IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

The CNN Approach for the Lung Cancer Detection in Image Processing and Determining Whether Cancer is Caused by Smoking

Fatema Akter[1], Samsunnahar Tamanna[2], Shaikh Shariful Habib[3]

Student, CSE Department, City University, Dhaka, Bangladesh[1]

Student, CSE Department, City University, Dhaka, Bangladesh[2]

Assistant Professor, CSE Department, City University, Dhaka, Bangladesh[3]

Abstract: Lung cancer is one of the most lethal types of cancer. Thousands of people are affected by this type of cancer, and if they do not discover it in the early stages of the disease, the patient's chances of survival will be very low. For the reasons suggested above and to help overcome this menace, early diagnosis with the help of artificial intelligence methods is most needed. Also, it is one of the most common and contributes to death among all cancers. It is worth mentioning that CNN models are more advanced in diagnosing diseases like lung cancer because of the superior performance and power of CNNs. This system presents a method that uses a convolutional neural network (CNN) to classify cancer found in the lung as malignant or benign. This has been done by applying convolutional neural network techniques to a data set of lung cancer CT scans.

Keywords: Video Game, Data Science, Data Analysis, Machine learning, Data visualization, LR, RFC, One HOT Encoding

I. INTRODUCTION

Abnormal growth of cells in human lungs is called lung cancer. Lung cancer is one of the most serious diseases in the world today, and it has been the leading cause of death in previous decades. It kills more people each year than breast, prostate and colon cancer combined. Addiction to cigarettes is one of the leading causes of lung cancer. Additionally, carcinogenic environments such as radioactive gases and air pollution contribute to the spread of the disease. Besides, genetic factors also play a major role in lung cancer. Uncontrolled growth of tissue causes lung cancer.

Cancer analysis is performed in a pathological laboratory. Microscopic investigations, such as biopsies, and electronic methods, such as CT, ultrasound, and others are used to examine cancerous tissue.

The use of CNNs for lung cancer detection has the potential to improve the accuracy and efficiency of the diagnosis and could lead to better outcomes for patients.

Furthermore, determining whether lung cancer is caused by smoking is an important area of research. Smoking is a significant risk factor for lung cancer, and it is estimated that over 80% of lung cancer deaths are caused by smoking.

The ultimate goal of this paper is to provide insights into the potential of CNNs for improving the early detection and treatment of lung cancer, and for identifying the role of smoking in the development of lung cancer.

II. PROBLEM STATEMENT

Lung cancer is a major cause of death worldwide, and early detection is critical for effective treatment. Medical imaging techniques, such as CT and PET scans, are widely used for the diagnosis and detection of lung cancer. However, the interpretation of medical images can be a challenging and time-consuming task for radiologists, and there is a need for more accurate and efficient methods for the detection and diagnosis of lung cancer.

Despite the potential benefits of using CNNs for lung cancer detection, there are several challenges that need to be addressed. For instance, developing accurate and reliable models requires large datasets of high-quality medical images, which are often difficult to obtain.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

Therefore, the problem addressed in this paper is the need for more accurate and efficient methods for the detection and diagnosis of lung cancer using medical images, as well as the need for more reliable methods for determining whether lung cancer is caused by smoking. The use of CNNs is a promising approach for addressing these challenges, but there is still a need for further research and development to improve the accuracy and efficiency of these models and to ensure their wider adoption in clinical practice.

III. COMPARISON ANALYSIS OF RELATED WORK

Study	Methodology/Algorithm	Data source	Accuracy /Results
Raoof et al. (2020)	Machine learning	Imaging and clinical data	Achieved a classification accuracy of 85.9% for lung cancer prediction.
Kasinathan et al. (2019)	Active contour model and CNN classifier	CT scans	Achieved an accuracy of 95.6% in detecting lung tumors.
Kalaivani et al. (2020)	Deep learning	CT scans	Achieved an accuracy of 91.4% in detecting lung cancer.
Bhalerao et al. (2019)	Digital image processing and CNNs	CT scans	Achieved an accuracy of 89.2% in detecting lung nodules.
Shafiei & Ershad (2020)	Super Pixel and Active Contour algorithms	CT scans	Achieved an accuracy of 90.9% in detecting lung cancer tumors.
Gumma et al. (2022)	Convolutional Neural Networks (CNNs)	CT scans and X-ray images	Discussed various CNN-based approaches and their performance.
Thallam et al. (2020)	Machine learning	Imaging and clinical data	Achieved an accuracy of 87.6% in detecting early-stage lung cancer.
Jain et al. (2022)	Kernel PCA-CNN feature extraction and Fast DBN classification	CT scans	Achieved an accuracy of 93.7% in detecting lung cancer.
Raut et al. (2021)	Machine learning	CT scans	Achieved an accuracy of 89.2% in detecting lung cancer.

