Full text open access online (Since 2001)
© KRE Kamla-Raj IJHG 2023
PRINT: ISSN 0972-3757 ONLINE: ISSN 2456-6330

Int J Hum Genet, 23(2-3): 168-176 (2023)
DOI: 10.31901/24566322.2023/23.2-3.853

A Novel Artificial Intelligence Framework for Analysis and Classification of Lung Cancer Images

T. Amudha, Subhiksha R. Mallya and J. Satheeshkumar*

Department of Computer Applications, Bharathiar University, Coimbatore 641 046, Tamil Nadu, India

KEYWORDS AI Modelling. Convolutional Neural Networks. Deep Learning. Image Classification. Lung Cancer

ABSTRACT Artificial Intelligence (AI) is an upcoming technology, which is becoming more prevalent in almost all walks of human life. It is no surprise that Artificial Intelligence is being used to solve one of the most complex and dynamic diseases known to mankind – cancer. Lung cancer is a fatal disease to a great degree, claiming millions of lives all around the world every year. Cancer detection and identification is almost always impossible at an early stage. AI is capable of identifying, diagnosing as well as predicting cancer occurrence. The AI framework employs the Deep Learning technique of Convolutional Neural Networks (CNN). This research paper demonstrates the ability of AI and Deep Learning Framework to identify and differentiate between three types of cases, that is, normal, benign and malignant cases. A comparison is done between existing research models applied on the same dataset and the proposed model surpassed the other models by achieving 100 percent training and testing accuracies.

INTRODUCTION

Cancer is a terrifying and cure-evading disease, triggered by the unconstrained and abnormal growth of cells in the human body. Detection of cancer is not as easy as it may sound, and quite often many cancer cases go undetected. Some of the factors that contribute towards cancer going undetected are that cancer does not exhibit symptoms until the Terminal Stage or Stage 4 of cancer, where the tumour starts being visible in the X-Ray or MRI (Magnetic Resonance Imaging). Pathologists and radiologists have seen fewer patients individually than combined consultations, and hence their experience falls short as the cancerous tumour is different in every other person. Doctors may diagnose cancer and set a treatment plan but since cancer's growth is unpredictable it becomes harder to save the patient's life. Now the need has finally arisen to detect and identify cancer in an early stage or better even predict the occurrence of cancer well in advance which enables the patient to undergo early treatment and increase the chances for survival. This paper focuses on developing an efficient AI framework using deep learning methodology to detect and identify the type of tumour in three categories, that is, benign, malignant and normal using benchmark Computed Tomography (CT) scan images.

A survey conducted by Global Cancer Statistics in 2020 has estimated a total death of 10.0 million people in 185 countries around the world due to 36 cancer types, in which breast cancer cases stands at the top with 23 lakh patients identified and followed by lung cancer with 22 lakh patients newly identified (Sung et al. 2021). About 90 percent of lung cancers are missed due to observer error, recognition, decision-making mistakes, and also due to the quality of the chest X-rays, radiographs and CT scans. The researchers suggested alternatives such as CAD (Computer-Aided Design) systems to aid radiologists to identify cancerous lesions from the CT scans and chest X-rays and thereby reducing the possibility of misdiagnosis as well as medicolegal problems (Ciello et al. 2017). AI is found to play a significant role in automating various clinical and medical processes, especially in Genomic Diagnostics such as Variant Calling, Genome Annotation and Variant Classification, Coding and Non-Coding Variants Classification, Phenotype-Genotype Mapping, Genetic Diagnosis and Electronic Health Record to Genetic Diagnosis in a study conducted by re-

**Address for correspondence:*
J. Satheeshkumar
Professor
Department of Computer Applications,
Bharathiar University,
Coimbatore 641 046, Tamil Nadu, India
E-mail: jsatheesh@buc.edu.in

searchers to find the depth of influence of AI. The tools employed by AI in the above mentioned processes include advanced usage of CNNs, Recurrent Neural Network and other Deep Learning technologies (Nesamalar et al. 2021b). A comparison between three Clustering algorithms on 700 cancer affected DNA datasets showed that hierarchical clustering algorithms performs better with high accuracy (Nesamalar et al. 2021a). In a review on AI in cancer data analysis, the researchers state that AI is a very efficient and cost effective form of fighting against cancer. AI is being implemented in all aspects of cancer related procedures from diagnosis, prognosis, treatment to drug discovery and repurposing. The paper also highlights various companies and establishments in pursuit of a cure for cancer (Amudha et al. 2021).

