Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 13 NO: 2 PAGE: 445-459 YEAR: 2024
DOI:10.17798/bitlisfen.1422869

Detection of Lung Cancer Cells Using Deep Learning Methods

Muhittin GENÇ [1*], Funda AKAR[2]

[1] *Institute of Science and Technology, Erzincan Binali Yıldırım University, Erzincan 24002, Turkey*
[2]*Department of Computer Engineering, Erzincan Binali Yıldırım University, Erzincan 24002, Turkey*
(ORCID: 0009-0002-5276-9244) (ORCID: 0000-0001-9376-8710)

Keywords: CNN, Deep Learning, Lung Cancer Detection, YOLO

Abstract

Lung cancer stands out as a high mortality, fatal disease worldwide. Early diagnosis is crucial for effective treatment of this disease; however, treatment options can be limited when it is often diagnosed in advanced stages. This study examines the role of artificial intelligence (AI) techniques in early diagnosis of lung cancer and emphasizes the advantages it provides. Particularly, the ability of deep learning algorithms to extract meaningful features from complex datasets indicates significant potential for detecting early stages of lung cancer. In this context, it is anticipated that AI-supported diagnostic systems have the potential to significantly improve lung cancer diagnostic methods by reducing the workload of radiologists and increasing accuracy rates. In this study, a total of 6 datasets were obtained by applying Gabor filter and Histogram Equalization+CLAHE filter to original datasets. The results obtained in the diagnosis of lung cancer using Convolutional Neural Networks (CNN) and YOLO algorithms are evaluated in two different categories. One of these categories is the investigation of the effect of image preprocessing methods. The other is the investigation of the effect of dataset partitioning into training, testing, and validation on success. According to the results obtained, the highest success rate in terms of F1 Score for the CNN model was achieved in both dataset partitioning (70%-20%-10% and 60%-20%-20%) with the datasets subjected to Histogram Equalization+CLAHE filter. It was obtained as 99%. For the YOLO model, the highest success rate was determined as 96% F1 Score with the same preprocessing technique and dataset partition. The effect of image preprocessing and dataset partitioning on success is not as high in the YOLO model as it is in the CNN model.

1. Introduction

Lung cancer is a serious disease that arises from the uncontrolled proliferation of cells in lung tissue, making normal breathing difficult for the patient. This type of cancer begins with the occurrence of DNA structure abnormalities, leading to the formation of an abnormal mass called a tumor, which in turn results in the excessive proliferation of cells [1]. Lung cancer is the deadliest among all cancer types worldwide. As seen in Figure 1, according to data released by the World Health Organization (WHO), approximately 1.8 million people worldwide lost their lives due to lung cancer in 2020, and 2.2 million new cancer cases were reported [2]. In Turkey, within the same year, the incidence rate of lung cancer among all cancer types was determined to be 17.6%, with 41,264 new cases detected [3]. Due to its higher mortality rate compared to other cancer types, early diagnosis of lung cancer is crucial.

In the evolving era of information technology, studies focusing on solutions to this issue have intensified, particularly with the use of artificial intelligence techniques. Artificial intelligence denotes the ability of computers to perform activities specific to human intelligence, such as carrying out various

*Corresponding author: mgenc@erzincan.edu.tr
Received: 20.01.2024, Accepted: 18.03.2024

tasks, solving complex problems, learning, and making decisions. Especially in the field of healthcare, the use of artificial intelligence aims to facilitate physicians in diagnosing diseases and increase operational efficiency. In this study aimed at this purpose, deep learning methods were employed to detect cancerous regions in lung images obtained from Computed Tomography (CT) scans.

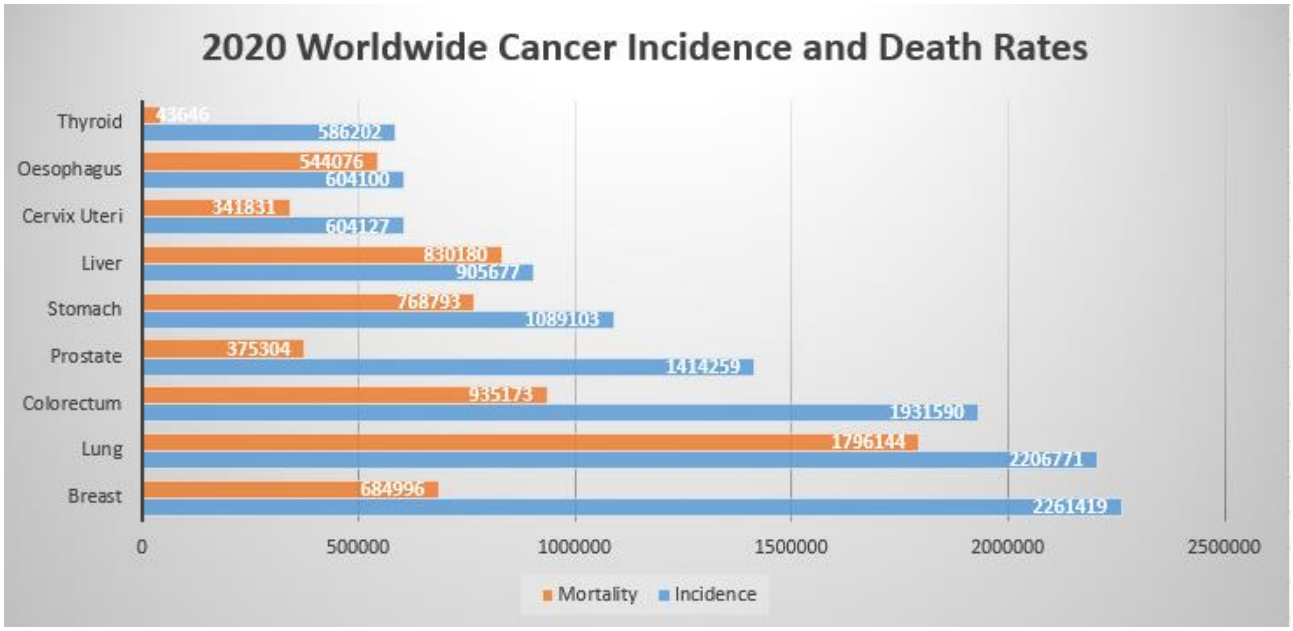

Figure 1. Global cancer cases worldwide according to 2020 data [2].

When examining the studies on this topic, it is noticeable that there has been a gradual increase in research, especially after 2019, and a concentration of studies offering solutions using various artificial intelligence methods. Keshani et al. conducted a study using CT images to segment cancerous lungs into regions through active contour modeling. Subsequently, various masking techniques were applied, and nodules were classified using Support Vector Machines (SVM) based on 2D stochastic and 3D anatomical features, achieving a detection accuracy of 89% [4]. Kuruvilla and Gunavathi performed lung segmentation from CT images, extracting various statistical parameters such as standard deviation, skewness, kurtosis, mean, fifth central moment, and sixth central moment. They utilized these parameters for classification using feedforward and backpropagation algorithms of artificial neural networks. The results indicated an accuracy of 91.1% for the first method and 93.3% for the second method [5]. De Carvalho Filho et al. developed a methodology for the detection of lung cancer nodules from CT images using pattern recognition and image processing techniques. They applied clustering algorithms to classify structures resembling lungs after segmenting the images. In the final classification, they employed Micro-Genetic Algorithm in addition to SVM, achieving sensitivity, specificity, and accuracy rates of 85.91%, 97.70%, and 97.55%, respectively [6]. Song et al. utilized three deep learning methods, Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), and Stacked Autoencoders (SAE), to detect benign and malignant nodules from lung CT images. The models achieved accuracies of 84.15% for CNN, 82.37% for DNN, and 82.59% for SAE [7]. Lustberg et al. investigated the use of computer-assisted segmentation in the delineation of organs at risk for lung cancer. They observed that manually performed segmentation took an average of 20 minutes, whereas the atlas-based software reduced this time to 7.8 minutes, and deep learning-based software further reduced it to 10 minutes [8]. In another study, various image processing techniques were applied to improve the quality of CT images, followed by cropping unnecessary details expressing redundancies in the images. The resulting images were modeled by a deep neural network algorithm with an increased number of layers. The obtained results showed an accuracy of 94.56%, sensitivity of 96.2%, and specificity of 94.2% [9]. To enhance the highlighting of cancerous regions in CT images, histogram equalization was performed in another study. They developed a deep learning method using Improved Profuse Clustering Technique (IPCT) and Deep Learning with Instantaneously Trained Neural Networks (DITNN), achieving a 98.42% accuracy after training [10]. A study in 2019 focused on detecting and classifying

