International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 12 Dec. 2024, pp: 306-312 www.ijaem.net ISSN: 2395-5252

A reliable lightweight deep learning base approach for lung cancer classification

Rahajaniaina Andriamasinoro[1], Ratiarison Adolph Andriamanga[2]

[1]*Associate Professor, Department of Mathematics, Computer Science and Applications, University of Toamasina, Madagascar*

[2]*Professor Emeritus, Department of physics and Applications, University of Antananarivo, Madagascar*

Corresponding Author: Rahajaniaina Andriamasinoro

Date of Submission: 10-12-2024 Date of Acceptance: 20-12-2024

ABSTRACT: The world suffers the impacts of the climate change caused by forest destruction and air pollution. Lung cancer is among of impacts where the population suffers mainly in the development country. The aim of this paper is to demonstrate the performance, the reliability and the effectiveness of using a lightweight deep learning base model for classifying lung cancer CT scan images. Furthermore, in the development country the lack of medical infrastructures and the specialist within the hospitals result numerous misdiagnoses. Or, some previous works were more complex and heavies. These situations conduct us to perform a lightweight deep learning base model inspired from the SSDLiteX approach for classifying lung cancer to aid specialist. The train data was balanced to obtain higher performance. Our model achieves 97% of accuracy, validation accuracy and F1-score. The result shows that the proposed approach is reliable for classifying lung cancer and could run on the low resource environment.

KEYWORDS: lung cancer classification, lightweight deep learning, lung cancer imaging.

I. INTRODUCTION

In the last decade, the world suffers the impacts of the climate change caused by forest destruction and air pollution. In fact, several problems have been occurred such as skin diseases, famine, drought and some cancers like lung, skin, and breast cancer. In 2002, 1.35 million people throughout the world were diagnosed with lung cancer, and 1.18 million died of this one [19] despite the advance of research in this domain. The advanced evolution of artificial intelligence has more impacts in several domains such as medical domain. Deep learning and transfer learning become essential to perform object classification task in some domain mainly medical one. The current trend for researchers uses evaluation of some methods or ensemble learning. So, in their work, [3] used ensemble learning to classify CT scan images. To achieve their goal, they used three different models such as CNN and its variants, Inception-V3 and ResNet50. Their model suffers from high variance on unsee data. [13] proposed the three CNN models for classifying lung cancer using VGG16, ResNet50V2, and DenseNet201 approaches. It performed to classify lung cancer into five different classes. The ensemble model achieved 91% of validation accuracy. Moreover, some deep learning models in [16] including ResNet50, ResNet101, EfficientNetB1, EfficientNetB3, EfficientNetB5, and EfficientNetB7 have been evaluated using a dataset of 1,000 lung CT-Scan images, contained four categories Adenocarcinoma, Large Cell, Squamous, and Normal. Their study revealed that EfficientNetB3 model gave higher performance achieving 97.78% of an accuracy, 97.34% of precision, a recall of 98.33%, and an F1-Score of 97.78%. However, these previews models have been more complex and heavies. Thus, the use of these methods requires expensive materials. Face to this situation, we propose a deep learning base model inspired from the SSDLiteX approach for detecting and classifying early lung cancer. Some pre-processing and balancing technics were applied to the train data to ameliorate the images qualities and to reduce overfitting. Then, the model trains and classifies the data into six class of lung cancer.

The rest of this study is organized as follows. A brief related work on the different lung cancer classification models is described in section II. Section III presents the material and methods. Results and discussion about our proposed approach are discussed in section IV and section V conclude this paper.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 12 Dec. 2024, pp: 306-312 www.ijaem.net ISSN: 2395-5252

II. RELATED WORK

[15] developed a hybrid lightweight deep neural network based on binary classification networks such as vanilla 2D CNN, 2D SqueezeNet, and 2D MobileNet in order to recognize lung cancer using CT images with and without early-stage lung cancer. The aim of their work is to evaluate the performance of each lightweight model. So, they used the Lung Image Database Consortium image collection (LIDC-IDRI) [12]. The dataset was formed by CT scans labelled 0 for non-cancer and 1 for cancer. 6055 samples are used for training and 1501 samples for validation. Some pre-processing technics such as down sampling, segmentation, normalization, and zero centering were applied. The experiment result shows that SqueezeNet is qualified the best model due to its quick processing time and sufficient accuracy. The proposed model provided an accuracy of 85.21%.