Researchers developed a Computer-aided system (CAD) using the principles of CNN and AlexNet architecture with Transfer Learning. The dataset used was IQ-OTH/NCCD and the model results in 95.714 percent sensitivity measure and 95 percent specificity measure (Al-Yasriy et al. 2020). In a comparison study conducted by researchers between CNN and Artificial Neural Network (ANN) on the dataset IQ-OTH/NCCD dataset for Lung Cancer Classification, CNN proved to yield better results with accuracy of 98.41 percent when compared to ANN (Prarthana and Bhavani 2021). Researchers developed a computer system using Image Processing techniques, which involves feature extraction, image segmentation and enhancement, and applied the Support Vector Machine (SVM) algorithm to identify three classes in the dataset IQ-OTH/NCCD, that is, normal, malignant, and benign. SVM used with Polynomial kernels gave an accuracy of 89.8876 percent (Kareem et al. 2021).

A comparison study between Manual CNN and AlexNet architecture on the dataset IQ-OTH/NCCD was conducted to classify the Chest CT scans into three categories, that is, benign, normal and malignant. The Manual CNN gave a 96.88 percent validation accuracy (Gupta et al. 2021). Researchers employed the Transfer Learning technique and GoogLeNet Deep Neural Network architecture aiming at the detection of lung cancer in the dataset IQ-OTH/NCCD and the proposed model achieved an accuracy of 94.38 percent (AL-Huseiny and Sajit 2021). From the UCI repository datasets, Faisal et al. (2018) evaluated eight different kinds of Classifiers and Ensembles in Machine Learning for timely prediction of the lung cancer. Gradient Boosted Tree technique achieved an accuracy of 90 percent (Faisal et al. 2018). An experiment was conducted on six different breast cancer datasets to define the best feature selection that helps the Machine Learning algorithms to achieve better prediction results. The proposed work with Logistic Regression model and feature selection using Genetic Algorithm gives better outcome and accuracy than k-Nearest Neighbour on all the six different datasets (Jijitha and Amudha 2020).

Objectives

The objectives of this research are:

1. To design and develop a novel AI framework to analyse and classify lung cancer data.
2. To test the developed model on a real-time dataset and measure its performance.
3. To substantiate the developed model efficiency by comparing the result with the existing research models.

RESEARCH METHODOLOGY

Concepts Applied

Artificial Intelligence (AI)

AI is the technique of duplicating the intelligence of mankind and other living beings in the form of machines where the machine is capable of doing the work of more than one human being. AI is said to be the mechanical model or replica of intelligent living beings.

Machine Learning (ML)

Machine Learning deals with the study of how machines can be made to learn, work and improve automatically through experience. This is possible through the use of numerous ML algorithms, which facilitate the detection of hidden relations and connections from the given data and brings new information to light. ML is also defined as training the machine to do something and learn from the results and produce better results.

Deep Learning

Deep Learning is a subdivision that comes under Machine Learning. Deep Learning is basically a neural network, which simulates the workings of the human brain. It is a part of Artificial Neural Networks, which contributes towards learning in three ways, namely, Supervised, Unsupervised and Reinforcement.

Convolutional Neural Network (CNN)

CNN falls under the category of Deep Learning algorithms built upon ANN, which is a computational representation of the human brain. CNN adopts Artificial Intelligence methodologies and is primarily applied for image analysis tasks.

Tensorflow

It is an open source software and library suite, created and maintained by Google Inc. and dedicated towards development of AI, ML and DL frameworks. Tensorflow helps in building, training, testing AI applications on a large-scale basis with a huge amount of data or Big Data in Python programming language.