diseases such as healthy lungs, Chronic Obstructive Pulmonary Disease (COPD), and fibrosis from CT images. Three newly emerged algorithms, namely Improvised Crow Search Algorithm (ICSA), Improvised Grey Wolf Algorithm (IGWA), and Improvised Cuttlefish Algorithm (ICFA), were employed to determine features. For classification, SVM, k-Nearest Neighbor (K-NN), and Decision Tree Classifier were used, and the best combination of classification models and feature extraction methods was observed to be IGWA + K-NN with a 99.4% accuracy [11]. Kasinathan et al. segmented the lungs using CT images, dividing them into subregions after applying Gaussian distribution for feature extraction. They achieved a classification accuracy of 97% using CNN [12]. In a study conducted in 2020, CNN was used to work on a lung CT dataset collected from Iraqi hospitals. A technique with an AlexNet architecture aimed to detect whether the target population in CT scans was benign or malignant, achieving an accuracy of up to 93.5% [13]. Nanglia et al. aimed to improve classification accuracy through a hybrid classification algorithm. They reported a 98.08% accuracy with the model they developed [14]. A study comparing Artificial Neural Networks (ANN), CNN, and Recurrent Neural Networks (RNN) on CT images found that the ANN model marginally outperformed RNN and CNN models with an accuracy of 71.18% [15]. Chen et al. applied the SegNet approach to facilitate the diagnosis of lung cancer from CT images, attempting to detect benign and malignant tumors. The manual detection accuracy for lung cancer was 86.25%, while SegNet achieved an accuracy of 92.50%, DeepLab v3 reached 80.41%, and VGG-19 had an accuracy of 79.58% [16]. In 2021, a research aimed at detecting lung cancer using artificial intelligence techniques employed three different CNN models on the LC25000 Lung and Colon Histopathological Image dataset. Inception_ResNet_V2 achieved a 99.7% accuracy, VGG-19 had 92.99%, and ResNet 50 showed 99.4% accuracy [17]. In the study conducted by Talukder et al. in 2022, they focused on early detection of colon and lung cancer using a hybrid feature extraction method. VGG16, MobileNet, and DenseNet201 models were employed, achieving a detection rate of 99.05% for lung cancer and 100% for colon cancer [18]. Another study on lung cancer utilized Internet of Things (IoT) technology in addition to CT images. Data from both CT and wearable technologies were combined, processed through an Extended Convolutional Neural Network (ECNN), resulting in an accuracy of 96.8% [19]. Haznedar and Simsek analyzed lung and renal cell cancer RNA-Seq data sets from the Cancer Genome Atlas (TCGA) using both classical machine learning methods and deep learning techniques. Among classical machine learning methods, Random Forest yielded the best results with an accuracy range of 93.51% to 91.83%. For deep learning methods, DNN-Adadelta achieved accuracy rates of 95.54% and 96.15% [20]. In another research effort, the VGG16 backbone was combined with Single Shot Detection (SSD) for detecting Osteosarcoma nodules metastasizing to the lungs, resulting in a model with an accuracy of 75.97% [21]. For rapid and accurate diagnosis of lung and colon cancers using CT images, Class Selective Image Processing (CSIP) was applied, and training was performed using an AlexNet neural network model. The accuracy increased from 89.8% before CSIP to 98.8% after its application [22]. A study focused on classifying abnormalities in lung nodules used an optimized hybrid approach of ICSA and CNN-based Long Short-Term Memory (LSTM) methods. Simulation results indicated that the proposed method achieved an accuracy rate of 98% [23]. Shanthi et al. aimed to improve the detection and classification of lung cancer cells by creating a new hybrid algorithm with Naïve Bayes, Decision Tree, and ANN classifiers, incorporating Stochastic Diffusion Search (SDS)-based feature selection. The results showed that the proposed TABU-SDS-NN achieved an accuracy of 94.07% [24]. In a study conducted in 2023, three different CNN models were created for detecting lung cancer cells from CT images. By averaging the results of these three models, a new approach was proposed, and the average accuracy of the three CNNs was measured at 95% [25].

In this study, a comparison of methods for lung cancer diagnosis using CNN and YOLO was conducted with a total of 6 datasets. The impact of image preprocessing methods such as Gabor filter and Histogram Equalization + CLAHE applied prior to the training process of the models on their success was investigated. Additionally, the effect of dataset partitioning, with ratios of 70%-20%-10% and 60%-20%-20%, on the accuracy of the models was also analyzed.

2. Material and Method

This study adopts a different approach from other studies in the literature concerning the importance of the mentioned topic. It transforms the dataset into three different datasets, including the original version, by applying various preprocessing methods. The aim of this approach is to determine the impact of the applied preprocessing methods on success.

The first dataset includes the original Cancer Imaging Archive (TCIA) dataset [26], [27]. To enhance success, a second dataset was obtained by applying a Gabor filter to the images. Additionally, images in the original dataset were subjected to histogram equalization followed by Contrast Limited Adaptive Histogram Equalization (CLAHE) to create a third dataset. After expanding the dataset, models were compared in terms of the metrics specified below using CNN and YOLO version 8, whose popularity has increased in recent years.

- Confusion Matrix: A table used to determine the success of models. TP represents true positives, FN represents false negatives, FP represents false positives, and TN represents true negatives (Figure 2).

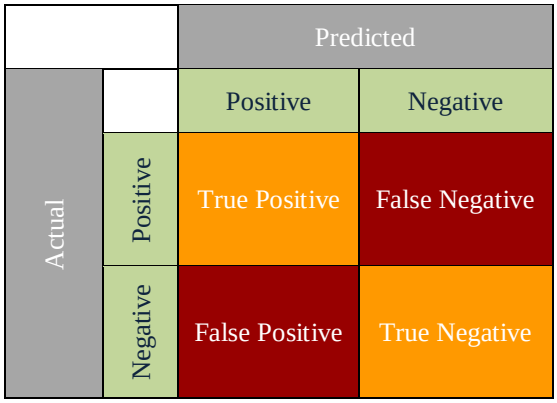

Figure 2. Representation of the Confusion Matrix

- Accuracy (ACC): A measure of how accurate the created model is.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Precision (P): Indicates the positive prediction rate of the model.

$$P = \frac{TP}{TP+FP} \quad (2)$$

- Recall (R): Also known as sensitivity or true positive rate, formalized as follows:

$$R = \frac{TP}{TP+FN} \quad (3)$$

- F1 Score: The harmonic mean of precision and recall, considered a better measure than accuracy.

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision*Recall} \quad (4)$$

- Average Precision (AP): Expressed as the area under the precision-recall curve.

$$AP = \int_0^1 p(r)dr \cong \sum_{k=1}^{n} p(k).\Delta r(k) \quad (5)$$

- Mean Average Precision (mAP): A metric used to evaluate the performance of the model. mAP measures the accuracy of detected objects and is often evaluated with a threshold value determined using a confidence threshold. While mAP50 represents the case where a confidence threshold of 0.5 is used, mAP50-95 calculates performance at different confidence thresholds between 0.5 and 0.95. This metric indicates the accuracy rate of the objects detected by the model. The mAP value is higher when the model detects objects more accurately. When calculating mAP, precision (AP) measurements are made separately for each class, and then the arithmetic average of these values is taken to obtain the overall mAP value.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (6)$$

Convolutional Neural Networks (CNNs): Convolutional Neural Networks are deep learning algorithms that take input images, pass them through various filters to extract features, and produce results by combining them with kernels [28]. The use of CNNs extends beyond image classification, object recognition, and face detection into the medical field, where these advancements facilitate disease diagnosis, ultimately improving survival rates.

You Only Look Once (YOLO): YOLO is a deep learning model used in object detection. It is a method that can detect objects in an image in a single pass, making it suitable for real-time applications such as video. Thanks to this feature, YOLO provides a fast and effective solution for object detection tasks.

The flowchart of the study is presented in Figure 3, and sample images for each dataset are shown in Figure 4.

