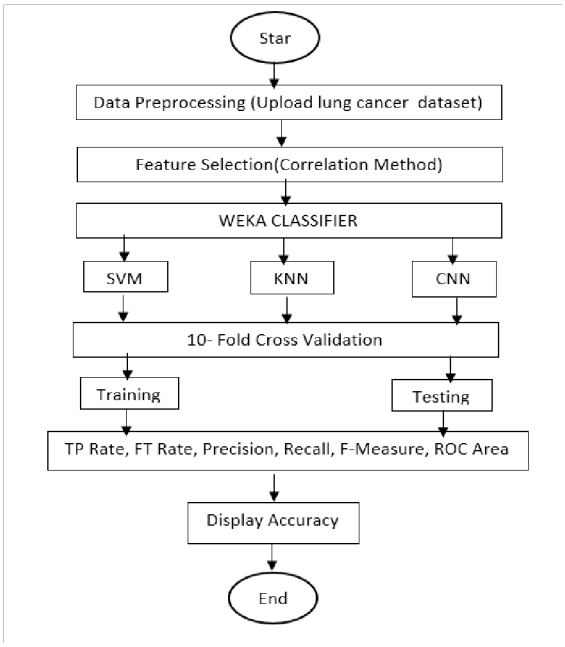

Figure 3. Lung cancer diagnosis using WEKA Technique and correlation method

In 2021, researchers investigated how well CNN classification performed as a method for identifying lung cancer. A deep CNN was skilled. using CT data provided from the Lung Image Database Consortium (LIDC) in order to identify between malignant and noncancerous lung nodules. This was accomplished by utilizing CT data to train the network. The inquiry led to a development that was superior to methods that had been applied in the past. The group preprocessed the CT images in order to standardize the file sizes and formats before importing them into the network model [18]. This was done before the CT pictures were imported into the network model. They were able to reach a degree of accuracy that was more than that of earlier research publications by using the LIDC dataset. This level of accuracy was 91%. According to the outputs of the research, deep CNN were successful in identifying malignant lung nodules and determining whether or not they were cancerous [19].

A group of researchers collaborated in 2022 to submit a study that was based on a method known as SVM analysis, which helps in the diagnosis of lung abnormalities by making use of a text dataset for an evaluation of lung abnormalities. In this particular research project, an SVM-based machine learning model was used in order to enhance the process of lung cancer diagnosis. Patients suffering from lung cancer are sorted into different categories according to the symptoms they exhibit along with SVM classifier, and the model is refined with the help of the computer language Python. Different tumors may be diagnosed using a variety of techniques [20]. However, there are only a few distinct methods for figuring out what populations are present in them. This essay will provide a technique for not only finding malignant tumors but also doing the necessary calculations to determine their location, shape, and size. Thus, in addition to counting and winning, the kind of tumor may also be determined. The development of lung cancer is predicted in this study using SVM. If the prognosis turns out to be true, the physician may be able to devise a more efficient treatment plan for the patient and arrive at a conclusion about their condition sooner. In this study, the recommended model was contrasted using the SVM and SMOTE methodologies. When compared to the current approaches, the suggested method has a 98.8% accuracy rate [21].

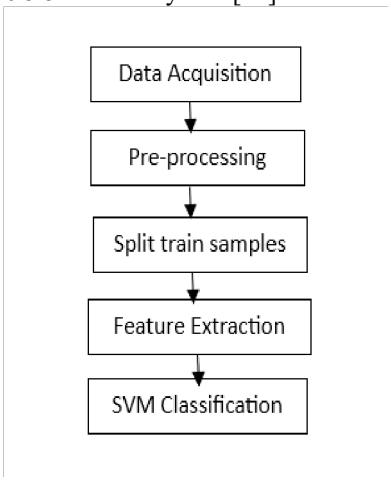

Figure 4. Lung cancer diagnosis using SVM and SMOTE techniques.