Keras

Keras is an API and open source software library, used as an interface for Tensorflow library. It focuses on development of the Artificial Neural Networks using Python.

Figure 1 gives a short description of the steps involved in the research methodology for the given problem.

Problem Definition

The problem has been identified as a need for detection of cancerous tumours at an early stage in the individual's lungs with the help of their lung CT scan images. AI should be able to find out what the human eye has missed. The main objective is to create an AI framework to learn and classify the tumours, types of tumours and healthy lungs from the images of the lung scans.

Selection of Dataset

Images of CT scans of healthy and cancer affected lungs are selected for this research. The dataset selected is a set of labelled images as it is supervised learning. The IQ-OTH/NCCD Lung Cancer Dataset is accessed from Mendeley Data repository. This dataset consists of a number of jpeg images of three types, that is, Benign Cas-

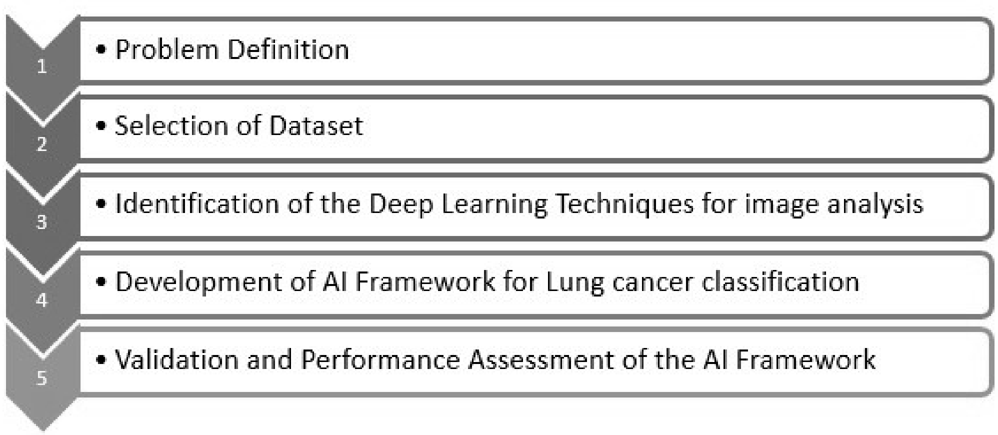

Fig.1. Diagrammatic representation of research methodology

es, Malignant Cases and Normal Cases. These data are collected by "The Iraq-Oncology Teaching Hospital/National Centre for Cancer Diseases (IQ-OTH/NCCD)" specialty hospitals from various provinces in Iraq. This dataset has 1097 CT scan images with 110 patients, of which 55 are normal patients, 40 patients are found to be malignant and 15 patients are found to be benign. Each scan consists of 80 to 200 slices of human chest scan images in different angles and sides. The dataset cases are of different ages, gender, employment, educational status, and area of residence (Mendeley Data 2021). Figure 2 depicts sample cases of cancer and non-cancer from the test dataset.

Identification of Deep Learning Technique

The proposed AI model uses CNN, which has successive layers of neural networks, each analysing a different part of the given data and collecting important information (features). CNN is mainly used for Image Classification, as it automatically extracts the features and trains the model accordingly. CNN works by applying a series of "convolutions" on the input image. A convolution is a filter or kernel applied on an image to extract its relevant information from each pixel. The relevant information consists of the internal variables, weights and biases. The filter is a matrix that moves on the image pixels to transform the input pixels into smaller parts. Figure 3 represents the working architecture of CNN.