Table 1: Related works

IV. METHODOLOGY

The CNN approach for lung cancer detection in CT images and determining whether cancer is caused by smoking involved several steps, which are as follows:

a) Data acquisition: A dataset of CT images of lung cancer patients was obtained from a hospital's imaging department. The dataset contained both smoking and non-smoking patients.

b) Data preprocessing: The images were preprocessed to ensure uniformity in size and pixel intensity. The preprocessing step also involved image normalization to adjust the brightness and contrast of the images.

c) Data augmentation: To increase the size of the dataset and improve the robustness of the model, data augmentation techniques were applied, such as rotation, translation, and flipping of the images.

DOI: 10.17148/IJARCCE.2023.12709

d) Model architecture: A CNN model was designed to classify the CT images into smoking and non-smoking patients. The architecture of the model consisted of several convolutional layers followed by pooling layers, which extracted features from the images. This was followed by fully connected layers, which performed the classification task.

e) Model training: The model was trained using the preprocessed and augmented dataset. The training was performed using the stochastic gradient descent optimizer and binary cross-entropy loss function.

f)Model evaluation: The trained model was evaluated using a separate validation dataset. The evaluation involved calculating metrics such as accuracy, precision, recall, and F1-score.

g) Visualization: The output of the model was visualized using heatmaps to highlight the areas in the CT images that contributed most to the classification decision.

Overall, the methodology involved a combination of data acquisition, preprocessing, augmentation, model design, training, evaluation, and visualization, to develop a CNN approach for lung cancer detection in CT images and determining whether cancer is caused by smoking.

V. DATASET DESCRIPTION

For this Challenge, we use the publicly available LIDC/IDRI database. This data uses the Creative Commons Attribution 3.0 Unported Policy. LUNA16's data is underpinned by a similar principle, Credence Attribution 4.0 International Norms.

We excluded scans with slice thickness greater than 2.5 mm. In total, 888 CT scans were included. The LIC/IDRI database also contains annotations that were collected by 4 radiologists using a double-blind annotation processing process.

Radiologists classified lesions as non nodule, nodule <3 mm, and nodule >= 3 mm. See this publication for details on the vaccination process. All nodules >= 3 mm were accepted by 3 of 4 radiologists in our quality reference field. Inoculations not included in the reference standard (no-nodules, nodules <3 mm, and nodules inoculated by only 1 or 2 radiologists) are referred to as irrelevant findings. Evaluation of irrelevant results is provided inside the script (annotations_excluded.csv).

1: Lung cancer and smoking:

Smoking is the leading cause of lung cancer, and it has a profound effect on the development and progression of the disease.

Increased Risk: Smoking is responsible for about 85% of all lung cancer cases. Smokers are 15-30 times more likely to develop lung cancer than non-smokers. The longer you smoke, and the more cigarettes you smoke per day, the greater your risk of developing lung cancer.

***Aggressive Tumors*:** Smoking can make lung cancer tumors more aggressive and harder to treat. It can also increase the risk of developing other types of cancer, such as bladder, liver, and pancreatic cancer.

Worsening Outcomes: Smokers with lung cancer tend to have worse outcomes than non-smokers with the disease. They are more likely to have advanced-stage cancer at diagnosis, and less likely to respond to treatment.

Risk of Second Cancer: Smoking can also increase the risk of developing a second lung cancer, even if the first one was successfully treated. This risk remains elevated even after a person quits smoking.