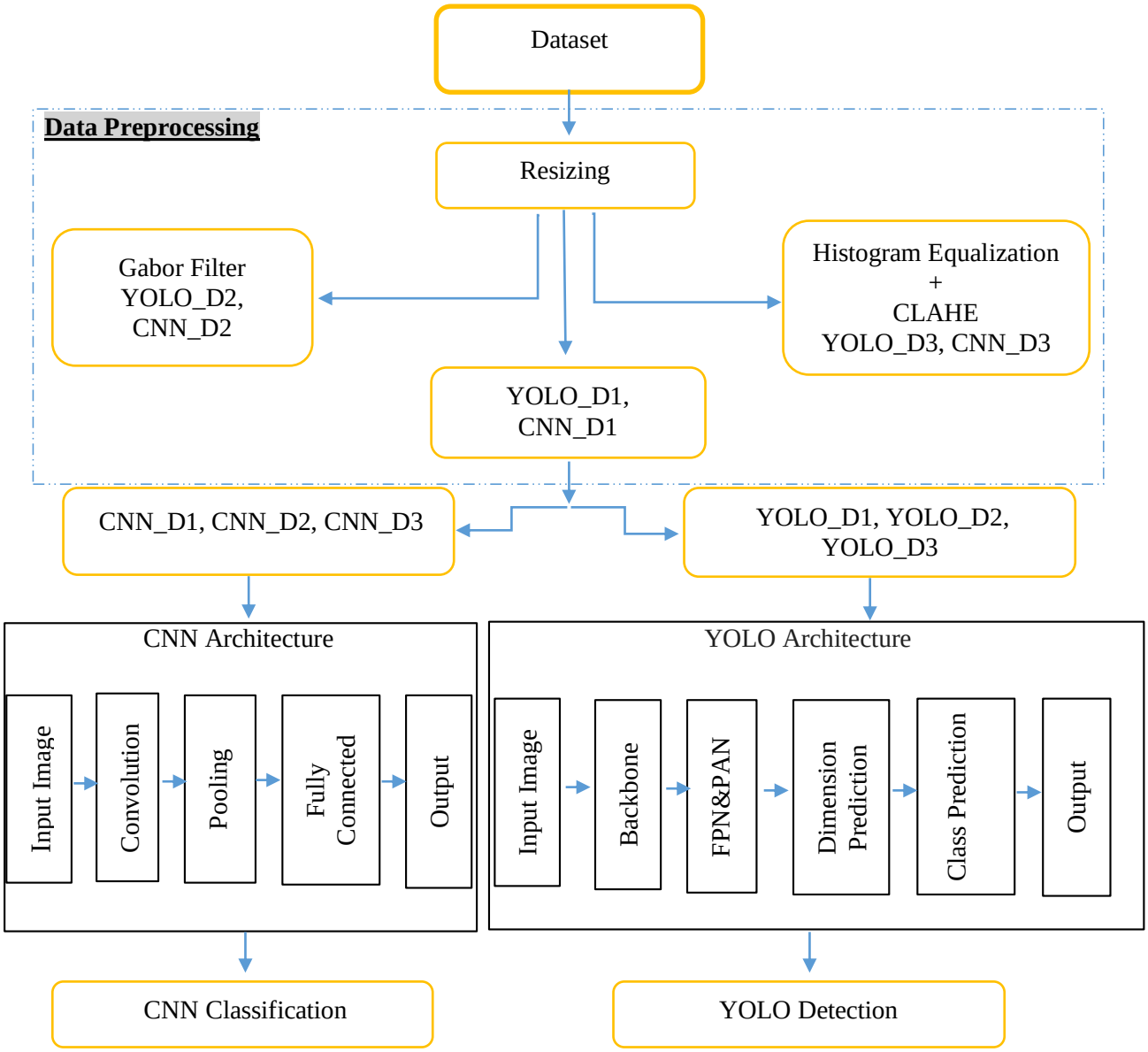

Figure 3. Workflow of the Application

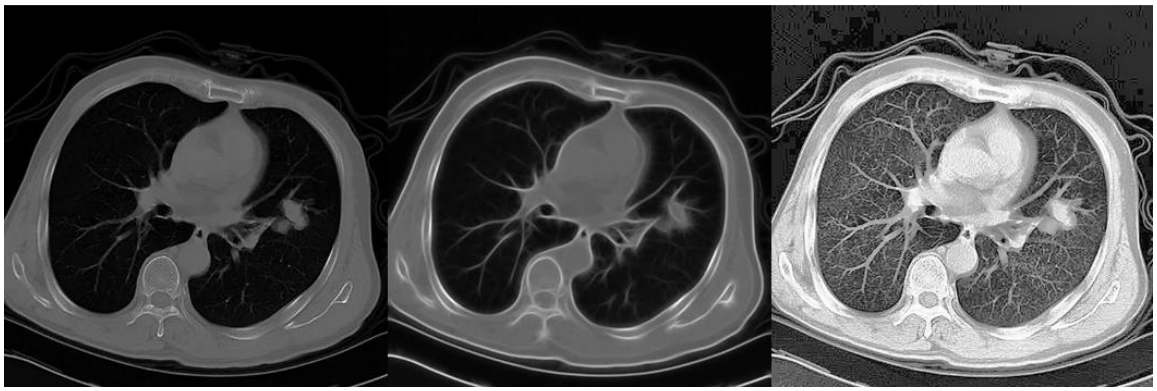

Figure 4. Example images from the dataset used for CNN and YOLO models: (a) original image, (b) image with Gabor filter applied, (c) image with histogram equalization and CLAHE applied

2.1. Data Preparation

The YOLO dataset consists of 4968 CT images with dimensions of 512x512 pixels. The CNN dataset comprises a total of 7940 images with dimensions of 512x512 pixels, half of which are cancerous and the other half normal. The images in the CNN dataset are reduced to a size of 64x64 pixels. This size reduction aims to save runtime and resource usage. Since the YOLO model achieves higher success on large-scale images, no dimension changes are made to the YOLO dataset.

Two different partitioning methods are employed to analyze the datasets. In the first partitioning, 70% of the dataset is used for training, 20% for validation, and 10% for testing. In the second partitioning, these ratios are set at 60% for training, 20% for validation, and 20% for testing. As the YOLO and CNN datasets are transformed into three separate datasets, the subsequent representations in this study are provided in the Table 1.

Table 1. Naming of datasets

Dataset	Description of the Dataset
D1	Original dataset
D2	The dataset created by applying the gabor filter
D3	The dataset created by applying histogram equalization + CLAHE

2.2. Data Preprocessing

Data preprocessing is a crucial step that helps improve the quality of images. In this study, various preprocessing steps, including resizing, formatting, and converting images to the desired format, were applied to the images in the dataset. In addition to the original dataset (D1), a Gabor filter was first applied to both datasets to investigate its impact on performance (D2). Subsequently, a new dataset was created for the original dataset by applying histogram equalization + CLAHE processing (D3). As a result, three different datasets, each subjected to different preprocessing methods for YOLO and CNN, were obtained. Classification processes were then carried out on these datasets.

CNN Model

Three different convolution processes of 32x3x3, 64x3x3, and 64x3x3 were applied to the input image. After each convolution process, pooling of 4x4, 2x2, and 2x2 was sequentially applied to reduce the height and width of the image. Additionally, a dropout parameter of 0.5 was used after each convolution to prevent overfitting and memorization by the model. Finally, the model was passed through the sigmoid activation function to obtain results. These processes aim to extract various features from the input image and support the model in learning more generally. The model was run for 50 epochs on a computer with an Intel Core i5 7200U 2.5 GHz processor and 8 GB RAM. This process was performed for three different datasets and two different split procedures. Training times are detailed in Table 2.

YOLO Model

The training of the YOLO model, based on the YOLO v8 architecture, was conducted for 60 epochs using Python programming language and Google Colaboratory with Google Drive. To expedite and facilitate the training process, a Tesla T4 GPU provided by Google was employed. The processing of the dataset and training duration are detailed in Table 2. The preference for the Tesla T4 GPU was motivated by the ability to complete the training in a few hours, compared to several days with a CPU. This choice aimed to save time and enhance the efficiency of the training process.

