

Lung Cancer Detection on Chest X-rays using Deep Convolutional Neural Networks

Jackson Kehoe
Queen's University
jkehoe00@gmail.com

John Zhou
Queen's University
johnzhou7913@gmail.com

David Nguyen
Queen's University
david.nguyen@queensu.ca

Frank Siyung Cho
Queen's University
20fsc@queensu.ca

Abstract—Early diagnosis of lung cancer is critical for early intervention and prognosis in cancer treatment. Chest X-rays are an accessible, cost-effective and low-radiation imaging method to investigate lung cancer. The team explored the application of machine learning algorithms to X-ray images to identify lung cancer. Our team used the a subset of the ChestXray-8 dataset and narrowed the data down to masses and nodules found in the lungs. The data was then split for an even distribution of cancer and non-cancer images. An AlexNet, a VGG-16 and a ResNet50 are the machine learning models that are investigated. A smaller subset of the data provided has bounding boxes, the bounding boxes will be included in the model it achieves small loss amounts. The team was able to achieve a 73.26% accuracy, a precision of 0.73, a recall score of 1.00, and an F1 score of 0.85 on the testing data. After training, the bounding box model achieved a 8.97×10^{-4} loss. Further work on the project includes exploring different models to improve the scores, obtaining more bounding box data to create a localization feature, and creating an interactive application that can be used in a medical setting.

I. INTRODUCTION

This is the paper for the QMIND team exploring machine learning algorithm applications to detect lung cancer. This paper contains all the information achieved through the design process and findings of the team.

A. Motivation

Lung cancer is the leading cause of cancer mortality in Canada and worldwide [1]-[2]. It is estimated that 30,000 were diagnosed with lung cancer in 2022, and a further 20,700 will die of it, representing approximately 24% of cancer deaths that year [3]. The Canadian health care system has been weakened due to the COVID-19 pandemic, with many health care providers facing burnout due to the stress of working overtime hours while being understaffed. In a survey of health care workers conducted in 2021, 83% of physicians said they felt more stressed, 65% stated they had an increase of workload, and 46% have had to work additional hours [4]. Due to this environment, 11% of doctors intend to leave health care or change careers within the next 3 years [4]. This loss in practitioners could further strain the system.

Wait times for medical imaging have significantly increased since the COVID-19 pandemic, with some provinces reporting median wait times of 7 to 8 weeks for a computed tomography (CT) scan and 12 to 20 weeks for a magnetic resonance imaging (MRI) scan compared to the recommended 30-day wait time [5].

Prevention, screening and early diagnosis are critical for reducing the burden of cancer and improving patient prognosis and survival rates [1]. While CT scans have been proven to be superior to chest X-rays for lung cancer screening, chest X-rays remain more accessible, cheaper and provide low exposure to radiation [6]. The highest quality studies in the literature suggest that the sensitivity of radiologist interpretation of chest X-rays for symptomatic lung cancer is approximately 77-80% [7]. Thus, augmentation of diagnostic accuracy using computer-aided detection software on chest X-rays may be warranted to further improve rates of early lung cancer diagnosis where low-dose CT scans are unavailable. These issues can be alleviated by leveraging deep convolutional neural networks (DCNNs) trained on chest X-rays to identify cancerous masses and nodules. Computer-aided detection systems using DCNNs can accelerate and assist medical providers during the diagnostic process and reduce the number of cancer cases that go undiagnosed.

B. Problem Definition

The problem is to correctly identify possible cancerous areas on the lungs based on chest radiographs and, if possible with a sufficiently high degree of accuracy, to additionally localize the cancerous tissue using bounding boxes. By classifying and localizing the presence of lung cancer on chest X-rays, the diagnostic process and accuracy of lung cancer detection can be improved. With the growing role of computer vision in radiology, we acknowledge the importance of ethical practices regarding data privacy, algorithmic biases, as well as safety and transparency.