A 2022 study examined lung cancer diagnosis and classification using CNN and Google Net deep learning systems. Their focus was using these algorithms to automatically identify and classify lung cancer CT data. CNN and Google Network were used to discover cancerous areas and classify them as normal or abnormal. The CNN algorithm was built using a VGG-16 deep convolutional network architecture. The study found that the algorithm identified and classified lung cancer patients with 98% accuracy. The researchers quantitatively examined the suggested network by computing the confusion matrix and classification accuracy tests. This study showed that deep learning algorithms can identify and classify lung cancer in CT images [22].

A group of scholars investigated using CNN algorithms to analyze CT scans and CT images in 2022. The study used medical and neural network domain knowledge to determine whether a patient had cancer. The deep learning machine generated masks and identified cancer in new instances despite its 65% training set accuracy. It learned from past data.

However, false positives were likely. The project's main tool was a web application that used the model. CNN algorithms can be used to analyze CT images for cancer diagnosis, but they require further development to improve accuracy [23].

In the year 2023, a group of researchers embarked on an investigation utilizing the deep learning methodology introduced as CNN to search for lung nodules in CT scan pictures. These nodules had the potential to be malignant. They created an ensemble technique in order to increase the accuracy of lung nodule identification and overcome the constraints that had been found in earlier studies [24]. They integrated the results of two or more CNNs in beneficial to make more precise predictions of the outcome of the analysis than they could have made using just one deep learning model. By using a 2D CNN strategy, their objective was to circumvent the restrictions imposed by previous research. The researchers integrated the results of the three CNN models, which were given the names CNN1, CNN2, and CNN3, and employed an ensemble deep learning technique to do so. The end result was an accuracy of 95%. [25].

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 [26]. 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 [27]. This method was followed by meticulously carrying out these procedures. While the testing phase allowed us to conduct an objective analysis of the model's performance, the prediction process allowed us to implement the newly designed system in real world contexts. This research primarily focuses on distinguishing between malignant and benign lung tumors, which is an important step toward developing more effective lung cancer treatments [28].

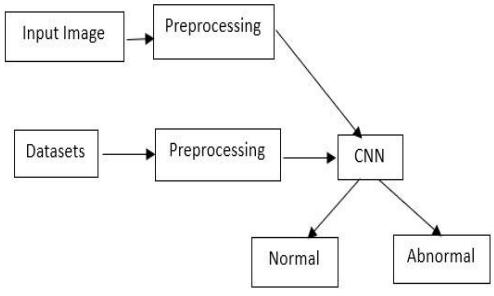

Figure 5. Flow chart of proposed method

3.1 Data Collection

The primary data source for our study was the LC25000 lung and colon histopathological image collection. These images were generated utilizing tissue samples from patients' lungs and colons. The collection provided a comprehensive set of histopathology photographs that were carefully chosen for our research. Throughout our study, our attention was on three specific categories of lung tissue: squamous carcinoma cells, adenocarcinoma cells, and benign tissue. These categories were of particular interest due to their relevance in lung cancer diagnosis and classification. To ensure a robust dataset, a total of 5,000 histopathology photographs from each category have been collected, resulting in a substantial and representative dataset for our analysis.

The histopathology photographs played a pivotal role in our research, serving as the foundation for our investigations and enabling us to accomplish our objectives. By analyzing these images, were able to study the characteristics and features of each tissue category, gaining valuable insights into the histopathological patterns associated with different lung conditions. The extensive collection of histopathology photographs allowed us to conduct a detailed and comprehensive analysis. Advanced image processing methodologies and ML algorithms have been utilized to extract relevant information and identify key patterns and abnormalities associated with each tissue category. This analysis was essential in developing our proposed methodology and in evaluating the effectiveness of the convolutional neural network (CNN) model employed.

By leveraging this diverse dataset, training and validation has been happened of CNN model, enhancing its accuracy and robustness in classifying lung tissue samples. The availability of such a substantial dataset allowed us to draw reliable conclusions and make informed decisions based on the results obtained from experiments.

