Kathford Journal of Engineering and Management,
ISSN: 2661-6106

https://kathford.edu.np/Journal/
VOL. 4 ISSUE 1 DECEMBER 2024, pp. 84–91

Efficacies of the CNN Algorithm in Predicting Lung Cancer

Prabin Acharya[1], Rashmi Tandukar[1], Suman Karki[1], Tripti Poudel[1], Jalauddin Mansur[1,*]

[1]*Department of Computer and Electronics, Communication & Information Engineering, Kathford International College of Engineering and Management (Affiliated to Tribhuvan University), Balkumari, Lalitpur, Nepal*

*Corresponding author: *er.jalauddin@kathford.edu.np*

ABSTRACT— Lung cancer is the leading cause of cancer-related death in the 21st century, and is highly expected to remain so in the future ages. It is possible to diagnose and treat the lung cancer if the proper symptoms of the diseases are detected in the early phases. A convolutional neural network (CNN)-based machine learning model is a recently emerged network design that has been functionalizing since the last decade mainly to optimize the detection processes of the lung cancer in the pre-collected medical examinations and the closely associated lung cancer datasets. Through the CNN classifier, the lung cancer patients are majorly classified based on their symptoms while the Python programming script drives the detection schemes as a whole through the effective implementation of the entire model. In fact, the CNN model enables the medical practitioners and hospital professionals to build the sustainable prototype mechanisms for the effective treatment of the lung cancer via the computational intelligence without any negative impacts on the societal environment. As it can reduce the amount of wasted resources and the amount of work required to complete the manual diagnosis tasks, the lung cancer patients can get the real-time treatment in a cost-effective manner with the minimal effort and latency from any location at any time. In this article, we overviews the performances of the CNN model designed by us, and assess its accuracy regimes in detecting the cancerous and noncancerous Lung cells. We believe that our model could mark the future prototype medical diagnosis scheme for the efficient lung cancer cells investigations, their progressive growing stages, and the timely assessments of the medical treatments.

KEYWORDS—***CNN, Machine Learning, Python Coding, Real Time Medical Diagnosis***

1. INTRODUCTION

1.1 BACKGROUND

Lung cancer is one of the leading causes of death and health problems in many countries with a 5-year survival rate of only 10─16%. The severe lung cancer would actually damage the cells of the respiratory system by hampering the entire tissues, which can lead to the various potential problems with breathing and difficulty in getting the fresh air, and these cases can even lead to the death of a person. Over the past few years, early detection of lung cancer has become a new area of research in the medical system. Accurate detection of the size and location of lung cancer plays a vital role in the diagnosis of lung cancer. The diagnostic method consists of the four major stages: pre-processing, characterization, extraction; and classification; the features are extracted based on the diverse computational algorithms. So far, various machine learning algorithms are used to detect the lung cancer and cancerous cells (Mandal & Banerjee, 2015, Taher & Sammouda, 2011). In this study, the most popular asset namely the convolutional neural network (CNN)-based machine learning algorithms are employed to detect the lung cancer cells and cancerous