II. METHODOLOGY

A. Data

The data used is from the Chest-Xray8 data set that is publicly available online [9]. The data set contains 108,948 frontal view X-ray images from 32,717 unique patients that have been de-identified by name and given subject IDs. It includes the subject's age, but contains no description of their lifestyle, medical history or other personal characteristics. The data has labels associated with each image for different lung and heart afflictions. The data was first tested using all lung-related labels, and subsequently narrowed down to explore the accuracy of DCNNs for classifying the presence or absence lung cancer. The labels of masses and nodules were the best

clinical indicators for predicting lung cancer on chest X-rays. The provided data had already been split into training, testing and validation data. Initially the data had approximately 1,000 cancer positive images with nearly 50,000 cancer negative images. This resulted in the model predicting negative every time. The data was further processed for an even dispersion of data. Shown below in Figure 1 is a graph showing the dispersion between cancerous and non-cancerous images.

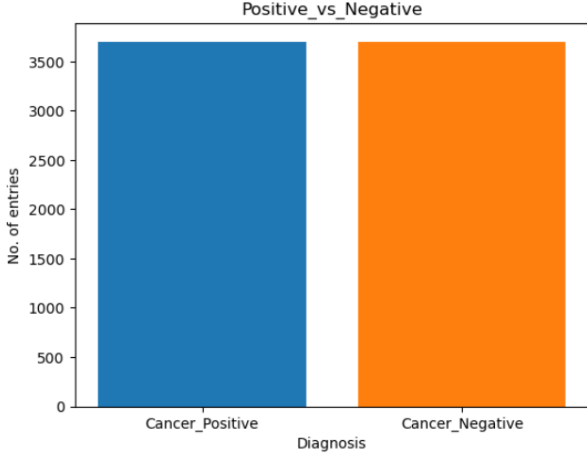


Fig. 1. Dispersion of the data used.

The images used for training and validation had a scale, shear, zoom, rotation, brightness range, horizontal and vertical flip applied to it. For testing, only a scale was applied. From these images, 6,300 images were used for the training set, 44 for the validation set, and 1,969 for the testing set.

B. Proposed Solutions

The proposed solution was to explore the accuracy of 3 different neural networks in predicting lung cancer. The chosen models were AlexNet, ResNet50 and a VGG-16 model due to their high accuracy in other machine vision applications. Each model will use the adaptive moment estimation (Adam) as their optimizer with a learning rate of 1e-4.

The AlexNet is based off a convolution neural network with 5 dense layers, 3 max pooling layers, 2 normalization layers, 2 fully connected layers and 1 softmax layer. The architecture can be seen in Figure 2.

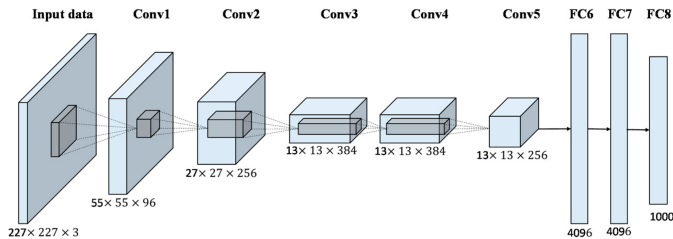


Fig. 2. Architecture of AlexNet[11].

The ResNet50 is a ResNet model consisting of 48 convolution layers with a MaxPool and Average Pool layer for a total of 50 layers. It uses ImageNet weights, which are initialized values for the layers to base their image classification on. The architecture can be seen in Figure 3.

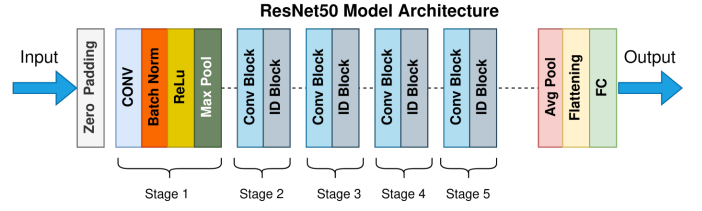


Fig. 3. Architecture of ResNet50[12].

The VGG-16 is an improvement to the AlexNet model. It has 13 convolutional layers and 3 fully connected layers, doubling the layers when compared with AlexNet. The architecture can be seen in Figure 4.