Preprocessing

The dataset obtained comprised histopathological photographs in the JPEG file format, with each image containing RGB color channels. To ensure consistency and facilitate the functioning of the convolutional neural network (CNN), resized all the images to have an aspect ratio of one and a pixel size of (180, 180). This resizing step aimed to maintain uniformity across the dataset and simplify CNN operations. In order to improve the effectiveness of the training process and expedite convergence, normalized the pixel parameters of the pictures to a range within 0 and 1. This normalization step played a crucial role in enhancing the model's learning capacity and

enabling it to extract relevant features from the data more effectively.

To locate the challenge of limited training data and mitigate the risk of overfitting, various image augmentation techniques have been employed. Specifically, applied horizontal and vertical flips and introduced zooming operations to the histopathological images. By incorporating these augmentation approaches, successfully increased the number of available training images and introduced additional variations in the dataset. This augmentation process played a vital role in enhancing the diversity of the data patterns and expanding the range of information available for the model during training.

It is crucial to remember that overfitting happens when a model becomes overly focused on the training set, which hinders its capacity to generalize effectively to brand-new, untried data. By leveraging data augmentation techniques such as flipping and zooming, aimed to create a more robust and diversified training set. These techniques enabled the neural network to learn a wider range of patterns and characteristics, thus mitigating the risk of overfitting. To illustrate the impact of these augmentation techniques, present an example of a histopathological image of an adenocarcinoma, along with its enhanced versions. These enhanced images were generated by applying horizontal and vertical flips as well as a zoom range of 0.2. By visually demonstrating the original image alongside its augmented counterparts, aim to highlight the influence of these strategies in increasing the diversity within the dataset and expanding the available training data for our model.

Data cleaning has been performed to remove the noisy data and missing value, after that data aggregation has been performed. Multiple features with missing value led to incorrectness so dimensionality reduction has also been performed. The reason behind this when the features are extracted carefully it helps in correctness and accuracy. So data cleaning or preprocessing is the most important step for feature extraction.

Deep Learning

Convolutional neural networks, sometimes known as CNNs, are a subcategory of feed-forward neural networks. These networks were developed based on models of the human visual system. According to our current knowledge of the nature of the visual system, CNNs are built using individual neurons that are structured to react to overlapping areas within their receptive fields. This design is in line with the principles of neural computation [29]. CNNs are able to produce a type of translational invariance thanks to the application of neurons that have the same parameters 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 primary benefit 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. A multi-layered feed-forward neural network is formed by the architecture of a CNN, which is comprised of numerous layers that are placed sequentially on top of one another. CNNs typically consist of convolutional layers followed by activation layers, and some CNNs also include pooling layers after the activation levels. The fundamental components and architecture of a CNN are shown in Fig. 1. This includes the network's three primary layers, which are referred to as the input layer, the convolutional layer, the pooling layer, and the fully connected layer [30].

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.

Deep convolutional neural networks may be used to detect lung cancer, and the method includes two different categories. The first category consists of preprocessing capabilities that are specifically designed to train and process pictures in the deep CNN, which makes feature extraction possible. The second category is concerned with the classification of input CT scans. Here, a deep CNN determines whether a particular nodule is benign or cancerous.