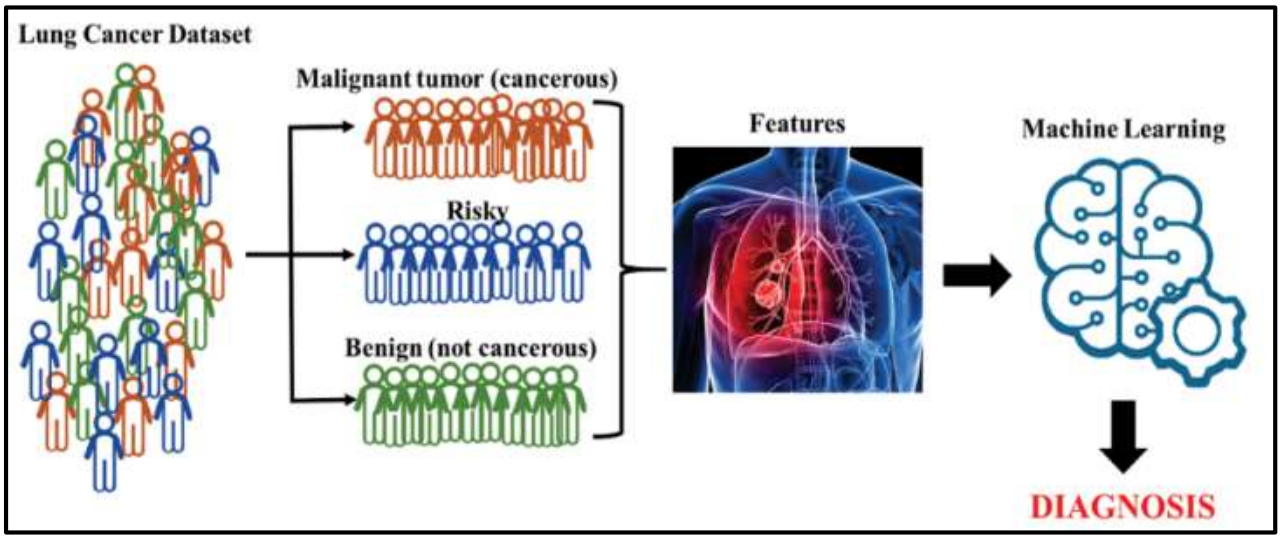

Figure 1. Outline of the research statement highlighting the overall steps of the CNN implementations strategies.

tissues. The detections of the computed tomography (CT) scans images using the machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Networks (BN) are few of the strategies to resolve the issues associated with the early phases detections (Sandhiya & Kalpana, 2019, Zhao, X. 2018). Some of the works are also reported by considering the proper usage of the histopathology images but are seriously used to distinguish between the carcinomas and non-carcinoma images with lower precisions accuracy. This research paper considers the use of the CNN architecture classify benign, adenocarcinoma and squamous cell carcinomas. To the knowledge of the present authors, the research works using the CNN model to classify only the given three various histopathological images and data model accuracy are very limited. Since the disorder with tumors in the lungs has the largest number of transients, the effective development of the node is the main cause for the prevalence of the disease. The four-year survival rate for this deadly lung cancer is almost 12%. The characteristics of the abnormal cell growth vary greatly at each typical stages, and depend on how dangerous the tissue is. It affects people 65-70% more than that caused by any other predominant cancers like breast or skin. The regular checkups and the constant medical monitoring are frequently required at every particular stages. If the early detections of the cancerous cells and tissues are made, the Cancer survival rates are already speculated as 45%-75% (Diaz & Solano, 2014). In the views of this, present work underscoring the effectiveness of the CNN model to detect the lung cancer and cancerous cells stands as a doctrine document expressing the facile means of diagnosing such a deadly cell and tissue types.

1.2. PROBLEM STATEMENT

The Lung cancer disease actually occurs when the cells in the lungs begin to grow rapidly out of control. This expedite cell growth takes place any parts of the lungs, and does affect any part of the respiratory tracks. In Nepal, the proper lung cancer detection stages lie in the range of 15 to 16 days after the absolute issuing of the samples in the cancer treatment hospitals, Bangalore, India. So, it not only makes the entire diagnosis and treatment processes prolonged but also let the cancerous segments spreading to the other parts of the body. The early detection of the cancerous cells and tissues is very indispensable to get the timely treatment of the lung cancer and its proper posttreatment cares. The machine learning algorithms such as those facilitated

by the CNN are one of the means to minimize the issues associated therewith. The present authors are very much triggered by the same statement of the problems because of which the original concept of the present work was initiated. With this, the present research work was mainly designed to predict weather the hospital patients bear the lung cancer cells and cancerous tissues or not in their live body. To accomplish the project successively, the CT image datasets (Paul, R, 2016) are practically assessed by examining the working performance and workability of the CNN driven algorithms and the underlying programmatic coding as summarized in Figure 1.