Table 2. Training durations for CNN and YOLO

	CNN		YOLO	
Dataset	**%70-%20-%10**	**%60-%20-%20**	**%70-%20-%10**	**%60-%20-%20**
D1	3 hours 23 min	3 hours 28 min	2 hours 27 min	2 hours 57 min
D2	3 hours 37 min	3 hours 36 min	2 hours 19 min	2 hours 50 min
D3	3 hours 05 min	3 hours 02 min	2 hours 31 min	2 hours 54 min

3. Research Findings

The results of the 12 studies conducted with a total of 6 datasets obtained through the division process of 2 different datasets for the 2 models used are presented in Table 3 and Table 4.

The images in the datasets used for CNN are labeled as cancerous and normal. Therefore, as

shown in Table 3, two values are observed for cancerous/normal in each of the F1 Score, Precision, and Recall values. However, in the YOLO datasets, since the training process is conducted solely on cancerous images, such a scenario is not applicable for YOLO.

Table 3. CNN Model Results

	Dataset	Preprocessing	Train Acc	Train Loss	Val Acc	Val Loss	Test Acc	Test Loss	F1-Score Cancer - Normal	Precision Cancer - Normal	Recall Cancer - Normal
CNN	Dataset 60-20-20	Original (D1)	0.985	0.061	0.894	0.480	0.89	0.47	0.89-0.90	0.97-0.84	0.81-0.98
		Gabor Filter (D2)	0.998	0.004	0.920	0.250	0.95	0.29	0.92-0.93	0.98-0.87	0.86-0.98
		Histogram Equalization+ CLAHE (D3)	0.996	0.01	0.989	0.0241	0.992	0.024	0.99-0.99	0.98-1.0	1.0-0.98
	Dataset 70-20-10	Original (D1)	0.985	0.685	0.935	0.307	0.80	0.67	0.93-0.94	0.98-0.90	0.89-0.98
		Gabor Filter(D2)	1.0	0.0027	0.925	0.256	0.98	0.082	0.92-0.93	1.0-0.87	0.85-1.0
		Histogram Equalization+ CLAHE (D3)	0.999	0.059	0.995	0.182	0.993	0.020	0.99-0.99	0.99-1.0	1.0-0.99

Table 4. YOLO Model Results

	Dataset	Preprocessing	Train mAP (50)	Train mAP (50-95)	Val mAP (50)	Val mAP (50-95)	F1-Skor/ Threshold	Precision/ Threshold	P-R	Recall
YOLO V8	Dataset 60-20-20	Original (D1)	0.980	0.621	0.98	0.621	0.96/ 0.471	1.0/0.862	0.980 mAP@ 0.5	0.98
		Gabor Filter (D2)	0.978	0.601	0.978	0.602	0.96/ 0.478	1.0/0.837	0.978 mAP@ 0.5	0.98
		Histogram Equalization+ CLAHE (D3)	0.981	0.616	0.981	0.615	0.96/ 0.512	1.0/0.831	0.981 mAP@ 0.5	0.98
	Dataset 70-20-10	Original (D1)	0.976	0.612	0.975	0.611	0.95/ 0.492	1.0/0.836	0.976 mAP@ 0.5	0.98
		Gabor Filter(D2)	0.974	0.604	0.973	0.603	0.96/ 0.361	1.0/0.838	0.974 mAP@ 0.5	0.97
		Histogram Equalization+ CLAHE (D3)	0.977	0.61	0.977	0.61	0.96/ 0.249	1.0/0.831	0.977 mAP@ 0.5	0.98

Ideally, both high Recall and high Accuracy are aimed for in a model, yet this situation may vary depending on the specific problem and application. In the CNN-based study, it was observed that the dataset with the applied Gabor filter (D2) achieved higher accuracy in both ways the dataset was split, although the training time with D2 was found to be longer compared to other datasets. Considering other evaluation parameters (F1-Score, Precision, and Recall), it can be seen that the dataset obtained by histogram equalization + CLAHE (D3) yields better results, and the processing time is significantly lower compared to D1 and D2 datasets. Additionally, the dataset split, while not having a highly significant impact on the model's success, led to a noticeable change. The split of 70%-20%-10% produced higher values. According to the model evaluation with the test data not used in the training stage, D3 outperformed all other datasets. In conclusion, preprocessing applied to the images in the dataset was found to enhance the model's performance.

In the case of YOLO-based studies, it was observed that the success rate of the D2 dataset decreased compared to D1 at the mAP (50) threshold, while D3 had a better accuracy compared to other methods. The dataset split also showed a small change in the model's success. Here, the 60%-20%-20% split yielded higher success rates in all three datasets. It was observed that the preprocessing applied to the images in the dataset was not as effective on the success of the YOLO model as it was in the CNN model.

Another success metric, the confusion matrix, shows how accurate the models predictions are. Figures 5 and 6 present confusion matrices and accuracy curves for the CNN model with D1, D2, and D3 datasets in the case of a 60%-20%-20% split. Similarly, Figures 7 and 8 show the confusion matrices and accuracy curves for the CNN model with D1, D2, and D3 datasets in the case of a 70%-20%-10% split. Upon examination, it can be observed that, for both split scenarios, D3 yields more successful results compared to other methods, and D2 is closer to D1 but exhibits a higher level of accuracy than D1.

When analyzing the accuracy curves in Figures 6 and 8, it is observed that, for both split scenarios, D3 exhibits less fluctuation compared to other methods, indicating a more stable performance. This analysis demonstrates that the D3 dataset performs more effectively for the CNN model, providing more reliable results.

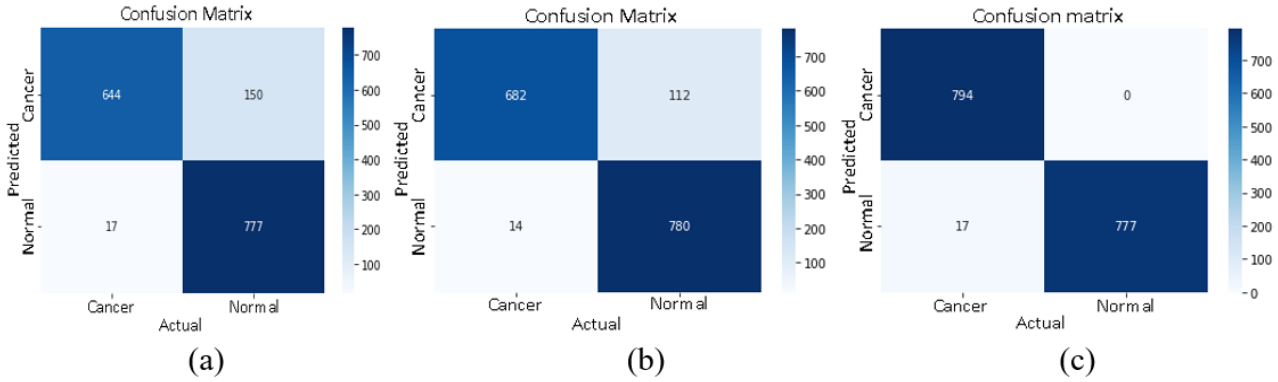

Figure 5. CNN model for 60%-20%-20% split: (a) confusion matrix for D1, (b) confusion matrix for D2, and (c) confusion matrix for D3

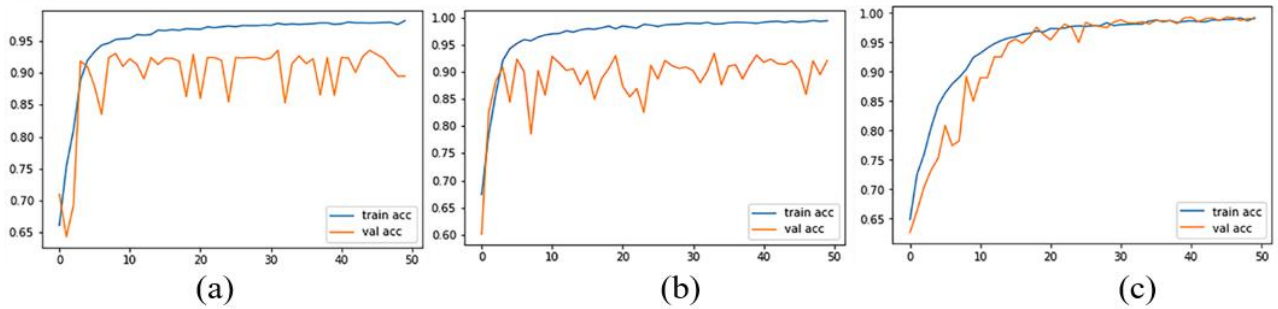

Figure 6. CNN model for 60%-20%-20% split: (a) accuracy curves for D1, (b) accuracy curves for D2, and (c) accuracy curves for D3