[8] proposed a model using U-NET with three parameter logistic distribution. The dataset is formed by 1018 thoracic CT scan images collection from the Lung Image Database Consortium image collection (LIDC-IDRI) [11]. The goal of this study is to evaluate a variety of labelling strategies, both monolithic (single label input) and hybrid (single label input with complementary labelling). To assess the proposed model, they made two experiments, without and with pre-processing such as CLAHE, wiener filter, and ROI segmentation. Then, data augmentation was applied. After experiments, the sensitivity improves from 90% to 91% after pre-processing and the dice coefficient increases from 91% to 92%. The result shows also that the monolithic strategy gave the best output.

In [7], the authors suggested a model based on Convolutional Neural Network (CNN) and DenseNet approaches for lung cancer detection and classification. In their study, 2073 radiology images were used. These ones were sorted into three classes: normal, benign, and malignant instances. Some pre-processing tasks such as resize and rescaling were applied to achieve best result. Among of 2073 images, 74% were used for training, 4% for validation and 20% for test data. Their provided 99.49% of accuracy.

The authors in [17] designed three types of deep neural networks like Convolution Neural Network (CNN), Deep Neural Network (DNN) and Stacked AutoEncoder (SAE) for lung cancer classification. Their goal was to evaluate the performance of each architecture. Thus, CT scan images collection from the Lung Image Database Consortium image collection (LIDC-IDRI) were used. This dataset contains 244,527 images of the 1010 cases. 4581 images of lung nodules were constituted the train set. Among them, 2265 cases were benign pulmonary nodules and the other one was malignant pulmonary nodules with 2311 images. The Cross-validation approach was used to assess the three architectures. 10% of train set were used to perform this validation task. Some pre-processing tasks and data augmentation were applied in order to obtain the higher performance and to avoid overfitting problem. The experimental results show that the CNN architecture achieved the best performance than the two other architectures. It gave an accuracy of 84.15%, sensitivity of 83.96%, and specificity of 84.32%.

In the paper, [5] introduced an augmented CNN for classifying lung cancer images. Their system was divided in three steps: data acquisition, data augmentation and classification. They used LIDC-IDRI dataset composed of malignant and Benign cases for training their model. The dataset contains 2066 images, of which 80% (1653) of them were used for train set and 20% (413) for test set. 20% (330) of the training data were used for Validation data. Many pre-processing like mean subtraction, normalization, and local contrast normalization were applied to obtain best performance. Then, data augmentation was performed. They carried out their model using Python libraries such as Keras and TensorFlow on a GPU NVIDIAMX450, an i5-1135G7 processor, with Windows as the operating system. The result shows an accuracy of 95%.

[1] used deep learning technique to classify lung cancer images. In their study, they utilized DenseNet-121 model for feature extraction and MobilNetV3 Small model as classifier model. Thus, they used the PET/CT (Lung-PET-CT-Dx) dataset formed by 31,562 annotated images publicly available in [8][9]. The CT image resolution was 512 × 512 pixels at 1mm × 1 mm, and the PET image resolution was 200 × 200 pixels at 4.07 mm × 4.07mm. Several pre-processing was utilized like images registration, symmetric normalization, resample image filter, … Then data augmentation was used to reduce overfitting. So, the authors applied more geometric transformations and multi-channel GAN model for achieving this one. Once the features were extracted, deep autoencoders was applied to minimize the feature dimensionality. To improve the accuracy quality, they performed quantization-aware training, early stopping strategies and the Adam optimization approach was utilized to fine-tune the hyper-parameters to reduce the training time. They carried out their experiment in Intel i7, 16 GB RAM, and GeForce GTX 950 GPU. The implementation was achieved with Python 3.8.3 using the Keras library. Their study led

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 12 Dec. 2024, pp: 306-312 www.ijaem.net ISSN: 2395-5252

to the result of an accuracy of 98.6 % and a Cohen's Kappa value of 95.8 % with 2.1 million parameters.