Development of Novel AI Framework for Lung Cancer Classification

The proposed Deep Learning based AI framework has a single layer of Convolutional 2D Layer with MaxPooling and "ReLu or Rectified Linear Unit" activation function. The conversion of

Fig. 2. Sample images of cancer and non-cancer cases from IQOTH/NCCD dataset

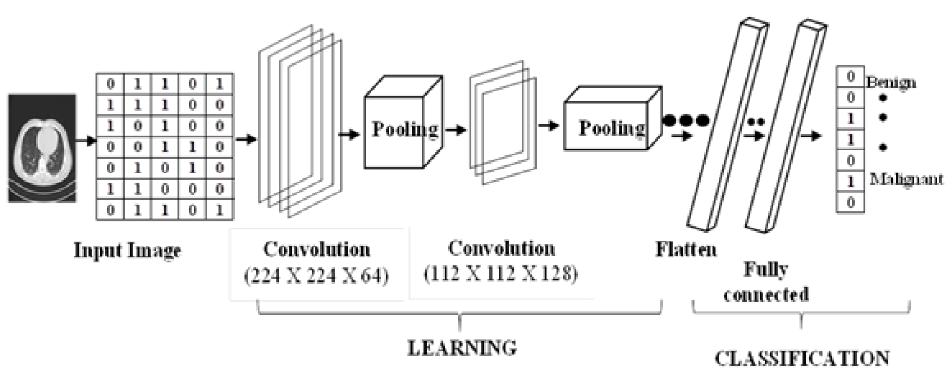

Fig. 3. CNN architecture

various layers into a one-dimensional array is performed by the Flatten layer, which is then combined by the fully connected dense layer that connects bias and kernel in all the neurons. Another activation function "Softmax" is applied with the dense layer. Softmax activation function provides probabilities for each possible output class. Table 1 gives a detailed description of the layers used in the proposed Deep Learning Model.

Conv2D Layer

The 2D convolutional layer applies a kernel or filter on the 2D image inputs to produce the weights and bias to be passed on to the next layer.

MaxPooling Layer

MaxPooling is a method by which many pixel values are converted into a single value by taking the maximum value among them. MaxPooling is applied after each convolutional layer in order to lessen the input image size. Two main components of MaxPooling are Pool Size and Stride.

Activation Function

An activation function is any mathematical function used to bring the nonlinearity into the output of each neuron. It calculates the weights and biases of each layer by activating the neurons. The activation function used here is ReLu or Rectified Linear Unit and Softmax.

Flatten

Flatten is a method of reducing the input to a 1-dimensional array for the next layer in the model. In convolutional neural networks, the Flatten is applied after the final convolutional layer to convert the layers into a single long 1-dimensional array.

Dense Layer

Dense layer is also fully connected and it combines the input, kernel, and bias to provide an output.

Loss Function

Loss function is a method for calculating the errors in a neural network. There are many mathematical loss functions available to calculate loss according to the given input. The Loss function used in this work for the model compilation is **"Sparse Categorical Cross Entropy"**, which is mainly used for multi-class classification, for comparing the predicted labels with true labels and calculation of the associated loss.

Optimiser

Optimiser is a method used to change the variables of a CNN such as weights and learning rate to reduce the loss. It is a kind of implementation of the Gradient Descent algorithm. The optimiser function used in this work is "ADAM or Adaptive with Momentum" or "Adaptive Moment Estimation", which considers an exponentially weighted average to adapt the learning rate of each weight in the neural network.

RESULTS

The proposed framework is tested with IQ-OTH/NCCD Mendeley dataset, which consists of three different types of lung images such as Benign, Malignant and Normal. The results obtained by the framework shows 100 percent ac-

Table 1: Deep Learning Model Layers applied

No. of layers	*Layer*	*Layer function*
Layer 1	Conv2D	Applies convolutions on 2D image input data using kernels.
Layer 1	ReLu	ReLu provides the model with ability to solve non-linear problems.
Layer 1	MaxPooling 2D	2D MaxPooling reduces the size of the input image.
Layer 2	Flatten	Flatten is a method of reducing the input to 1D array for the next layer.
Layer 3	Dense(64)	Dense Layer combines the input and kernel and bias.
Layer 4	Dense(3)	Dense Layer combines the input and kernel and bias.
Layer 4	Softmax	Softmax in the last layer provides the probabilities for each class.

curacy in training as well as validation with the above mentioned dataset on a ratio of 60:40 with training and validation division. plots the training accuracy and the validation accuracy achieved by the proposed AI framework are plotted in Figure 4. The plots of the training loss and the validation loss are shown in the Figure 5.