1.3. LITERATURE REVIEWS

Muthazhagan B, Ravi T, *et al.* (2021) states that with the help of current lung cancer prediction technologies, predicting and detecting lung cancer at an early stages is a difficult and challenging task. Even though the early prediction of the lung cancer could extend the personal life by one to five years, the images diagnosed through the traditional means were only classified as "abnormal" or "normal", and none of the cancerous stages [Stage 0 ─ Stage IV] were practically considered. In this views, present work does address to some extents. The research work reported by Masud M, Sikder N, *et al.* (2021) uses a CNN-based model to classify the diagnosis images into one of five types: colon adenocarcinomas, benign colon tissues, lung adenocarcinomas, squamous cell lung carcinomas, and benign lung tissues. While a maximum accuracy of 96.33% was achieved in classification, the authors reported that two of the five classes can perform much better with further experimentation. The dataset used therewith was linked to the histopathology, a microscopic examination invasive process of the biopsy. Median and Gaussian filter were applied in the preprocessing stages and the required image-datasets were then segmented using the Watershed algorithm. They went on to the use of support vector machines to identify the diagnosed cancer nodes as benign or malignant. This upgraded model outperformed the previously designed model approaches by 5.4% with 92% precision accuracy. The only flaw of the model was its inabilities to differentiate the cancer cells residing into the cancer stages of I to IV.

T. Atsushi, T. Tetsuya, *et al.* applied the Deep Convolutional Neural Network (DCNN) to cytology images to automate the lung cancer type classification. In their dataset, they considered the images of small cell carcinoma, squamous cell carcinoma, and adenocarcinoma. The DCNN architectures with 3 convolution and pooling layers and 2 fully connected layers with a dropout of 0.5 were used. Thus developed model was able to achieve an overall accuracy of 71.1%, which is quite low though. Similarly, S. Sasikala, M. Bharathi, and B. R. Sowmiya proposed the detection scheme using the CNN on CT scan images to detect and classify the lung cancer. They exclusively used the MATLAB for their work and had the two phases in training while extracting the valuable volumetric features from the essential input data as the first phase and classification as the second phase. The system proposed by them was able enough to classify the cancerous and non-cancerous cells with the 83% accuracy. In the present work, we vow to assess the lung cancer predictions using the CNN algorithm associated image datasets.

2. COMPUTATIONAL AND THEORETICAL DETAILS

2.1 Overviews of the CNN

A convolutional neural network (CNN) (outlined in Figure 2) is a type of deep learning algorithm designed to process the structured grid data such as images or video (Jain, 2020).The CNNs are particularly powerful in computer vision tasks where the

goal is to recognize the patterns, objects, or features in the visual datasets. The CNN architecture is actually inspired by the visual processing that occurs in the human brain (Westaway *et al*. (2013), American Lung Association) Here are the lists of the key components and concepts of a typical CNN:

Convolutional Layers

The basic building blocks of CNNs are convolutional layers. These layers apply a convolution operation to the input data. Convolution involves passing a filter (also known as a kernel) over the input data to extract the local patterns or features. The result is a set of the feature maps that highlight the important features in the input.

Pooling Layers

Pooling layers follow the convolutional layers and reduce the spatial dimensions of the input data. Maximum pooling is a common type of pooling where the maximum value within a certain window is kept and the rest are discarded. This helps to reduce the computational complexity, and makes the entire network more robust towards changing inputs.

Activation Function

The nonlinear activation functions such as Rectified Linear Unit (ReLU) are applied after the convolution and pooling operations to introduce the nonlinearity into the model. This allows the network to learn complex patterns and relationships in the data.

Fully Connected Layers

After the several convolution and pooling layers, one or more fully connected layers are usually added. These layers connect each neuron to each neuron in the previous and subsequent layers, allowing the network to make predictions based on high-level features learned previously in the network.