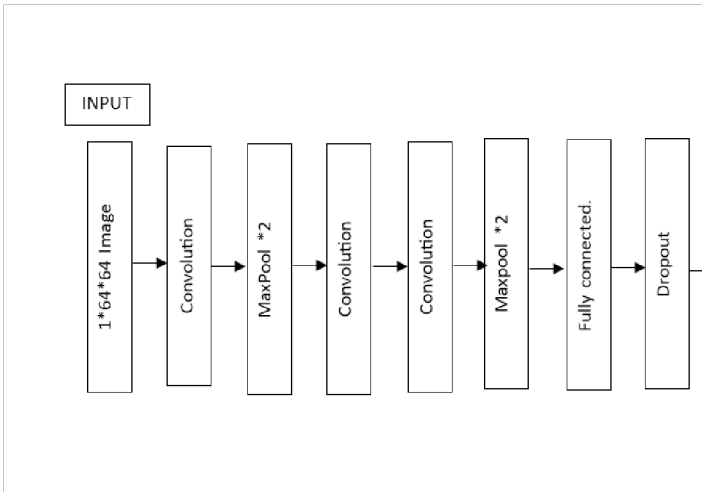

Figure 6. Architecture of CNN Model

The architecture of a deep convolutional neural network (CNN) is a complex and powerful system designed to extract meaningful features from input images and classify them accurately. Fig. 6 provides a visual representation of this architecture, which includes several key components such as ReLU layers, max-pooling layers, a fully connected layer with a SoftMax layer, and an additional fully connected layer. In the CNN architecture, information flows through the network in a sequential manner. Fig. 1 illustrates this flow, starting with the input layer that receives images of a suitable size for the network. The input layer serves as the entry point for the data, allowing it to be processed and transformed throughout the subsequent layers.

The second layer in the CNN architecture is the convolutional layer, which plays a crucial role in feature extraction. It takes the input images, typically sized at 256 x 256, and applies convolutional kernels of size 3 x 3 to them. These convolutional kernels act as filters, scanning the input images and producing feature maps as output. Each feature map represents a specific feature or pattern present in the input images. The convolutional layer is followed by rectified linear unit (ReLU) layers, which introduce non-linearity and help improve the transformation of images into feature maps. ReLU layers apply an activation function that selectively activates or deactivates certain neurons based on their input values, enhancing the network's ability to learn complex patterns. After the convolutional and ReLU layers, the architecture includes max-pooling layers. These layers serve the purpose of down sampling the feature maps generated by the previous layers. The max-pooling layer applies a pooling operation, typically with a filter size of 2 x 2 and a stride of 2 pixels. This operation reduces the spatial dimensions of the feature maps while preserving the most salient features. By down sampling the feature maps, the network becomes more computationally efficient and less susceptible to overfitting as it focuses on the most important information. Following the max-pooling layer, the outputs are passed to a fully connected layer. This layer connects every neuron from the previous layer to every neuron in the subsequent layer, enabling the network to learn complex relationships and capture high-level features. The fully connected layer generates outputs with a dimensionality of 1024, allowing for more abstract representations of the input data. In some cases, an additional fully connected layer may be added, providing further processing and feature refinement.

Finally, a SoftMax layer is introduced at the end of the architecture. The SoftMax layer transforms the outputs of the fully connected layer into probabilities. It applies the SoftMax activation function, which normalizes the values across the classes and ensures that the resulting probabilities sum up to 1. This layer provides the classification probabilities for determining whether the detected cancer type is benign or malignant. By examining the probabilities, a decision can be made regarding the classification of the input images. The effectiveness of the deep CNN architecture is evaluated through experimental studies, and the results are presented in Figs. 7 and 8. These figs. display examples of images after being classified as either malignant or benign based on the trained CNN model. The ability of CNN to accurately classify these images is a testament to its potential in aiding the diagnosis and diagnosis of cancer.