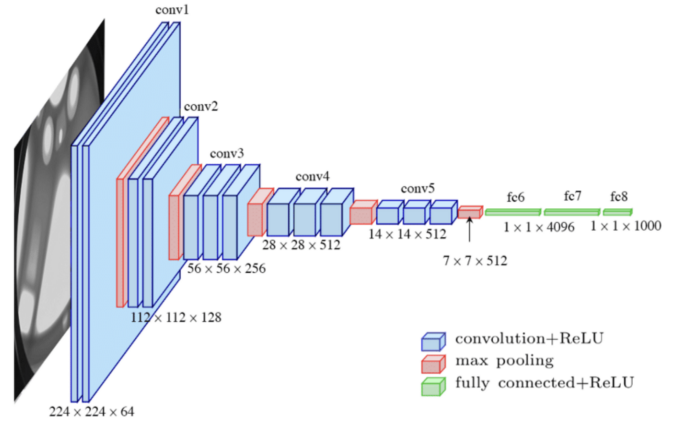


Fig. 4. Architecture of VGG-16[13].

C. Evaluation

The models are evaluated based on their accuracy when predicting positive and negative instances of cancer. 3 different accuracy measures are collected for the training, validation and testing. Additionally, measures for the precision, recall and F1-score will be collected to evaluate the effectiveness of each model. The equations for the accuracy measures are shown respectively below.

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

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

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

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

D. Bounding Boxes

Bounding boxes will be applied to the model if a regression model is able to accurately predict the area of concern. The data for bounding boxes is limited for the X-rays with approximately 160 entries for mass and nodules annotated. Additionally, applying bounding boxes to X-rays is difficult as it does not provide perfectly clear images when compared to CT scans. Shown below in Figure 5 is the annotated bounding box on an X-ray image with a mass/nodule.

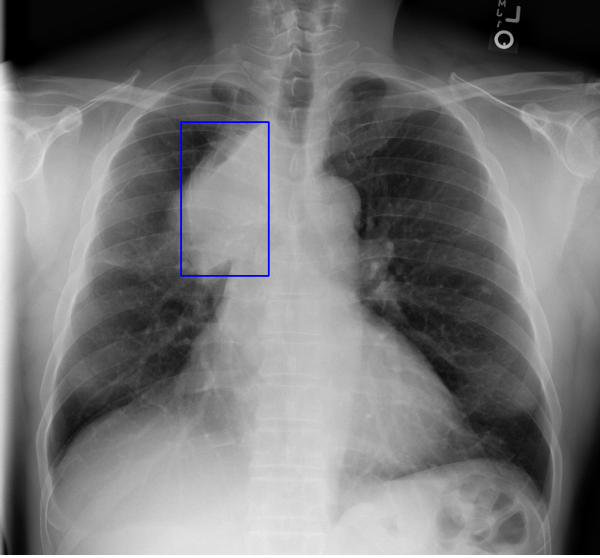


Fig. 5. Bounding Box on Training Image.

III. RESULTS

A. Classification

The performance of the 3 models were decided based on 4 metrics: Accuracy, Precision, Recall and F1. These 4 metrics were decided to be the most relevant however, other metrics such as Specificity or Negative Predictive Value (NPV) could have been used.

The accuracy metric was used to determine how many images the model could correctly classify as having masses/nodules from the total amount of sample images; this is demonstrated below in Equation 1.

The precision metric was used to determine the number of samples which were predicted to have masses/nodules from the total amount of samples with masses/nodules; this is demonstrated in Equation 2.

The recall metric also known as sensitivity or true positive rate was used to determine how many predicted images and actual images had masses/nodules from the total amount of images with masses/nodules; this is demonstrated in Equation 3.

Finally, the F1 metric was used to determine the number of correct predictions the model made; this is demonstrated in Equation 4.

The results for the classification task for all models can be found in Table 1.

TABLE I
METRICS FOR BINARY CLASSIFICATION USING DIFFERENT MODELS

Model	Testing Accuracy	Precision	Recall	F1
VGG-16	73.26%	0.73	1.00	0.85
ResNet50	57.29%	0.50	0.85	0.62
AlexNet	49.21%	0.49	1.00	0.66

Out of the models we tested for binary classification, VGG-16 performed the best overall. VGG-16 achieved an accuracy of 73.26%, precision of 0.73 and F-1 score of 0.85 for predicting cancer versus no cancer on the testing set.