Flattening

Before the fully connected layers, the output from the previous layers is combined into a one-dimensional vector. This step is necessary to connect the convolution and pooling layers to the fully connected layers.

Outage

Dropout is a regularization technique commonly used in the CNNs to avoid overfitting. During training, random neurons are "dropped out" by setting their weights to zero. This helps the model to better generalize to the unseen data.

Loss function and optimization

The choice of loss function depends on the specific task (classification, regression, etc.). The common options include cross-entropy loss for classification problems. An

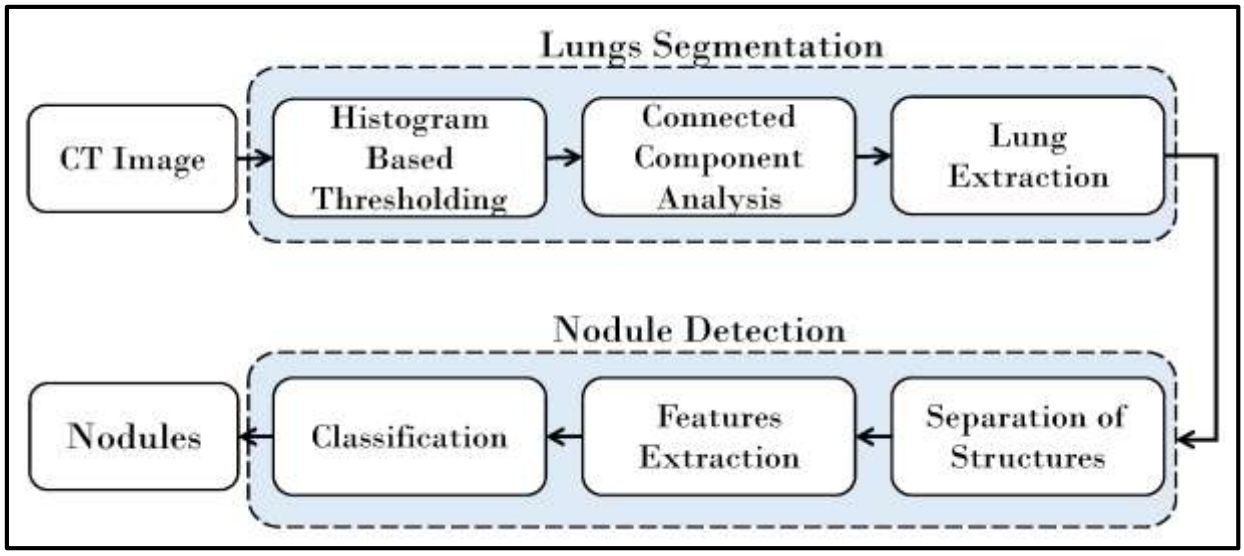

Figure 2. General block diagram for Machine Learning

optimization algorithm (e.g. stochastic gradient descent) is used to minimize the loss function during the training.

Training

CNNs are trained using the labeled datasets. The process involves feeding the input data over the network, calculating the loss, and adjusting the weights using backpropagation. This process is repeated iteratively until the model performs well on the training data. The CNNs have demonstrated the state-of-the-art performances in a variety of computer vision tasks, including image classification, object detection, and image segmentation. Their ability to automatically learn hierarchical representations of the features makes them effective for capturing the complex patterns in the visual data.