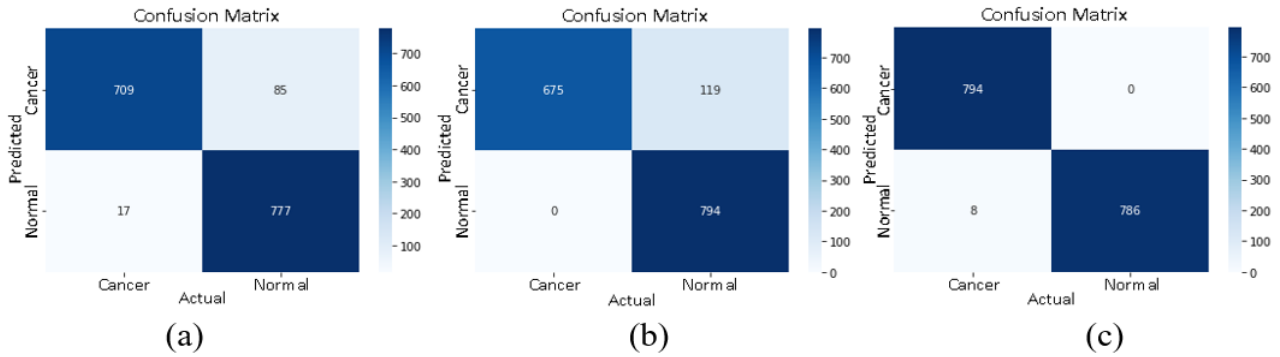

Figure 7. CNN model for 70%-20%-10% split: (a) confusion matrices for D1, (b) confusion matrices for D2, and (c) confusion matrices for D3

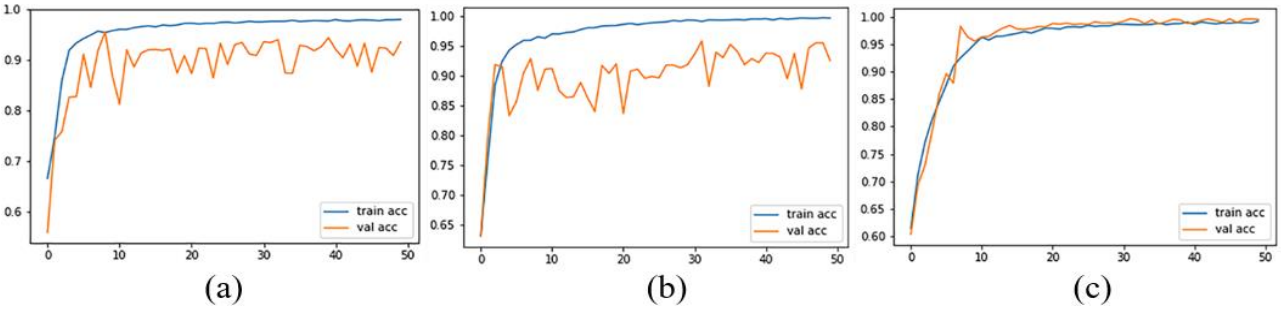

Figure 8. CNN model for 70%-20%-10% split: (a) accuracy curves for D1, (b) accuracy curves for D2, and (c) accuracy curves for D3

Figure 9 and Figure 10 show the confusion matrices of the YOLO model for the 60%-20%-20% and 70%-20%-10% split scenarios, respectively, for D1, D2, and D3 datasets. When examining the confusion matrices, it is observed that D3 achieved better results by correctly predicting nodule 1 with 96% and nodule 2 with 100% in Figure 9, and nodule 1 with 98% and nodule 2 with 97% in Figure 10 compared to other datasets. Additionally, in cases where the model could not make any detections, it was noted that it perceived and labeled the images as the background. This observation is crucial for assessing the detection capability of the model.

For the YOLO model in the 60%-20%-20% split, the accuracy and loss curves for the D1, D2, and D3 datasets, along with changes in performance parameters with increasing epoch counts, are respectively illustrated in Figures 11, 12, and 13. In the 70%-20%-10% split, the accuracy and loss curves for the D1, D2, and D3 datasets are shown in Figures 14, 15, and 16. The mAP(50) and mAP(50-95) values started to increase as the epoch count increased, reaching over 97% for mAP(50) and up to 62% for mAP(50-95). Additionally, it was observed that the classification loss dropped below 0.5. Moreover, in both splits, the mAP curves for the D3 dataset exhibited a more stable pattern compared to the other datasets. As can be seen from Figure 11, 12, and Figure 14, 15, the performance metrics of YOLO obtained after training with D1 and D2 datasets indicate less fluctuation in the graphs of the D1 dataset for both dataset partitioning methods (%60-%20-%20 and %70-%20-%10) compared to the D2 dataset. This suggests that the D1 dataset exhibits a more stable structure for both partitioning methods compared to the D2 dataset.

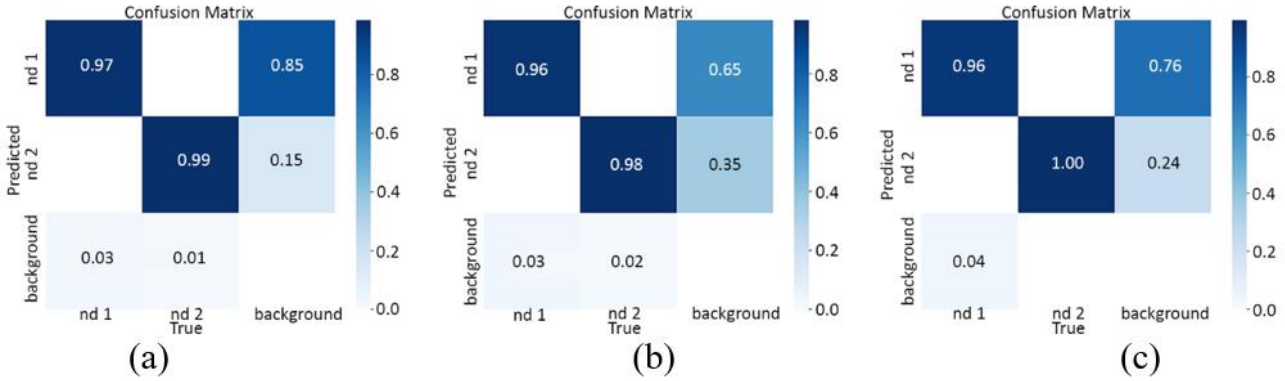

Figure 9. Confusion matrices for the YOLO model in the 60%-20%-20% split for (a) D1, (b) D2, and (c) D3

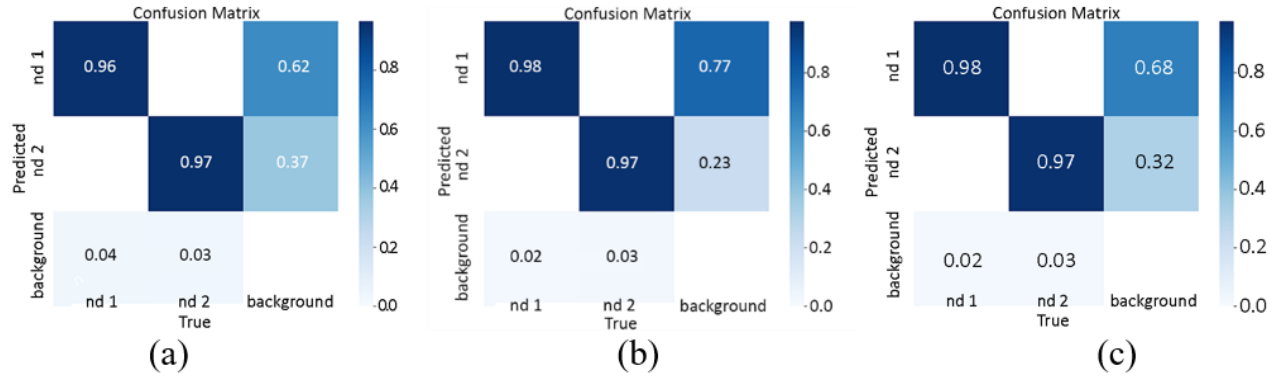

Figure 10. Confusion matrices for the YOLO model in the 70%-20%-10% split for (a) D1, (b) D2, and (c) D3