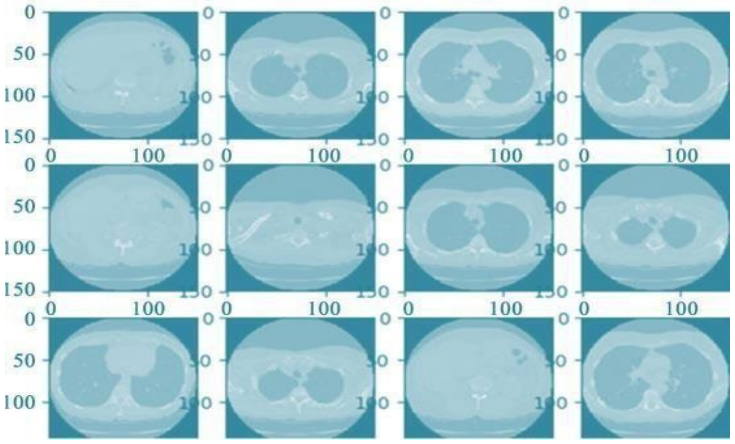

Figure 7. Benign Images

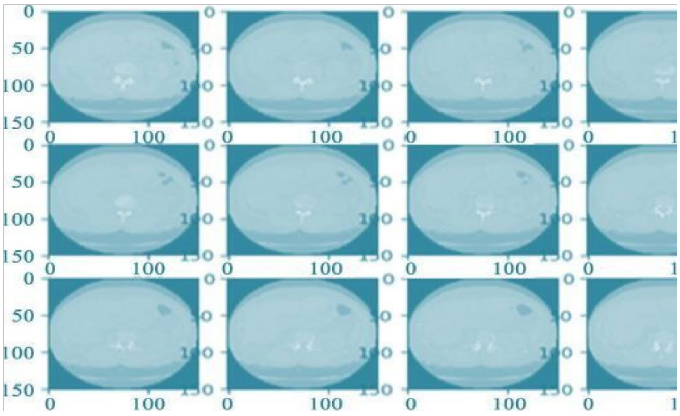

Figure 8. Malignant Images

3.2 Model Training, Evaluation, and Forecasting

CNNs have proven to be powerful tools for image classification and recognition tasks. In the context of CNNs, ConvNet-layered convolutional layers are commonly used to process input images and extract meaningful features. This reconfiguration of the traditional neural network architecture has significantly improved the ability of CNNs to handle complex visual data.

During the training and testing processes of a CNN, multiple convolutional layers are utilized. These layers consist of kernel filters, which are convolved with the input data to capture relevant patterns and features. The use of max pooling further reduces the computational load by down sampling the feature maps obtained from the convolutional layers. Lastly, fully connected layers combine the extracted features and make predictions using the SoftMax activation function, which assigns probabilities to different classes. In some implementations, such as in the case of using Google Collaboratory GPUs, the CNN model can benefit from accelerated processing power, particularly when assigned to "device: GPU:0." This optimization can significantly speed up the training and evaluation of the CNN model. A typical CNN architecture comprises several layers, including an input layer, one or more hidden layers, and an output layer. In the specific neural network described, there was one input layer and three hidden layers. The input layer received images with dimensions of 180 × 180 pixels, with a training/validation data split of 90:10 to ensure appropriate model evaluation. Each convolutional layer in the network employed a 3 x 3 kernel matrix to convolve with the input data. The activation function used in the network was ReLU (Rectified Linear Unit), defined as max (0, x). This activation function introduces non-linearity to the model and aids in capturing complex patterns. To reduce the number of parameters and computational complexity, a maximum pooling operation with a size of 2 x 2 was applied after each convolutional layer.

The final output class probabilities were calculated using the sigmoid activation function, with a dense value of three to match the number of output classes. Additionally, a dropout value of 0.1 was applied to prevent overfitting by randomly dropping out a portion of the neurons during training. The Adam optimizer was employed 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.

$$C.E = -\log(\Sigma_{ce^{Sj}}^{eSp}) \quad (1)$$

The categorical cross-entropy loss function is defined by the equation (1), where C represents 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 guide the learning process of the CNN by penalizing incorrect predictions and encouraging convergence towards accurate class probabilities.

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

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

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

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

To assess the effectiveness of the newly created CNN model, a confusion matrix plot was generated. This matrix allows for a visual representation of the model's performance by comparing the predicted labels with the true labels. Numerous evaluation metrics, like f1-score, precision, recall, and accuracy, were calculated to further analyze the model's performance. Equation (2) measures 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)) combines precision and recall into a single metric that accounts for both factors. 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.

Once the architecture of the CNN model is established and trained, the weights of the trained model are typically saved in the HDF5 file format. These saved weights can be later used for making predictions on new, unseen data. This step allows for the deployment of the trained model in real-world applications, where it can accurately classify and recognize images.

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%. Fig. 9 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. Fig. 10 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.

The high accuracy observed during the validation phase indicates that the CNN architecture shows promising capability in effectively distinguishing between benign and malignant lung tumors. These results suggest that the proposed model could serve as a reliable and valuable instrument to aid radiologists and clinicians in diagnosing lung cancer. Furthermore, the training and validation graphs generated in this study provide valuable insights into the performance of the CNN model throughout the training process. These visual representations allow for a comprehensive analysis of the model's progression and can help identify any potential issues or opportunities for further development.