Table 2 presents the accuracy metrics of the developed Deep Learning Model.

Table 3 presents the Classification Report for each class in the dataset.

Figure 6 depicts a descriptive heat map of the confusion matrix of the model.

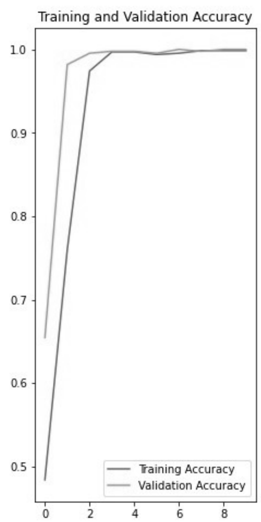

Fig. 4. Training and validation accuracy achieved by the proposed AI framework

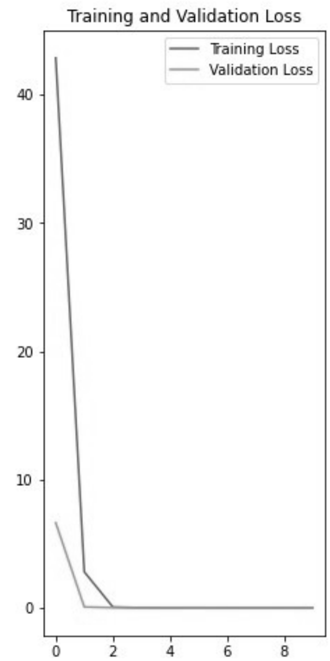

Fig. 5. Training and validation loss in the proposed AI framework

DISCUSSION

The Deep Learning model developed and proposed in this work has attained 100 percent of accuracy on a single layer of 2D Convolutional Layer on non-augmented data. The dataset was divided into 60:40 and 10 Epochs

Table 2: Accuracy metrics

Accuracy metrics	*Benign cases*	*Malignant cases*	*Normal cases*
False Positive	0	0	0
False Negative	0	0	0
True Positive	48	225	167
True Negative	392	215	273
Sensitivity/ True Positive Rate	1	1	1
Specificity/ True Negative Rate	1	1	1
Precision/ Positive Predictive Value	1	1	1
Negative Predictive Value	1	1	1
False Positive Rate	0	0	0
False Negative Rate	0	0	0
False Discovery Rate	0	0	0
Overall Accuracy for each class	1	1	1

Table 3: Classification Report of the proposed Deep Learning Model

	Classification report			
Case name	*Precision*	*Recall*	*F1- score*	*Support*
Benign Cases: 0	1.00	1.00	1.00	48
Malignant Cases: 1	1.00	1.00	1.00	225
Normal cases: 2	1.00	1.00	1.00	167
Accuracy	-	-	1.00	440
Macro avg	1.00	1.00	1.00	440
Weighted avg	1.00	1.00	1.00	440