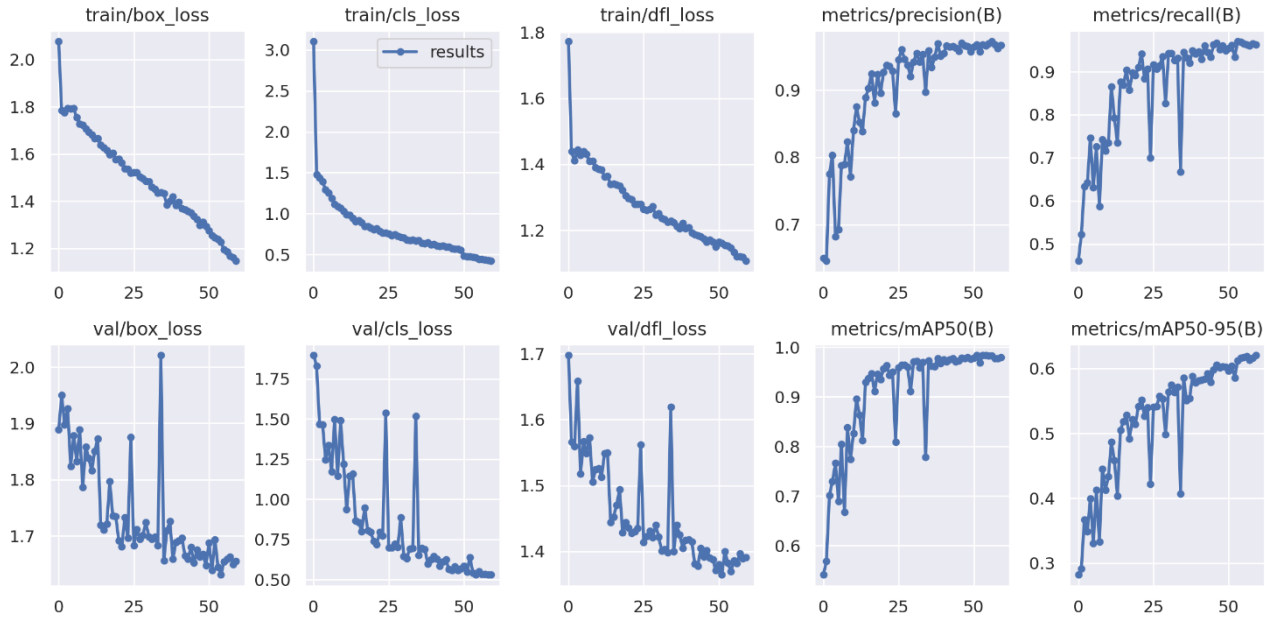

Figure 11. Accuracy and loss curves for YOLO model with %60-%20-%20 split for dataset D1

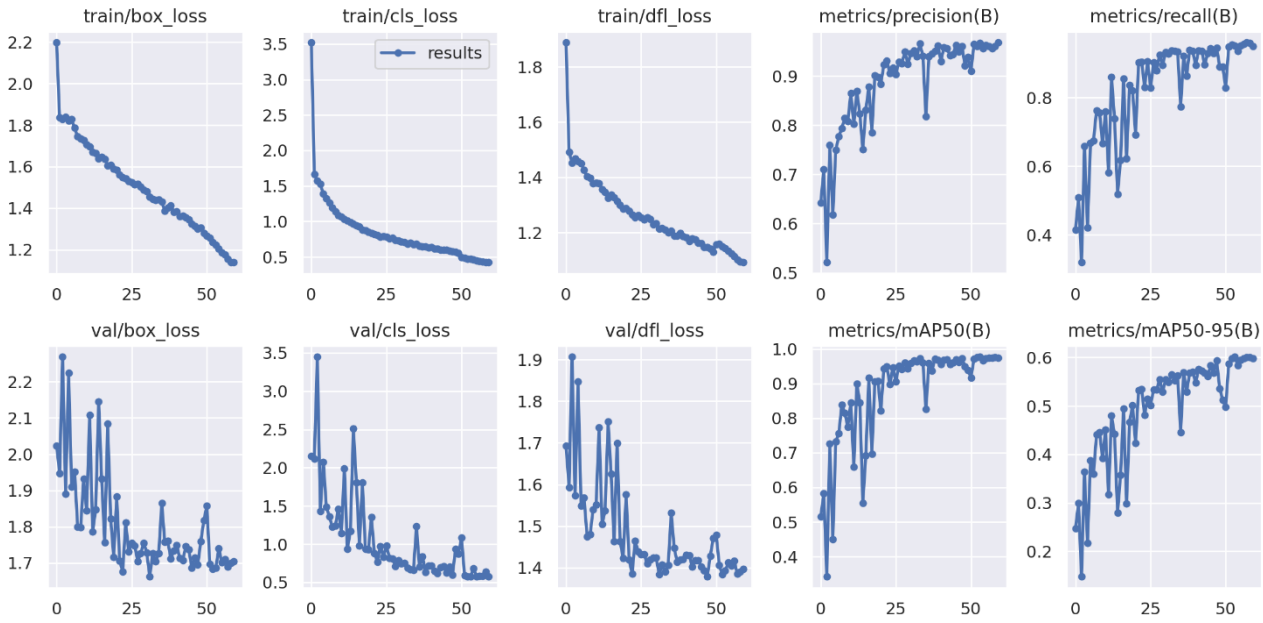

Figure 12. Accuracy and loss curves for YOLO model with %60-%20-%20 split for dataset D2

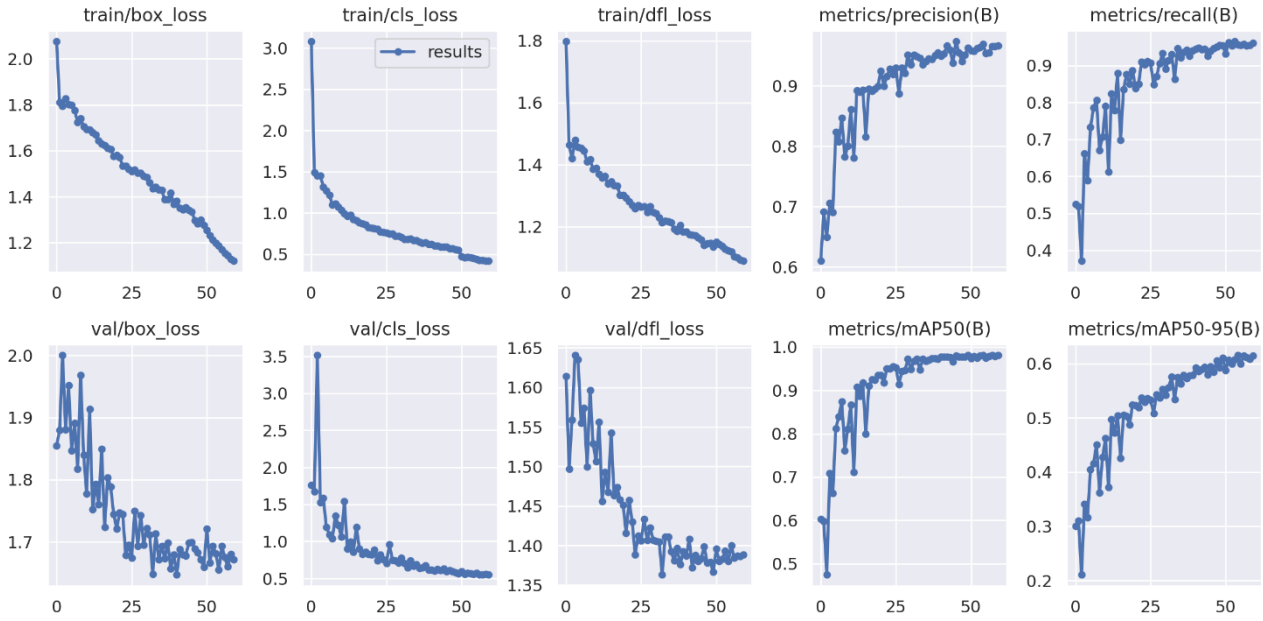

Figure 13. Accuracy and loss curves for YOLO model with %60-%20-%20 split for dataset D3