Overall, this research demonstrates the potential of the trained CNN model as a robust diagnostic tool for accurately identifying lung cancer. The accuracy achieved, along with the visualization of the training and validation progress, contributes to a deeper understanding of the model's performance and its potential applications in clinical settings.

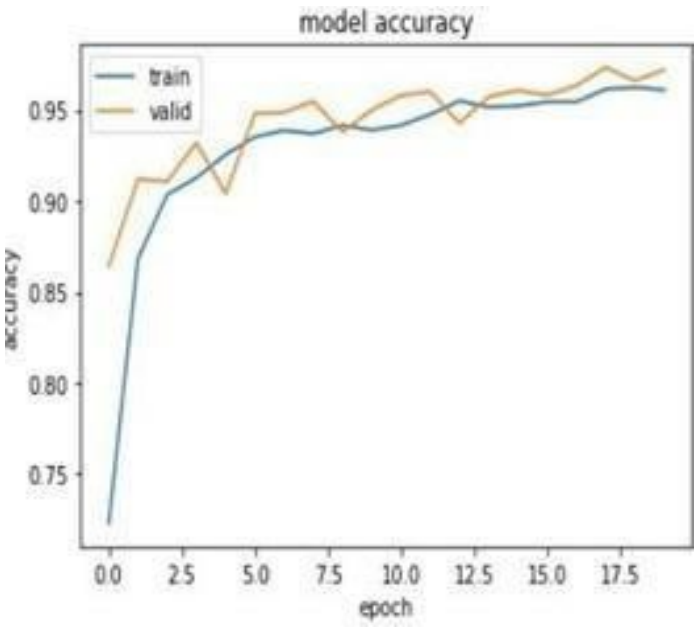

Figure 9. The plot shows the Model accuracy as a function of time for both training and validation images.

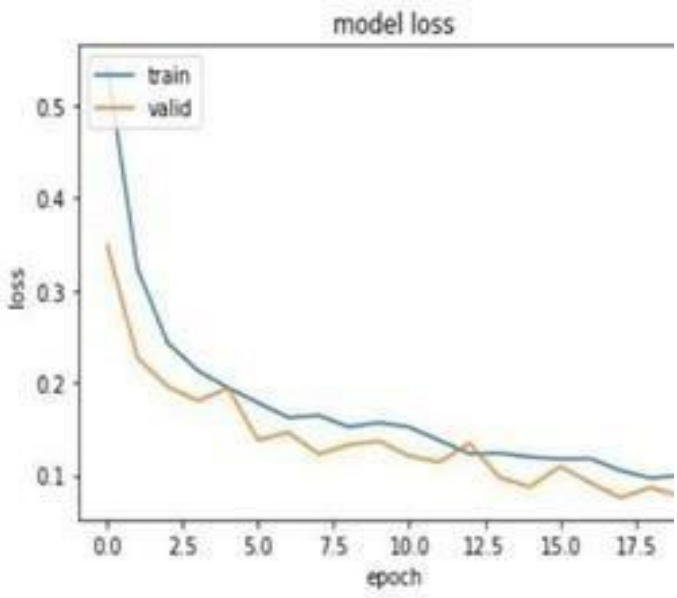

Figure 10. The plot shows the Model Loss as a function of time for both training and validation images.

Table 1. A table that provides a comparison of the eight different review articles published between 2019 - 2022.

References	Year	Methods	Result
[1]	2019	presented a novel architecture that merged 3D CNNs with metaclassifiers.	The fusion approach has 91.23% accuracy and 3.99 false positives per scan.
[3]	2020	Deep learning	This DL framework has an accuracy rate of 97.27%.
[4]	2021	CNN	Using the LIDC dataset, they reached 91% accuracy.
[7]	2021	Correlation Method and WEKA Technique.	SVM – 95.56% KNN – 88.40% CNN – 92.11%
[5]	2022	Using CNN and Google Net deep learning systems.	The algorithm identified 98% of lung cancer patients.
[6]	2022	CNN	65% accuracy
[8]	2022	SVM and SMOTE	The accuracy of this model was 98.8%
[12]	2023	CNN1, CNN2, and CNN3 used ensemble deep learning for their outcomes.	The end result was an accuracy of 95%.