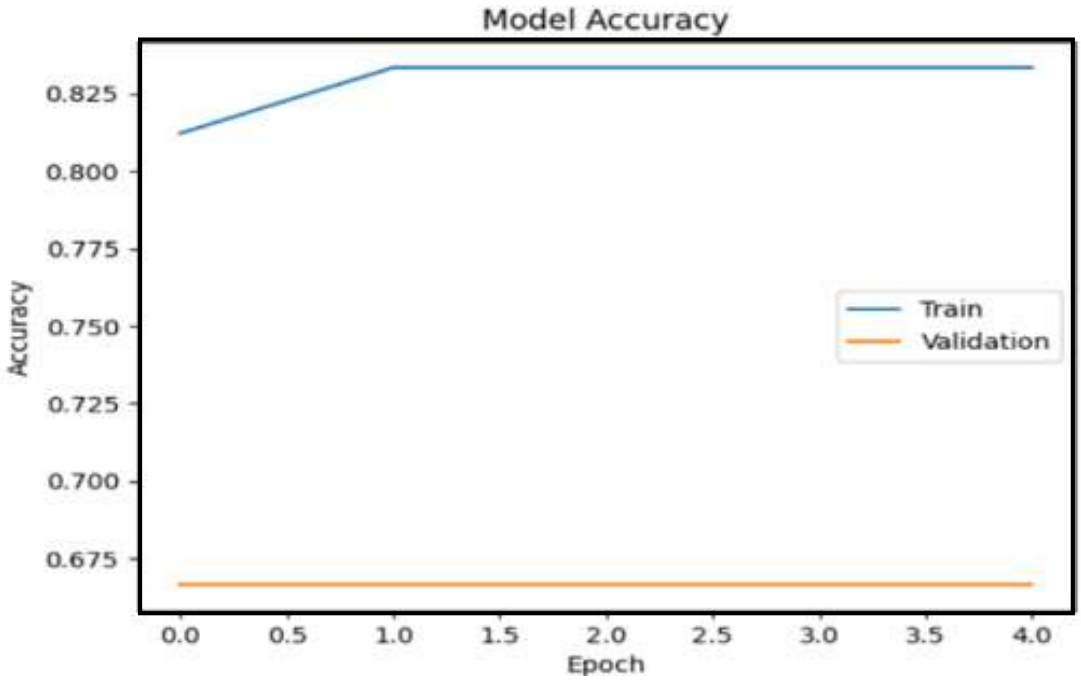

Figure 3. Model accuracy (Train vs Validation of the model)

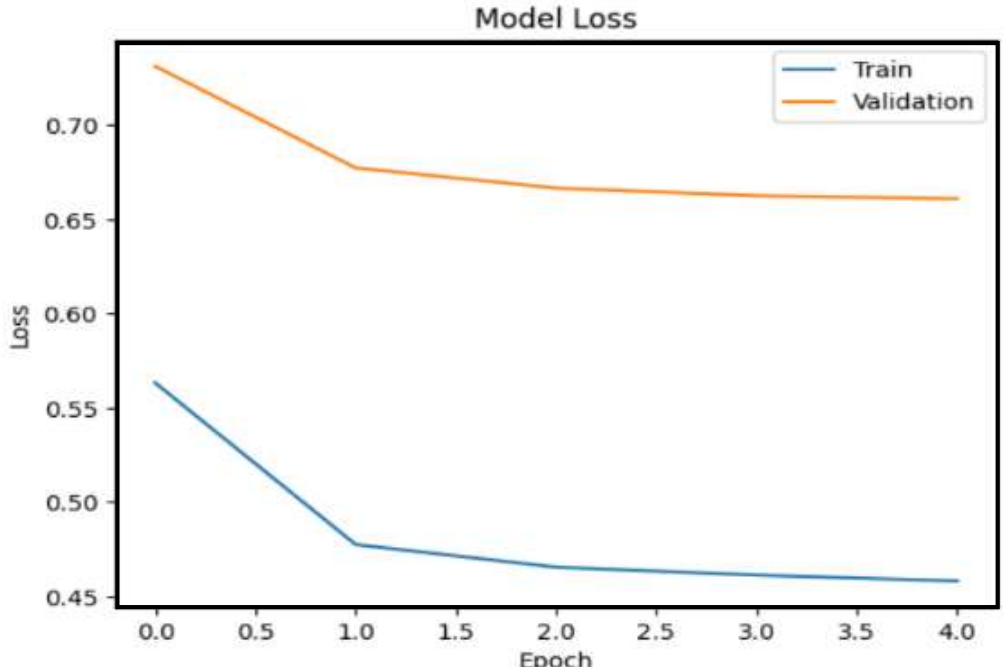

Figure 4. Model loss (Train vs Validation of the model)