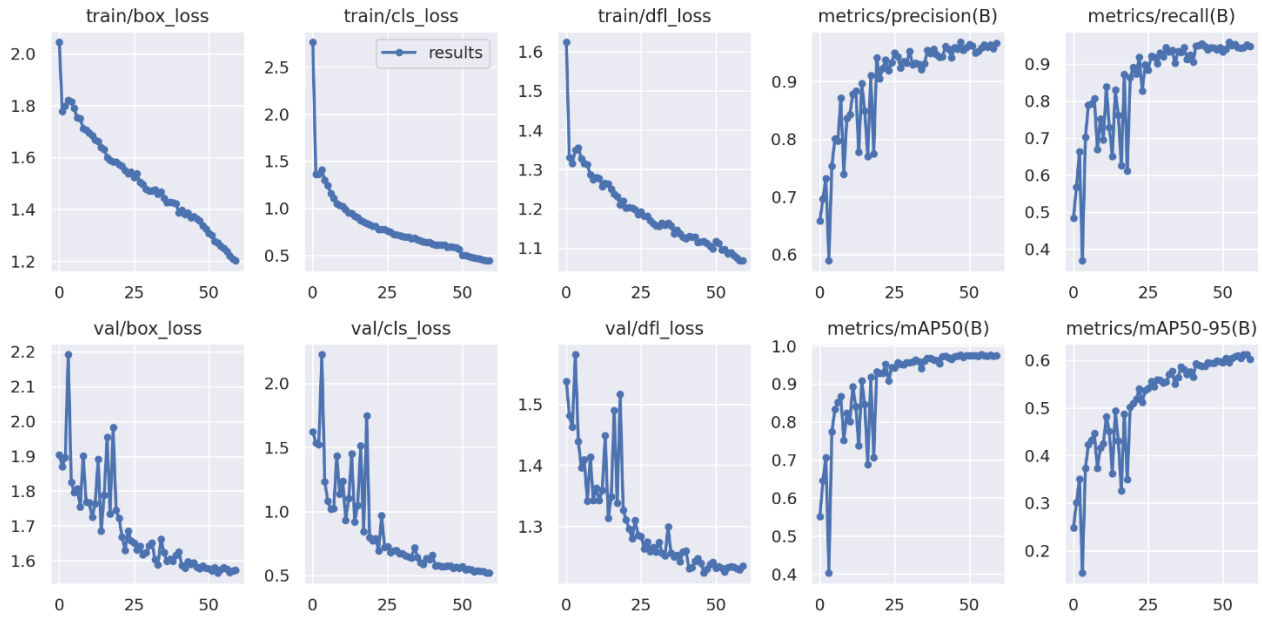

Figure 14. Accuracy and loss curves for YOLO model with %70-%20-%10 split for dataset D1

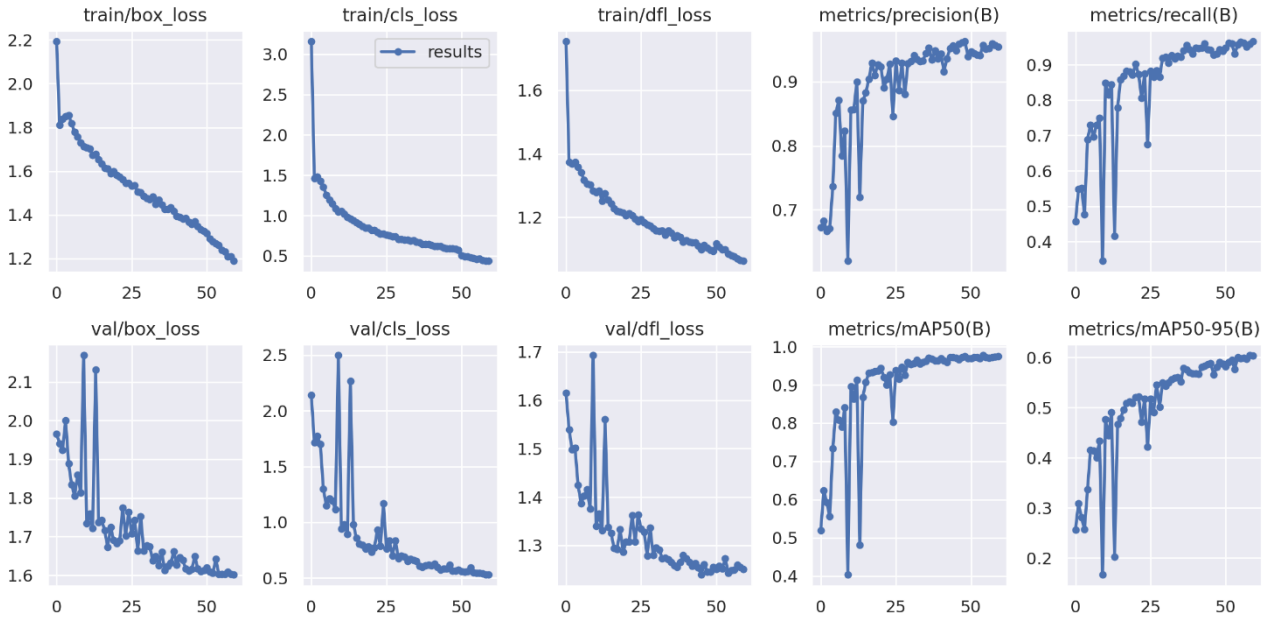

Figure 15. Accuracy and loss curves for YOLO model with %70-%20-%10 split for dataset D2

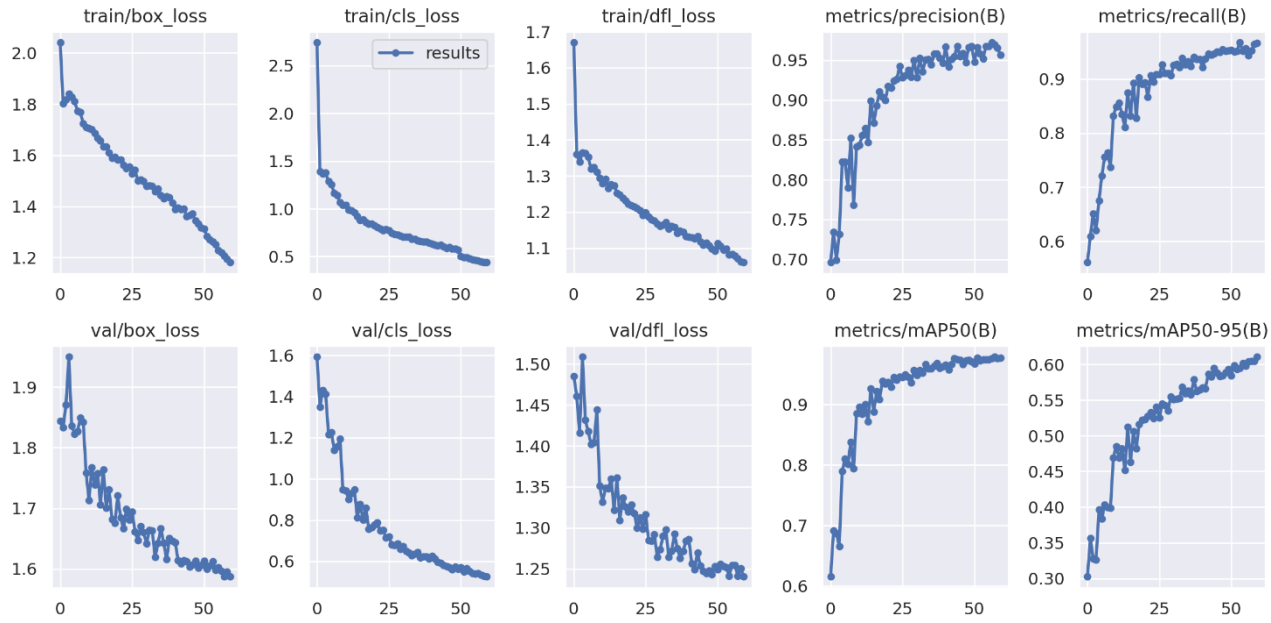

Figure 16. Accuracy and loss curves for YOLO model with %70-%20-%10 split for dataset D3

4. Conclusion and Suggestions

This research investigates the use of artificial intelligence components, namely CNN and YOLO algorithms, for the early diagnosis of lung cancer and evaluates the effectiveness of these algorithms. The dataset used for the CNN model consists of a total of 7940 images, with half being cancerous and the other half healthy. For the YOLO model, the dataset comprises 4968 CT images. The datasets were segmented using different image preprocessing techniques, and the impact of these techniques on the performance of models was thoroughly examined. The proposed CNN model achieved an accuracy of 98.5% when trained with the original dataset (D1), 99.85% with the dataset obtained by applying histogram equalization + CLAHE filter (D3), and 100% with the dataset where the Gabor filter was applied (D2). The F1 score, a metric considered better than accuracy, reached the highest value of 99% with the D3 dataset. The F1 score values for the other datasets were observed to be 94% for D1 and 93% for D2.