[10] designed the use of the "Decision Tree" algorithm and the Customised "VGG16". For this work, Decision Tree was used to evaluate the severity of Lung Cancer according to the answer of the patient. Then, VGG16 was utilized to examine the CT scan images of the patient and validate the result. For this study, they used two datasets from Kaggle. The first dataset consists of 1000 unique entries about the conditions of numerous patients and their diagnosed risk of cancer. It was used by the decision tree model to provide the severity of the carcinogen found in the patient's body. The second one was formed by CT scan images available from [14]. It was utilized by VGG16 model to predict the three types of cancer found in the lungs such as Adenocarcinoma, Large Cell Carcinoma, Squamous Cell Carcinoma or the Normal Lung image. Before the training phase, some pre-processing was performed: resize, random rotation, skew, random zoom. Then, data augmentation was applied. The decision tree approach gave an accuracy of 99.67% for the prognosis of the severity of Lung Cancer and VGG16 achieved an accuracy of 90.79% in predict task.

An overview of lung cancer classification method was performed by [18]. They compared the performance of seven machine learning model such as CNN, Support Vector Machine (SVM), Artificial neural network (ANN), Multi-Layer Perceptron (MLP), K-Nearest Neighbour (KNN), Entropy degradation method (EDM) and Random Forest (RF). They utilized an accuracy, sensitivity and specificity as evaluation metrics. For achieving their goal, they used small and large dataset. After experiment, for small dataset, CNN and SVM gave the best result rather than other models. CNN architecture able to classify the benign and malignant cases with an accuracy of 96% with 30 images for test data and 96% also for SVM with 32 images for test one. For large dataset, the KNN and RF achieved an accuracy of 98.31%, 98.30% respectively.

A Deep neural network for classifying lung cancer as normal or malignant using CT images was developed in [19]. Thus, they used a densely connected convolution neural network (DenseNet) and adaptive boosting algorithm. Moreover, a dataset containing 201 lung images was used to evaluate their approach. Among of 201 images, 85% are used for training and 15% for testing and classification. Resize blur removal and histogram equalization have been performed before training phase. The result of training and testing were utilized by ADABOOST for classification. They obtained an overall accuracy of 90.85%.

III. MATERIALS AND METHOD

3.1. Our approach

In this section we describe our approach to achieve our goal. According to the situation with previews works, we propose a lightweight model. We combined MobilNetV3Small and auxiliary stage of SSDLiteX with some changes in the number of the filter. For validating our model, we merged two publicly available dataset from Kaggle web site. Our dataset contains 3174 images, 70% were used for training and the rest of this one was divided equally in to validation and test set. So, we balanced the data in train set for having better results. Then, this data is passed through to the model to be classified into one of five types of lung cancer or normal case. The Figure 1 shows our approach.

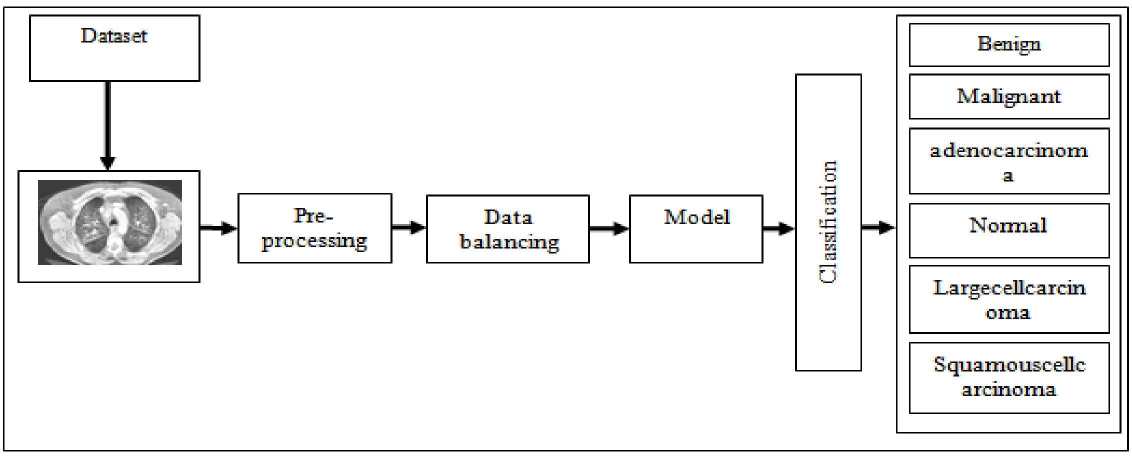
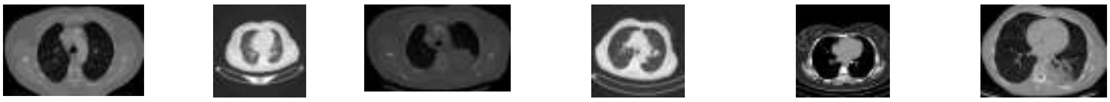