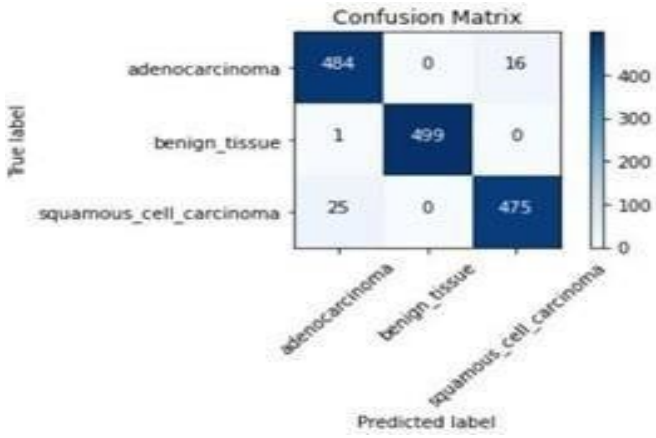

Figure 11. Confusion Matrix for Validation Images Comprised of the Many Different Image Categories

Table 2. Results of CNN

Accuracy of plotted image		Loss function validation with respect to Time	
Training	Validation	Training Time	Validation Time
96.1%	97.2%	0.45	0.35

Fig. 11 shows the confusion matrix, which displays the relationship between the true labels and the predicted labels of the images in the validation data for the labeled categories. The confusion matrix provides a visual representation of the model's efficiency by showing the count of correct and incorrect probabilities for each category. It allows for a comprehensive assessment of the classification accuracy and potential misclassifications. By examining the confusion matrix, researchers can gain insights into the strengths and weaknesses

of the model, identify patterns of misclassification, and make informed decisions regarding further improvements or adjustments to the model. The confusion matrix acts as a valuable method for evaluating the efficiency of the classification model and gaining a deeper understanding of its predictive capabilities.

5. Conclusion

The primary goal of this study was to create an advanced diagnostic tool that could accurately identify cases of lung cancer using histological scans and a CNN. The CNN model was trained on a large dataset of annotated lung cancer photos, and its performance was thoroughly assessed during both the training and validation phases. The CNN model achieved remarkable accuracy, with a training precision of 96.11% and a validation precision of 97.20%. These impressive figures reveal the model's ability to classify and differentiate between cancerous and non-cancerous lung tissue samples. Additional evaluation computes were also performed to thoroughly assess the model's efficacy. The findings of this study highlight the developed CNN model's tremendous potential as a valuable diagnostic tool in clinical settings for precise lung cancer diagnosis. Healthcare professionals can benefit from improved cancer detection capabilities by leveraging machine learning techniques like the CNN used in this study. This, in turn, has the potential to significantly improve patient outcomes through earlier diagnosis and intervention. Overall, this study showcases the power and promise of utilizing advanced machine learning methodologies to aid in cancer identification. The successful application of the CNN model to accurately diagnose lung cancer underscores the importance of leveraging technology to augment and support medical professionals in their critical decision-making processes. A Machine learning based web application or Android application can be developed for the user benefits in future work.

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., Ramya, M., Prajwala, G., & Ramya Geetha, S. V. (2020). Early lung cancer detection using machine learning and image processing. J Eng Sci, 11(7), 510-514.

[3] Bhandary, A., Prabhu, G. A., Rajinikanth, V., Thanaraj, K. P., Satapathy, S. C., Robbins, D. E., ... & Raja, N. S. M. (2020). Deeplearning 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] Jain, D., Singh, P., Pandey, A. K., Singh, M., Singh, H., & Singh, A. (2022, November). Lung Cancer Detection Using Convolutional Neural Network. In 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) (pp. 1-4). IEEE.

[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, 33(4), 401-409.

[17] Khouja, 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.

[19] Ojha, T. R. (2023). Machine Learning based Classification and Detection of Lung Cancer. Journal of Artificial Intelligence, 5(2), 110-128.