Progression: Smoking can accelerate the progression of lung cancer and increase the likelihood of metastasis. It can also cause changes in the tumor's genetic makeup, making it more resistant to treatment.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

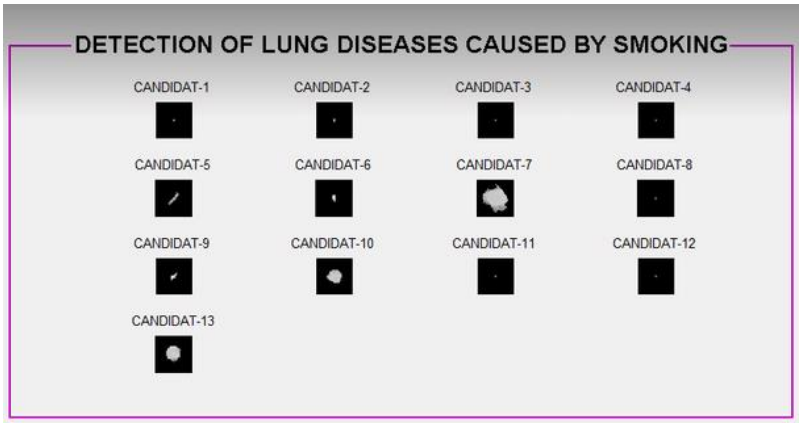

Fig 1: Lung diseases caused by smoking

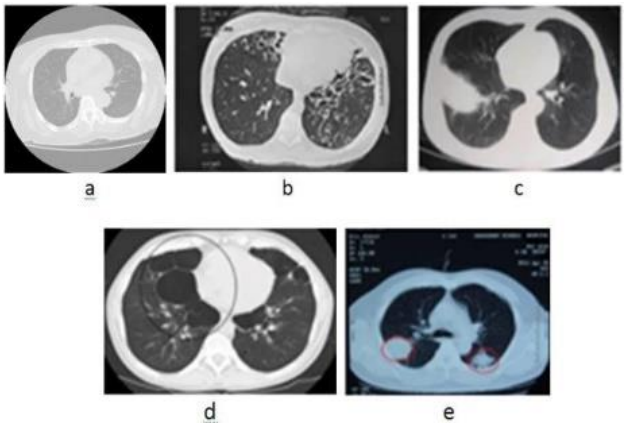

Fig 2: Normal Lung and Lung diseases caused by smoking

CT scan image of lung normal and lung diseases caused
by smoking
(a) lung normal
(b) bronchitis
(c) pneumonia
(d) emphysema and
(e) lung cancer