Figure 1- Proposed approach

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 12 Dec. 2024, pp: 306-312 www.ijaem.net ISSN: 2395-5252

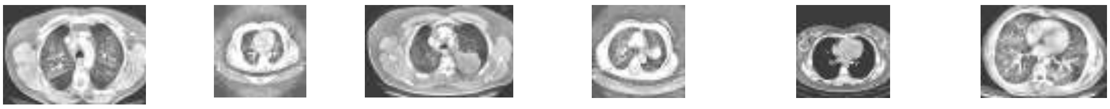
3.2. Dataset description

The dataset consists of 3174 images constitutedby six classes in which 220 are benign, 1122 are malignant, 338 are adenocarcinoma, 187 are large cell carcinoma, 260 are squamous cell carcinoma and 1047 are normal case. This dataset was obtained after merging two datasets downloaded from Kaggle website [20][21]. The first dataset consists of 1292 CT scan images divided into three classes such as benign, malignant and normal case. The second one was constituted by 1882 CT scan images formed by six classes such as benign, malignant, adenocarcinoma, large cell carcinoma, squamous cell carcinoma and normal case. Thus, 70% of 3174 images were used for train set and the rest was divided equally into validation and test set. Some pre-processing such as sharpening, clahe, brightness adjustment and median filter were applied to ameliorate images qualities. Then, the dataset was balanced in order to have a best result. The Figure 2 illustratessome images from the dataset before pre-processing and the Figure 3 shows some ones after pre-processing.



adenocarcinoma Benign Large cell carcinoma Malignant Normal Squamous cell carcinoma

Figure 2- Sample images from the dataset before pre-processing



adenocarcinoma Benign Large cell carcinoma Malignant Normal Squamous cell carcinoma

Figure 3- Sample images from the dataset after pre-processing

3.3. Proposed method

The proposed model was formed by two components: the backbone and the auxiliary stage. The backbone was performed by MobilNetV3Small which is the lighter version of MobileNetV3. In order to have the best features, we used the last feature extractor layer that has an output stride of 16 called C4 and the last one that has an output stride of 32 called C5. Moreover, the SSDLiteX auxiliary stage was composed of a 3 × 3 depthwise convolutional layer and a 1 × 1 convolutional layer was used. To achieve our goal, we made a reduction of the filter numbers. Rather than used 256 filters, our system was used 32, 32 and 16 filters at the first, the third and the last stage respectively. These tricks ameliorate the performance and reduce the complexity of the model. It has only 0.9 million (3.44 MB) of parameters and has a size of 6.17 MB. The Figure 4 shows the architecture of our model.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 12 Dec. 2024, pp: 306-312 www.ijaem.net ISSN: 2395-5252

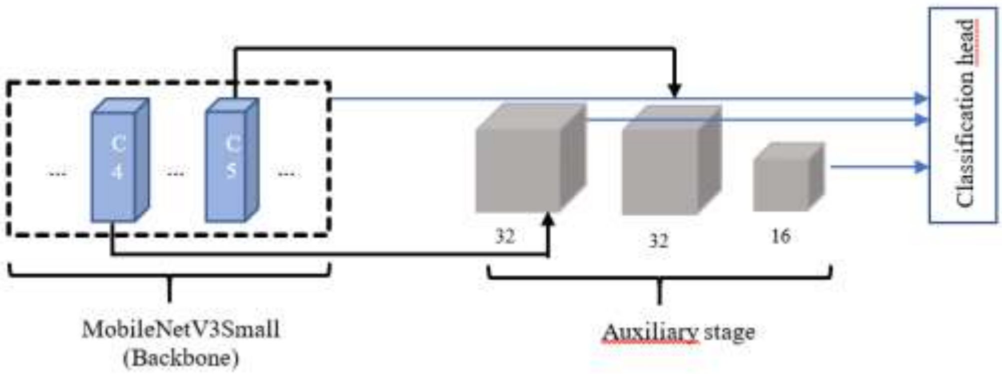


Figure 4 - Model architecture [2]