Figure 5. The present CNN model presumes this cell as a cancerous

2.2 Platform Selection

PYTHON

The Python's standard library is very extensive. The library contains built-in modules (written in C) that provide direct access to system functionality such as file I/O that would otherwise be inaccessible to Python programmers, as well as modules written in Python that provide standardized solutions for many problems that occur in everyday programming. The Python installers for the Windows platform usually include the entire standard library and often also include many additional components. For Unix-like operating systems, Python is normally provided as a collection of packages, so it may be necessary to use the packaging tools provided with the operating system to obtain some or all of the optional components. Following are the components of python:

1. Pandas: Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

2. NumPy: The name "NumPy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like Tensor Flow use NumPy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

3. Scikit-Learn: It is a famous Python library to work with the complex data. Scikit-learn is an open-source library that supports machine learning. It supports variously supervised and unsupervised algorithms like linear regression, classification, clustering, etc. This library works in association with the NumPy and SciPy.

4. Matplotlib: This library is very responsible for plotting the numerical data. And that's why it is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs, etc.

3. RESULTS AND DISCUSSIONS

The complete prediction of lung cancer using CNN algorithm taking image dataset is outlined in Figure 1. User can submit their dataset in the model as shown schematically, and achieves the immediate results. From the beginning, we have selected the topic for our project as lung cancer prediction using the CNN algorithm, our main intension was to predict whether the patient have lung cancer cells or not. We at first searched out on lung cancer, its symptoms, effects and its mitigating requirements. And then the feasibility of the model was verified. We prepared multiple diagrams like, class, use case, sequence, ER etc. After that, we managed to do code using python in Google colab. We collected the required image datasets from kaggle (Hamdalla F., 2020. We trained the CNN model developed by us using the multiple tools and functions. Then we worked on the assessment of the accuracy of our model. The algorithm developed and tested in the present work achieved an accuracy of 79.56% as shown in Figure 3 (Figure 4 displays the model loss percentage during train and validation). The relatively lower performances of the presently designed CNN model in the validation and test sets compared with the training dataset are most probably due to the sample size of 1250 and 134 respectively. Then we linked our dataset in the model and we could verify weather our model gives the correct output or not. As shown in Figure 5 and 6, the CNN model proposed in this study is able to detect the both cancerous and noncancerous cells almost in good precisions.