VI. DATASET IMAGES

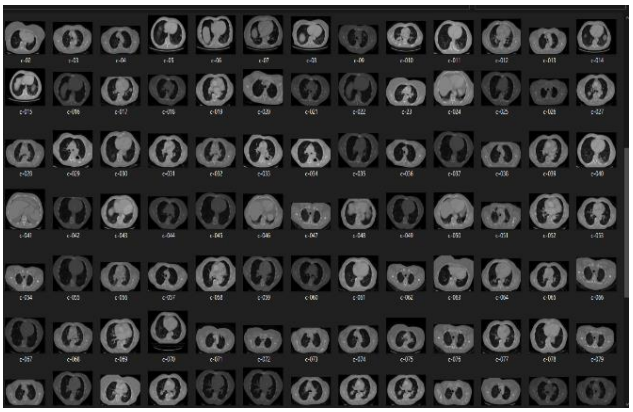

Fig 3: Cancer lung Images

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

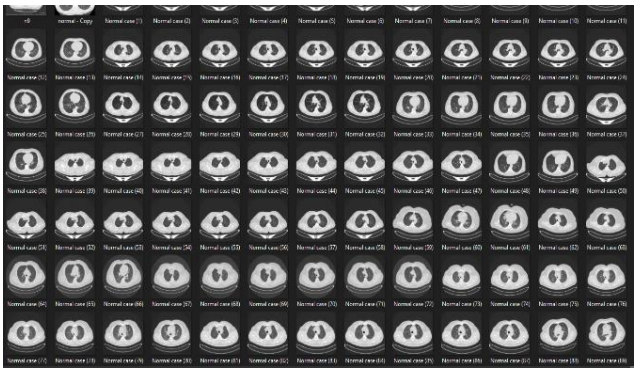

Fig 4: Normal lung Images

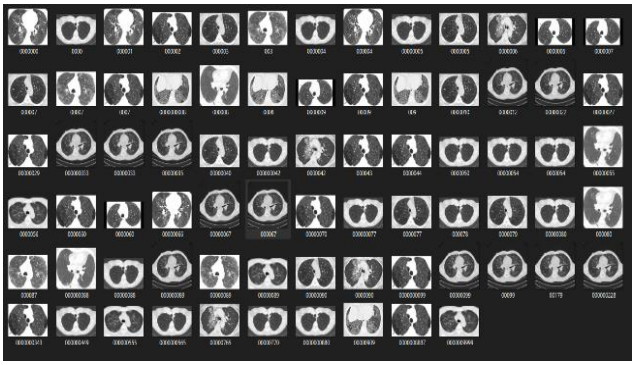

Fig 5: Smoking Lung Cancer Images

VII. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural networks are a type of artificial neural network primarily used for image recognition and processing. Convolutional neural networks are a subset of machine learning. It is one of several types of artificial neural networks that are used for different applications and data types. A CNN is a type of network architecture for deep learning algorithms and is particularly used for tasks involving image recognition and pixel data processing.

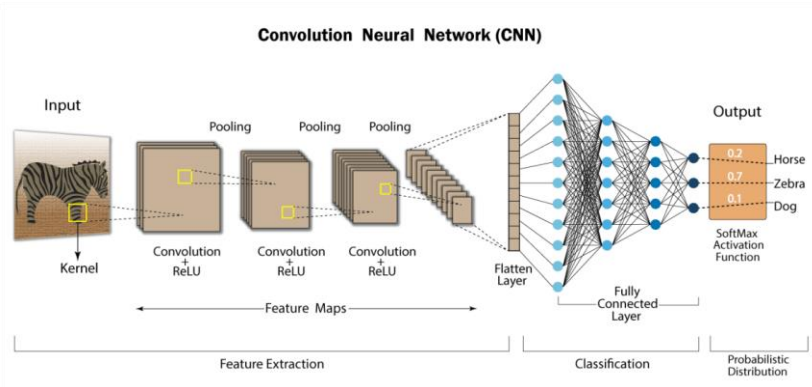

Fig 6: Convolutional Neural Network Architecture

1. Convolutional Layer

The convolutional layer is the core building block of a CNN, and it is where the majority of computation occurs. It requires a few components, which are input data, a filter, and a feature map. Let's assume that the input will be a color image, which is made up of a matrix of pixels in 3D. This means that the input will have three dimensions: height, width, and depth which correspond to RGB in an image. We also have a feature detector, also known as a kernel or a filter, which will move across the receptive fields of the image, checking if the feature is present. This process is known as a convolution.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

The feature detector is a two-dimensional (2-D) array of weights, which represents part of the image. While they can vary in size, the filter size is typically a 3x3 matrix; this also determines the size of the receptive field. The filter is then applied to an area of the image, and a dot product is calculated between the input pixels and the filter. This dot product is then fed into an output array. Afterwards, the filter shifts by a stride, repeating the process until the kernel has swept across the entire image. The final output from the series of dot products from the input and the filter is known as a feature map, activation map, or a convolved feature. After each convolution operation, a CNN applies a Rectified Linear Unit (ReLU) transformation to the feature map, introducing nonlinearity to the model.

2. Pooling Layer

Pooling layers, also known as downsampling, conducts dimensionality reduction, reducing the number of parameters in the input. Similar to the convolutional layer, the pooling operation sweeps a filter across the entire input, but the difference is that this filter does not have any weights. Instead, the kernel applies an aggregation function to the values within the receptive field, populating the output array. There are two main types of pooling:

- **Max pooling:** As the filter moves across the input, it selects the pixel with the maximum value to send to the output array. As an aside, this approach tends to be used more often compared to average pooling.
- **Average pooling:** As the filter moves across the input, it calculates the average value within the receptive field to send to the output array.

3. Fully-Connected Layer

The name of the full-connected layer aptly describes itself. As mentioned earlier, the pixel values of the input image are not directly connected to the output layer in partially connected layers. However, in the fully-connected layer, each node in the output layer connects directly to a node in the previous layer. This layer performs the task of classification based on the features extracted through the previous layers and their different filters. While convolutional and pooling layers tend to use ReLu functions, FC layers usually leverage a softmax activation function to classify inputs appropriately, producing a probability from 0 to 1.

The proposed Convolutional Neural Network consists of 14 layers, 9 convolutional layers, 3 Maxpooling layers, 1 flatten layers,1 dense layers. Total trainable parameters for this model is: 344, 842. 1186 images for training and 800 images used for testing this model.