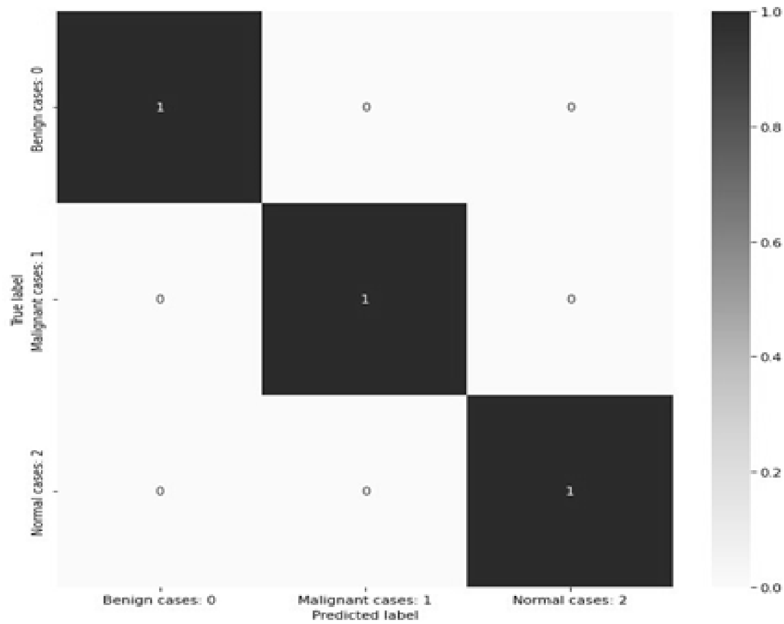

Fig. 6. Descriptive heat map of Predicted Label and True Label of the proposed Deep Learning Model

are used in training and validation. The images were in their original, grayscale format, rescaled and normalised all the images to the same range [0,1]. It is observed that the validation loss is low when compared to the training loss. To measure and verify the correctness of the proposed model, various accuracy metrics are used. These accuracy metrics help to identify where exactly the model went wrong with the classification. A Confusion Matrix is a method to visualise errors in the classification done by the model. It displays the correctly and incorrectly predicted images in the validation dataset.

Validation and Assessment of the Proposed AI Framework

Accuracy Metrics

True Positive (TP): This measure specifies the correct predictions of Positive classes.

True Negative (TN): This measure specifies the correct predictions of Negative classes.

False Positive (FP): This measure specifies the incorrect predictions of Positive classes.

False Negative (FN): This measure specifies the incorrect predictions of Negative classes.

Sensitivity

Sensitivity is a measure that says the ratio of the rightly classified positive classes.

$$Sensitivity = \frac{TP}{TP = FN}$$

Specificity

Specificity is a measure that says the ratio of the rightly classified negative classes.

$$Specificity = \frac{TN}{TN = FP}$$

Precision

Precision is a measure that says the proportion of rightly classified positive samples to the total samples classified as positive, irrespective of the prediction correctness.

$$Recall = \frac{TP}{TP + FP}$$

Recall

Recall is a measure that says the proportion of rightly classified positive samples in the positive class to the total positive samples. High Recall value indicates that more positive samples have been identified.

$$Precision = \frac{TP}{TP + FN}$$

Table 4 draws a comparison between various models applied on the same dataset and the proposed Deep Learning Model.

Lung cancer image classification of Mendeley dataset was done by some of the earlier research works by using CNN with AlexNet Architecture, CNN, ANN, SVM with Polynomial Kernel, Manual CNN, AlexNet, and GoogleNet with DNN. A comparative study was done between the performance of these existing research techniques and the proposed Deep Learning model. It is clear from the comparison that the proposed Deep Learning model outperforms the other models especially for lung CT images.

CONCLUSION

This work is aimed at the development of an AI framework for Lung Cancer Classification, which can aptly identify the three types of cases, that is, Benign, Malignant and Normal. A review of literature is done to determine the latest research developments in lung cancer detection through AI. A detailed study on various research methods and techniques to build a strong AI framework is carried out. A comparative study is done with other research models to demonstrate the capability of the proposed framework. It has been found that the proposed Deep Learning model has outpaced the existing research methods by achieving 100 percent accuracy in training and validation.

RECOMMENDATIONS

This study identified the proposed Deep Learning model with CNN technique as a good contender for classification of Lung Cancer CT images. A number of datasets were considered for testing the framework, of which the IQ-OTH/

Table 4: Comparison between various models applied on the same dataset

Research technique	*Dataset*	*Sensitivity*	*Specificity*	*Precision*	*Accuracy*
CNN with AlexNet Architecture (Al-Yasriy et al. 2020)	IQ-OTH/NCCD	95.714%	95%	97.1015%	95%
CNN ANN (Prarthana KR et al. 2021)	IQ-OTH/NCCD	-	-	-	98.41% 70.92%
SVM with Polynomial Kernel (Kareem et al. 2021)	IQ-OTH/NCCD	97.143%	97.5%	98.551%	89.8876%
Manual CNN AlexNet (Gupta et al. 2021)	IQ-OTH/NCCD	-	-	-	96.88% 51.56%
GoogLeNet with DNN (AL-Huseiny et al. 2021)	IQ-OTH/NCCD	95.08%	93.7%	-	94.38%
Proposed Deep Learning Model	IQ-OTH/NCCD	100%	100%	100%	100%

NCCD-Mendeley dataset with CT scan images of the lungs is selected. The proposed AI framework will be well suitable for classifying a wide variety of CT images as it is proven to produce the best results in this research work. The developed AI framework can also be further tested with other types of image datasets such as PET, MRI, functional MRI, etc., for further improvement and wider applicability.