[20] Pogue, J. A., Cardenas, C. E., Harms, J., Soike, M. H., Kole, A. J., Schneider, C. S., ... & Stanley, D. N. (2023). Benchmarking automated machine learning enhanced planning with Ethos against manual and knowledge-based planning for locally advanced lung cancer. Advances in Radiation Oncology, 101292.

[21] Chassagnon, G., De Margerie-Mellon, C., Vakalopoulou, M., Marini, R., Hoang-Thi, T. N., Revel, M. P., & Soyer, P. (2023). Artificial intelligence in lung cancer: current applications and perspectives. Japanese Journal of Radiology, 41(3), 235-244.

[22] Huang, S., Yang, J., Shen, N., Xu, Q., & Zhao, Q. (2023, January). Artificial intelligence in lung cancer diagnosis and prognosis: current application and future perspective. In Seminars in Cancer Biology. Academic Press.

[23] Saravanan, T. M., Jagadeesan, M., PA, S., Mubarak, M. M., Kumar, S. P., & Sanjay, S. (2023, January). A CNN Based Machine Learning Scheme to Detect Lung Cancer Detection from CT Scan Images. In 2023 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.

[24] Bhattacharjee, S., Saha, B., & Saha, S. (2023, January). Prediction of Recurrence in Non Small Cell Lung Cancer Patients with Gene Expression Data Using Machine Learning Techniques. In 2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE) (pp. 1-8). IEEE.

[34] P.Durgadevi, S.Vijayalakshmi, Munish Sabharwal. (2021). Fetal Brain Abnormality Detection through PSO (Particle Swarm Optimization) and Volume Estimation. Annals of the Romanian Society for Cell Biology, 2700–2714.

[25] Fatoki, F. M., Akinyemi, E. K., & Phlips, S. A. (2023). Prediction of Lungs Cancer Diseases Datasets Using Machine Learning Algorithms. Current Journal of Applied Science and Technology, 42(11), 15-23

[26] Mohammed, B. A., Senan, E. M., Alshammari, T. S., Alreshidi, A., Alayba, A. M., Alazmi, M., & Alsagri, A. N. (2023). Hybrid Techniques of Analyzing MRI Images for Early Diagnosis of Brain Tumours Based on Hybrid Features. Processes, 11(1), 212. MDPI AG. Retrieved from http://dx.doi.org/10.3390/pr11010212

[27] Minna, J. D., Roth, J. A., & Gazdar, A. F. (2002). Focus on lung cancer. Cancer cell, 1(1), 49-52.

[28] Leiter, A., Veluswamy, R. R., & Wisnivesky, J. P. (2023). The global burden of lung cancer: current status and future trends. Nature Reviews Clinical Oncology, 20(9), 624-639.

[29] Oliver, A. L. (2022). Lung cancer: Epidemiology and screening. Surgical Clinics, 102(3), 335-344.

[30] Al Bakir, M., Huebner, A., Martínez-Ruiz, C., Grigoriadis, K., Watkins, T. B., Pich, O., ... & Swanton, C. (2023). The evolution of non-small cell lung cancer metastases in TRACERx. Nature, 1-10.

[31] Azween Abdullah, S. Nithya, M. Mary Shanthi Rani, S. Vijayalakshmi and Balamurugan Balusamy, "Stacked LSTM and Kernel-PCA-based Ensemble Learning for Cardiac Arrhythmia Classification" International Journal of Advanced Computer Science and Applications(IJACSA), 14(9), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0140905

[32] Tanwar, Sushama & Vijayalakshmi, S. & Sabharwal, Munish & Kaur, Manjit & Ali, Ahmad & Lee, Heung-No. (2022). Detection and Classification of Colorectal Polyp Using Deep Learning. BioMed Research International. 2022. 1-9. 10.1155/2022/2805607.

[33] Gokul Rajan V., S. Vijayalakshmi, Munish Sabharwal. (2021). Modified CNN based Feature Extraction for Sclera Recognition of Off . Annals of the Romanian Society for Cell Biology, 2579–2590.