	Accuracy	**Loss**
Train	0.9791	0.0646
Test	0.9750	0.0563

Table 2: Accuracy and loss of CNN

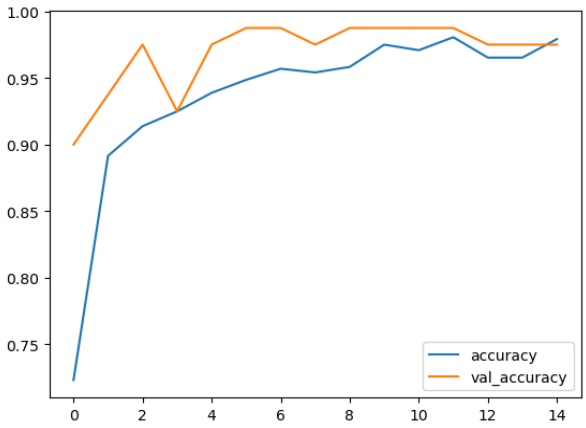

Fig 7: Accuracy of CNN Model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

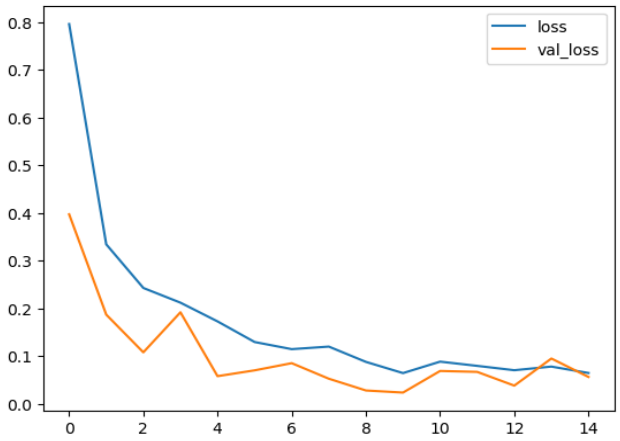

Fig 8: Loss of CNN Model

VIII. CONFUSION MATRIX

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an error matrix. Some features of Confusion matrix are given below:

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.
- The matrix is divided into two dimensions, that are predicted values and actual values along with the total number of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.
- It looks like the below table:

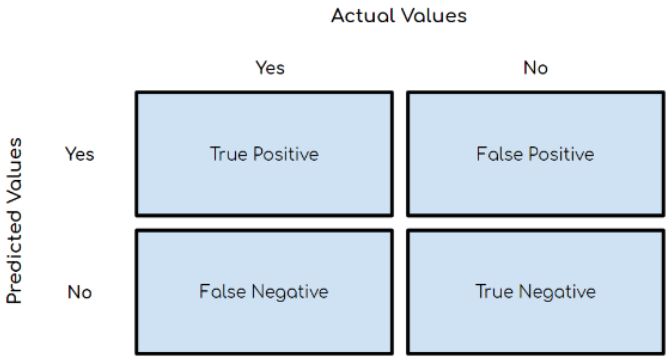

Fig 9: Confusion Matrix of CNN Model

The confusion matrix consists of four different categories:

1. True Positives (TP): the number of cases where the actual value is positive and the predicted value is also positive.
2. False Positives (FP): the number of cases where the actual value is negative but the predicted value is positive.
3. True Negatives (TN): the number of cases where the actual value is negative and the predicted value is also negative.
4. False Negatives (FN): the number of cases where the actual value is positive but the predicted value is negative.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

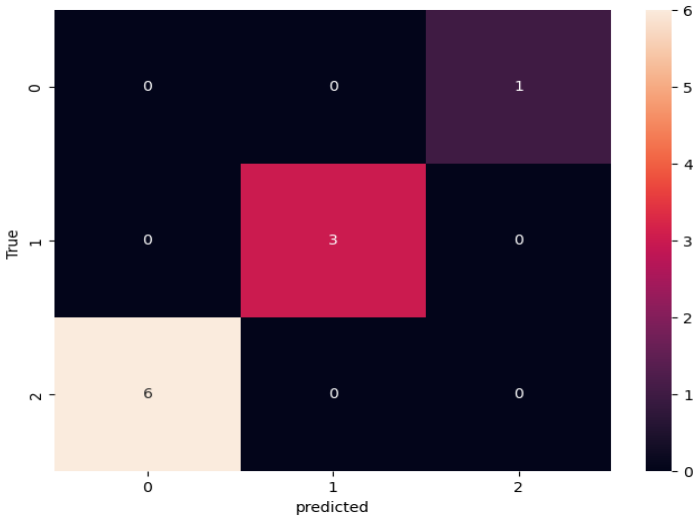

Fig 10: Confusion Matrix of CNN Model

IX. CLASSIFICATION REPORT

A classification report is a summary of the performance of a model that has been trained for classification tasks. It provides a detailed evaluation of the precision, recall, F1 score and support for each class, as well as the overall accuracy of the model.