For the YOLO model, the highest success rate in terms of mAP scale was achieved with the D3 dataset, where histogram equalization + CLAHE was applied (98.1%). The mAP values for the other datasets were measured as 98% for D1 and 97.8% for D2. In terms of F1 score, all datasets had approximately the same value for the YOLO model. Success parameters and graphical analyses indicate that the histogram equalization + CLAHE method outperforms other methods. Additionally, it was observed that image preprocessing and dataset segmentation have a significant impact on success rates.

References

[1] C. Yan and N. Razmjooy, “Optimal lung cancer detection based on CNN optimized and improved Snake optimization algorithm,” *Biomed Signal Process Control*, vol. 86, p. 105319, 2023, doi: https://doi.org/10.1016/j.bspc.2023.105319.

[2] Global Cancer Observatory, “World,” 2020. Accessed: Dec. 02, 2023. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf

[3] Global Cancer Observatory, “Turkey,” 2020. Accessed: Jan. 18, 2024. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/792-turkey-fact-sheets.pdf

[4] M. Keshani, Z. Azimifar, F. Tajeripour, and R. Boostani, “Lung nodule segmentation and recognition using SVM classifier and active contour modeling: A complete intelligent system,” *Comput Biol Med*, vol. 43, no. 4, pp. 287–300, May 2013, doi: 10.1016/j.compbiomed.2012.12.004.

[5] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," *Comput Methods Programs Biomed*, vol. 113, no. 1, pp. 202–209, Jan. 2014, doi: 10.1016/j.cmpb.2013.10.011.

[6] A. O. De Carvalho Filho, W. B. De Sampaio, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index," *Artif Intell Med*, vol. 60, no. 3, pp. 165–177, 2014, doi: 10.1016/j.artmed.2013.11.002.

[7] Q. Z. Song, L. Zhao, X. K. Luo, and X. C. Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images," *J Healthc Eng*, vol. 2017, 2017, doi: 10.1155/2017/8314740.

[8] T. Lustberg *et al.*, "Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer," *Radiotherapy and Oncology*, vol. 126, no. 2, pp. 312–317, Feb. 2018, doi: 10.1016/j.radonc.2017.11.012.

[9] S. K. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar, and G. Ramirez, "Optimal deep learning model for classification of lung cancer on CT images," *Future Generation Computer Systems*, vol. 92, pp. 374–382, Mar. 2019, doi: 10.1016/j.future.2018.10.009.

[10] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks," *Measurement (Lond)*, vol. 145, pp. 702–712, Oct. 2019, doi: 10.1016/j.measurement.2019.05.027.

[11] N. Gupta, D. Gupta, A. Khanna, P. P. Rebouças Filho, and V. H. C. de Albuquerque, "Evolutionary algorithms for automatic lung disease detection," *Measurement*, vol. 140, pp. 590–608, 2019, doi: https://doi.org/10.1016/j.measurement.2019.02.042.

[12] G. Kasinathan, S. Jayakumar, A. H. Gandomi, M. Ramachandran, S. J. Fong, and R. Patan, "Automated 3-D lung tumor detection and classification by an active contour model and CNN classifier," *Expert Syst Appl*, vol. 134, pp. 112–119, Nov. 2019, doi: 10.1016/j.eswa.2019.05.041.

[13] H. F. Al-Yasriy, M. S. Al-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, "Diagnosis of Lung Cancer Based on CT Scans Using CNN," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1757-899X/928/2/022035.

[14] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, "A hybrid algorithm for lung cancer classification using SVM and Neural Networks," *ICT Express*, vol. 7, no. 3, pp. 335–341, 2021, doi: https://doi.org/10.1016/j.icte.2020.06.007.

[15] S. Doppalapudi, R. G. Qiu, and Y. Badr, "Lung cancer survival period prediction and understanding: Deep learning approaches," *Int J Med Inform*, vol. 148, p. 104371, 2021, doi: https://doi.org/10.1016/j.ijmedinf.2020.104371.

[16] X. Chen, Q. Duan, R. Wu, and Z. Yang, "Segmentation of lung computed tomography images based on SegNet in the diagnosis of lung cancer," *J Radiat Res Appl Sci*, vol. 14, no. 1, pp. 396–403, Dec. 2021, doi: 10.1080/16878507.2021.1981753.

[17] N. Baranwal, P. Doravari, and R. Kachhoria, "Classification of Histopathology Images of Lung Cancer Using Convolutional Neural Network (CNN)." [Online]. Available: https://orcid.org/0000-0002-1113-7884

[18] Md. A. Talukder, Md. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, "Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning," *Expert Syst Appl*, vol. 205, p. 117695, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117695.

[19] A. B. Pawar *et al.*, "Implementation of blockchain technology using extended CNN for lung cancer prediction," *Measurement: Sensors*, vol. 24, p. 100530, 2022, doi: https://doi.org/10.1016/j.measen.2022.100530.

[20] B. Haznedar and N. Y. Simsek, "A Comparative Study on Classification Methods for Renal Cell and Lung Cancers Using RNA-Seq Data," *IEEE Access*, vol. 10, pp. 105412–105420, 2022, doi: 10.1109/ACCESS.2022.3211505.

[21] C. Loraksa, S. Mongkolsomlit, N. Nimsuk, M. Uscharapong, and P. Kiatisevi, "Development of the Osteosarcoma Lung Nodules Detection Model Based on SSD-VGG16 and Competency Comparing With Traditional Method," *IEEE Access*, vol. 10, pp. 65496–65506, 2022, doi: 10.1109/ACCESS.2022.3183604.

[22] S. Mehmood *et al.*, "Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing," *IEEE Access*, vol. 10, pp. 25657–25668, 2022, doi: 10.1109/ACCESS.2022.3150924.

[23] M. Kanipriya, C. Hemalatha, N. Sridevi, S. R. SriVidhya, and S. L. Jany Shabu, "An improved capuchin search algorithm optimized hybrid CNN-LSTM architecture for malignant lung nodule detection," *Biomed Signal Process Control*, vol. 78, p. 103973, 2022, doi: https://doi.org/10.1016/j.bspc.2022.103973.

[24] S. Shanthi, V. S. Akshaya, J. A. Smitha, and M. Bommy, "Hybrid TABU search with SDS based feature selection for lung cancer prediction," *International Journal of Intelligent Networks*, vol. 3, pp. 143–149, 2022, doi: https://doi.org/10.1016/j.ijin.2022.09.002.

[25] A. A. Shah, H. A. M. Malik, A. Muhammad, A. Alourani, and Z. A. Butt, "Deep learning ensemble 2D CNN approach towards the detection of lung cancer," *Sci Rep*, vol. 13, no. 1, p. 2987, 2023, doi: 10.1038/s41598-023-29656-z.

[26] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," Journal of Digital Imaging.

[27] Mehmet Fatih AKCA, "Veri seti." Accessed: Jan. 19, 2024. [Online]. Available: https://universe.roboflow.com/mehmet-fatih-akca/yolotransfer/dataset/2

[28] R. Chauhan, K. K. Ghanshala, and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 2018, pp. 278–282. doi: 10.1109/ICSCCC.2018.8703316.