IV. RESULTS AND DISCUSSION

In this section we describe the proposed model's experiment result. A lot of experiments were conducted during the training phase but the batch size of 30 and an epoch of 100 performed the best result. The model was implemented on Intel(R) Core (TM) i5-5300U CPU @ 2.30GHz (4 CPUs), ~2.3GHz and 8 Gb RAM. We have been developed our system using python,TensorFlow, NumPy, Matplotlib, Scikit-learn, OpenCV, Panda and Keras libraries. We applied a shuffling technique with a seed of 123 in order to minimize loss, to have a lower variance and ameliorate the generalization of the proposed model. To evaluate our approach, we refer to two metrics: the accuracy and F1-score. The accuracy gives the proportion of the correct prediction made by the proposed model. So, this statistical measure computes the ration for both true positives and true negatives to the true total number cases. F1-score on the over hand can be interpreted as a harmonic mean of the precision and recall. It gives a balance between these two metrics. Outcome of the training and the testing phases, the proposed approach achieves 97% for overall accuracy and the global F1-score. The F1-score value demonstrate that the model is reliable to classify each class within the dataset. Furthermore, our result is comparable to the previous works even if we used a lightweight model. Thus, it could be deployed on a mobile system to resolve the problem in the development country. The figure 5 below shows the curve of results.

Figure 5–the training and validation loss - the training and validation accuracy - the training and validation F1 score

The confusion matrix below illustrates the global result of the classification task. It shows the high performance of the model into classifying each class.

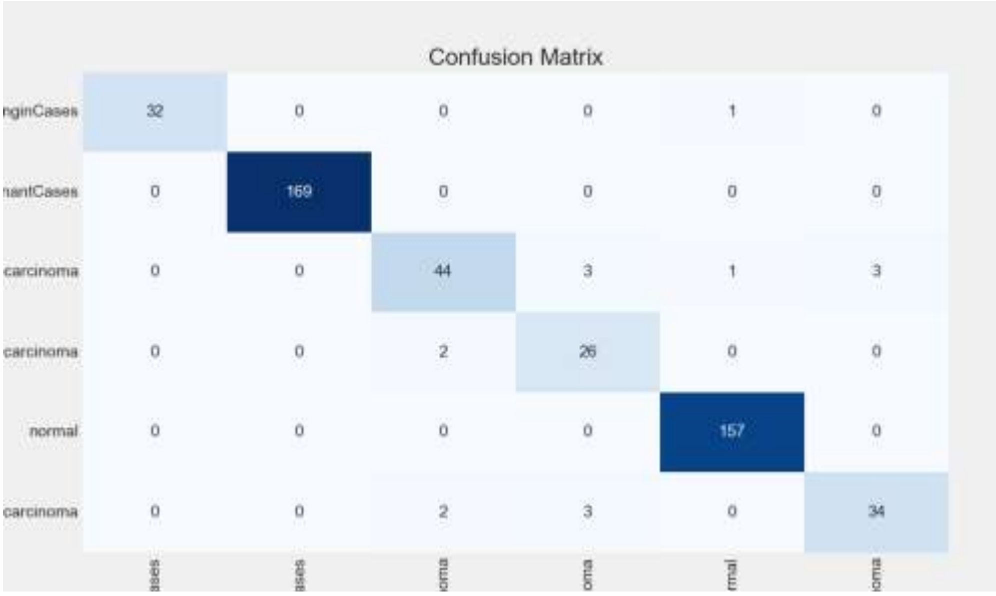


Figure 6 – Confusion Matrix

V. CONCLUSION

In conclusion, we have described in this paper the result of our research about lung cancer classification. Firstly, we have made a survey of previews work in order to know the state of the art about this team. Thus, we have remarked that the majority of previous model was complex and heavier and these required an infrastructure expensive. This situation led us to perform a lightweight system inspired from the SSDLiteX approach to classify lung cancer into six classes using CT scan images. Our model compared with previews work achieve a performance comparable. In its nature, it could run on the low resource environment and answer the need of the specialist within the hospitals.