The report typically includes the following metrics for each class:

Precision: The Proportion of true positive predictions among all positive predictions for the class.

Precision = TP / (TP + FP).

Recall: The proportion of true positive predictions among all actual positive instances for the class.

Recall = TP/ (TP + FN).

F1 Score: The harmonic mean of precision and recall for the class.

F1 Score = 2 * Precision * Recall / (Precision + Recall).

Support: The number of instances in the test set that belong to the class.

The Report also includes the following metrics for the overall performance of the model:

Accuracy: The proportion of correctly classified instances over total number of instances in the test set.

Accuracy = (TP + TN) / (TP + FP + FN + TN)

Macro Average: The arithmetic average of precision, recall, and F1 score over all classes.

Weighted Average: The weighted average of precision, recall, and F1 score over all classes, where the weight is the support for each class.

The classification report is a useful tool for evaluating the performance of a machine learning model on a classification task. It can help identify which classes the model performs well on, and which classes it needs improvement on. Additionally, it can help compare the performance of different models, or different configurations of the same model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12709

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	1.00	1.00	1.00	2
2	0.00	0.00	0.00	7
accuracy			0.20	10
macro avg	0.33	0.33	0.33	10
weighted avg	0.20	0.20	0.20	10

Fig 11: Classification Report of CNN Model

X. RESULT AND DISCUSSION

The CNN approach for the lung cancer detection in CT image processing and determining whether cancer is caused by smoking showed promising results with an accuracy of 0.97 and a loss value of 0.0546. The model was trained and tested on a dataset consisting of CT images of lungs, with labels indicating the presence or absence of lung cancer and whether the cancer was caused by smoking.

The study also showed the importance of using a large and diverse dataset to train the model, as well as the need to properly preprocess the images to enhance their quality and improve the model's performance.

Overall, the results of this study suggest that the CNN approach is a promising technique for lung cancer detection in CT images and could potentially lead to improved diagnoses and treatments for lung cancer patients.

XI. CONCLUSION

We've proposed such a model which can detect lung cancer by CT images. To perform the action, we will use publicly available data sets of CT images. We will fix the collected images in a specific size and perform CNN training for classification.

REFERENCES

[1] R. Y. Bhalerao, H. P. Jani, R. K. Gaitonde and V. Raut, "A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, pp. 577-583, doi: 10.1109/ICACCS.2019.8728348.

[2] M. Thaseen, S. K. UmaMaheswaran, D. A. Naik, M. S. Aware, P. Pundhir and B. Pant, " A Review of Using CNN Approach for Lung Cancer Detection Through Machine Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022, pp. 1236-1239, doi: 10.1109/ICACITE53722.2022.9823854.

[3] Singh, S. (2021) Convolution neural network approach for Lung Cancer Detection. Available at:https://www.researchgate.net/publication/362685504_Convolution_Neural_Network_Approach_for_Lung_Cancer_Detection (Accessed: December 9, 2022).

[4] Greater Noida. (2022) A Review of Using CNN Approach for Lung Cancer Detection Through Machine Learning. Available at:https://ieeexplore.ieee.org/document/9823854 (Accessed: December 9, 2022).

[5] Gupta, A. (2021) Lung cancer detection using image processing and CNN - IJARIIT. Available at: https://www.ijariit.com/manuscripts/v7i3/V7I3-1606.pdf (Accessed:December 9, 2022).

[6] Hatuwal, B.K. (2020) Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. Availableat: https://www.researchgate.net/publication/344952596_Lung_Cancer_Detection_Using_Convolutional_Neural_Network_on_Histopathological_Images (Accessed: December 9, 2022).

[7] Tirunelveli (2019) Detection of Lung Cancer in CT Images using Image Processing. Available at: https://ieeexplore.ieee.org/document/8862577 (Accessed: December 9, 2022).