ACKNOWLEDGEMENT

The authors would like to thank and acknowledge the project funding provided by MHRD – RUSA 2.0 – BCTRC (Bharathiar Cancer Theranostic Research Centre), Bharathiar University, Coimbatore, India.

REFERENCES

Amudha T, Subhiksha R Mallya, Satheesh Kumar J 2022. Artificial intelligence tools and techniques in cancer data analysis. *NeuroQuantology*, 20: 273-286.

Anubha Gupta, Kathikeya Manda VL, Ida Seraphim B 2021. Lung cancer detection using image processing and CNN. *International Journal of Advance Research, Ideas and Innovations in Technology*, 7: 949- 957.

Ciello DA, Franchi P, Contegiacomo A, Cicchetti G, Bonomo L, Larici AR 2017. Missed lung cancer: When, where, and why? *Diagn Interv Radiol*, 23: 118-126.

Faisal MI, Bashir S, Khan ZS, Hassan Khan F 2018. An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer. *3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*, Karachi, Pakistan, 2018, 1-4. doi: 10.1109/ICEEST.2018.8643311

Hamdalla F Al-Yasriy, Muayed S AL-Husieny, Furat Y Mohsen, Enam A Khalil, Zainab S Hassan 2020. Diagnosis of Lung Cancer Based on CT Scans Using CNN. IOP Conference Series Materials Science and Engineering, 2nd In*ternational Scientific Conference of Al-Ayen University*, Thi-Qar, Iraq, 15-16 July, 2020, 928: 022035, pp.1-10. doi:10.1088/1757-899X/928/2/022035

Hamdalla Fadil Kareem, Muayed S AL-Husieny, Furat Y Mohsen, Enam A Khalil, Zainab S Hassan 2021. Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, 21: 1731-1738.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, Freddie Bray 2021. Global Cancer Statistics 2020. GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *ACS Journals*, 71: 209-249.

Jijitha S, Amudha T 2020. Breast cancer prognosis using machine learning and genetic algorithm. Experiment on six different datasets. *Evolutionary Computing and Mobile Sustainable Networks*, 53: 703-711.

Kiruba Nesamalar E, Satheeshkumar J, Amudha T 2021a. Influence of an efficient hierarchical clustering algorithm in analyzing cancer affected DNA dataset. *Journal of Physics: Conference Series, 3rd International Conference on Smart and Intelligent Learning for Information Optimization (CONSILIO 2021)*, Hyderabad, India, July 9-10, 2021, 1998 012030: 1-12.

Kiruba Nesamalar E, Satheeshkumar J, Amudha T 2021b. Influence of artificial intelligence in clinical and genomic diagnostics. *Artificial Intelligence Theory, Models and Applications.* Florida, United States of America: CRC Press, Auerbach Publications, 323-335.

Mendeley Data Page of The IQ-OTHNCCD Lung Cancer Dataset 2021. From <https://data.mendeley.com/datasets/bhmdr45bh2/2> (Retrieved on 15 November 2021).

Muayed S AL-Huseiny, Ahmed S Sajit 2021. Transfer Learning with GoogLeNet for detection of Lung Cancer. *Indonesian Journal of Electrical Engineering and Computer Science*, 22: 1078-1086.

Prarthana KR, Bhavani K 2021. Lung cancer classification techniques. *International Journal of Engineering Science and Computing (IJESC)*, 11: 28054-28058.

Paper received for publication in January, 2023
Paper accepted for publication in May, 2023