REFERENCES

[1]. Abdul Rahaman Wahab Sait, 2023. "Lung Cancer Detection Model Using Deep Learning Technique". Applied Sciences 2023, Volume 13. https://doi.org/10.3390/app132212510

[2]. Andriamasinoro and Adolph, 2024. "Plant leaf disease classification using a lightweight model". Journal of Emerging Technologies and Innovative Research 2024, Volume 11, Issue 1. Pp: 225-229.

[3]. Bokefode et al., 2022. "Ensemble Deep Learning Models for Lung Cancer Diagnosis in Histopathological Images". Procedia Computer Science Volume 215, 2022, pp: 471-482.https://www.sciencedirect.com/journal/procedia-computer-science

[4]. Bushara et al., 2023. "An ensemble method for the detection and classification of lung cancer using Computed Tomography images utilizing a capsule network with Visual Geometry Group". Biomedical Signal Processing and Control Volume 85, August 2023.https://doi.org/10.1016/j.bspc.2023.104930

[5]. Bushara and Vinod, 2022. "Deep Learning-based Lung Cancer Classification of CT Images using Augmented Convolutional Neural Networks". Electronic Letters on Computer Vision and Image Analysis, Volume 21 Issue 1, pp: 130-142. https://doi.org/10.5565/rev/elcvia.1490.

[6]. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J. Digit. Imaging 2013, 26, pp: 1045–1057.

[7]. Damayanti and al., 2023. "Lung cancer classification using convolutional neural network and DenseNet", Journal of Soft Computing Exploration, Volume 4 N° 3 September 2023, pp: 133-141. https://doi.org/10.52465/joscex.v4i3.177

[8]. Ealistephen Neal Joshua, 2021. "Lung Cancer Classification using deep learning".

International Journal of Food and Nutritional Sciences, Volume 10 Issue 1, 2021

[9]. Kalaivani et al., 2020. "Deep Learning Based Lung Cancer Detection and Classification". IOP Conf. Series: Materials Science and Engineering 994 (2020), 6 pages.https://doi.org/10.1088/1757-899X/994/1/012026

[10]. Krishna et al, 2024. "Lung Cancer Prediction and Classification Using Decision Tree and VGG16 Convolutional Neural Networks". The Open Biomedical Engineering Journal.https://doi.org/10.2174/0118741207290271240322061032.

[11]. Li, P.; Wang, S.; Li, T.; Lu, J.; HuangFu, Y.; Wang, D. A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) [Dataset]. Cancer Imaging Arch. 2020.

[12]. Lung Image Database Consortium image collection (LIDC-IDRI).https://www.kaggle.com/datasets/justinkirby/the-cancer-imaging-archive-lidcidri.Access on 25/11/2024 at 8:12 a.m.

[13]. Manop, 2021. "Ensemble Transfer Learning for Lung Cancer Detection". DSIT 2021: 2021 4th International Conference on Data Science and Information Technology pp: 438 - 442https://doi.org/10.1145/3478905.3478995.

[14]. Mohamed Hany, Chest CT-Scan images Dataset, Version 1 2023https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images?datasetId=839140

[15]. Rashmi Mothkur a and B N Veerappa, 2023. Classification Of Lung Cancer Using Lightweight Deep Neural Networks. International Conference on Machine Learning and Data Engineering, Procedia Computer Science 218 (2023), pp: 1869–1877.

[16]. Sandag et al., 2024. "Comparative Analysis of Lung Cancer Classification Models Using EfficientNet and ResNet on CT-Scan Lung Images". COGITO Smart Journal – Volume 10, No. 1, June 2024, pp:259-270.

[17]. Song and al., 2017. "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images". Hindawi, Journal of Healthcare Engineering, Volume 2017, Article ID 8314740, 7 pages.https://doi.org/10.1155/2017/8314740.

[18]. Taher et al., 2021. "An Overview of Lung Cancer Classification Algorithms and their Performances". International Journal of Computer Science, Volume 48 Issue 4. 7 pages.

[19]. Youlden et al., 2008. "The International Epidemiology of Lung Cancer Geographical Distribution and Secular Trends". Journal of Thoracic Oncology • Volume 3, Number 8, August 2008, pp: 819-831.

[20]. https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-datase. Access on 28/11/2024 at 13:12 p.m.

[21]. https://www.kaggle.com/datasets/dishantrathi20/ct-scan-images-for-lung-cancer. Access on 30/11/2024 at 17:20 p.m.