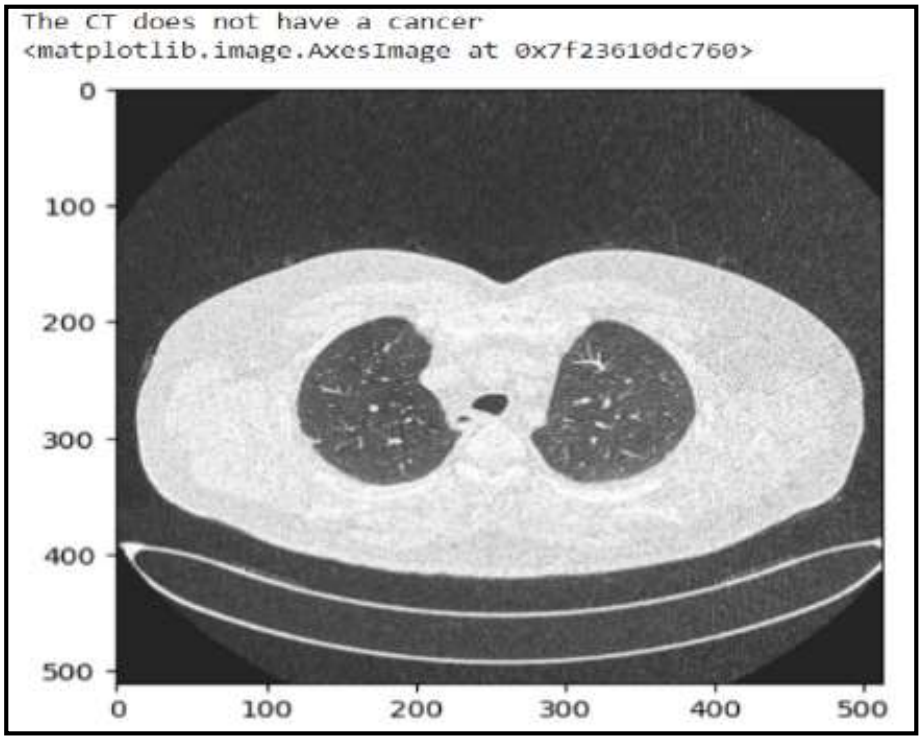

Figure 6. The present CNN model detected this cell as a noncancerous

4. CONCLUSION

This work was primarily based on evaluating the CNN detection ranges of the lung cancer cells and noncancerous cells. We at first designed the model, and validated it based on its detection abilities to the pool of the cancerous and noncancerous cell CT images. We found the accuracy range of around 80% with the continuous decrement in the model loss during the both training and validation of the model. The model architectures, and the

computational programmatic codes designed in this study would definitely become the prototype model to advance the functionality of the other predominant CNN models and the ML algorithms. As the worldwide demands of the efficient real-time medical diagnosis of the deadly cancerous cells is increasing day to day, the possible schematic layouts and the model designs presented and implemented throughout this study may track the traditional medical diagnosis processes to the human friendly yet progressive and innovative directions.

5. ACKNOWLEDGEMENT

We would like to thank Department of Computer and Electronics, Kathford International College of Engineering and Management, for giving us a useful feedback in the course of this work. We are thankful to our seniors, our dearest friends for giving proper suggestions and advices. At last, we acknowledge Prof. Dr. Anant Babu Marahatta, a Professor of Chemistry at Kathford International college of Engineering and Management, for his thorough revisions and rewritings of this manuscript.

REFERENCES

1. Mandal, S., & Banerjee, I. (2015). Cancer classification using neural network. *International Journal*, 172, 18–49.

2. Taher, F., & Sammouda, R. (2011). Lung cancer detection by using artificial neural network and fuzzy clustering methods. In *2011 IEEE GCC conference and exhibition (GCC)*, IEEE, 295–298.

3. Sandhiya, S., & Kalpana, Y. (2019). An artificial neural networks (ANN) based lung nodule identification and verification module. *Medico-Legal Update*.

4. Westaway, D. D., Toon, C. W., Farzin, M., *et al*.(2013). The International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society grading system has limited prognostic significance in advanced resected pulmonary adenocarcinoma. *Pathology*, 45(6):553-8.

5. American Lung Association (2025). Available: *www.lung.org*

6. Diaz, J. M. P., & Solano, R. C. (2014). Lung cancer classification using genetic algorithm to optimize prediction models. The 5th International Conference on Information, Intelligence, Systems and Applications (IEEE).

7. Hamdalla F. (2020). The IQ-OTH/NCCD lung cancer dataset. Available:https://www.kaggle.com/hamdallak/the-iqothnccd-lung-cancer-dataset

8. Jain, A (2020). Deep Learning for Computer Vision - Introduction to Convolution Neural Networks. *Analytics Vidhya.*

9. Zhao, X., Liu, L., Qi, S., Teng, Y., Li, J., & Qian, W. (2018). Agile convolutional neural network for pulmonary nodule classification using CT images. *International Journal of Computer Assisted Radiology and Surgery*, 13(4), 585–595.

10. Paul, R., Hawkins, S. H., Balagurunathan, Y., et al. (2016). Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography*, 2(4), 388–395.