Our results could have been limited by the data set. There was a low distribution of data and low amount of data available for cancer positive versus negative. With the 50/50 distribution we utilized to disperse data, the models were not exposed to the entire range of X-rays for training and were thus not generalizable. Moreover, further limitation of this model is that the Chest-Xray8 dataset used auto-annotated labels and were estimated to only be 90% accurate, potentially contributing to inaccuracies [9]-[10]. The model could be correct in its assessment, but the labelling on the data set may have been wrong.

It must also be noted that authors of the Chest-Xray8 data set tested various models on multi-label classification had significantly lower accuracies (56.44% and 71.64% respectively) for mass and nodule classification compared to other diseases [9].

The training results of the models can be seen through figures 6 to 8. 20-30 epochs of the training and validation sets were used for training. All 3 differed in accuracies for training and validation. It is important to reiterate that the validation set only had 44 images, so interpreting the results should mainly be focused on the training accuracy.

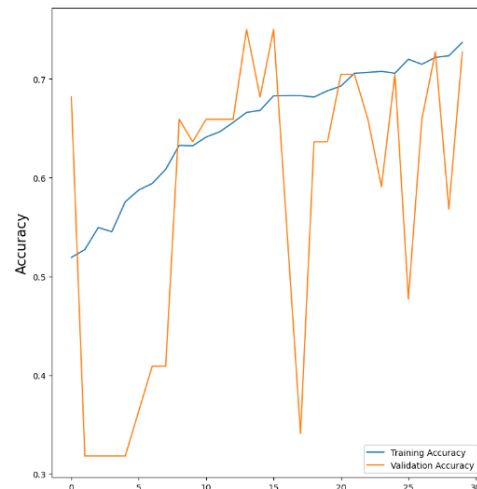


Fig. 6. ResNet50 Training and Validation Accuracy

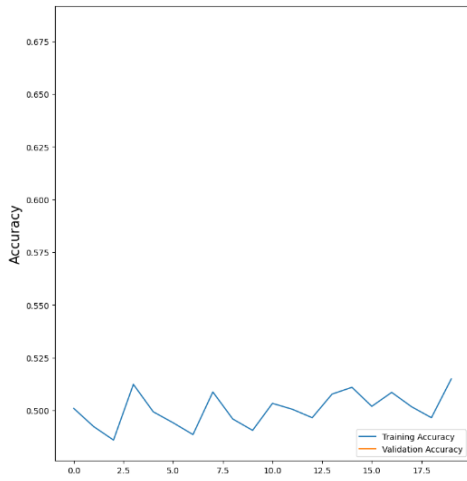


Fig. 7. AlexNet Training and Validation Accuracy

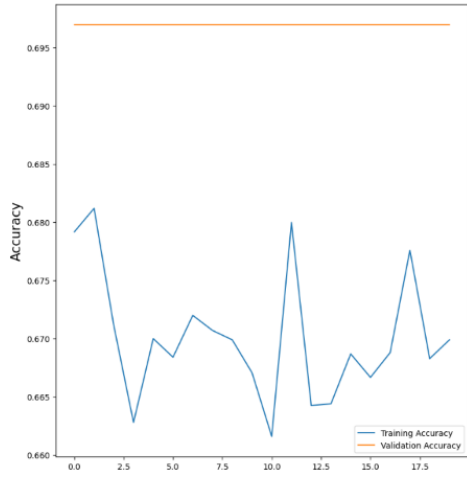


Fig. 8. VGG-16 Training and Validation Accuracy

ResNet50 gradually improves its training accuracy from about 50% to about 72%, but its validation accuracy remains inconsistent. As the accuracy has not plateaued by the end of training, this indicates that the accuracy could further improve with more epochs. AlexNet training accuracy remains steady around 50% throughout training. Due to error in running, the validation was not tracked, but the results suggest that it would be consistent with VGG-16's validation accuracy. VGG-16 produces consistent results throughout training, retraining a training accuracy around 67% and a validation accuracy around 70%. The AlexNet and VGG-16 consistent results suggest that the model have been overfitted early on and were not able to improve. This suggests that the models are too simple and could benefit from more layers, a different optimizer, a different loss metric, and/or a smaller learning rate.

B. Localization

The bounding boxes were trained on a basic CNN and loss was measured. The loss by the end of training was 8.97 X e-

04 on the training data and 0.0362 on the validation data. The values on the bounding boxes were normalized before being passed into the model. So these values are considered to be large. Good results were unable to be achieved in the bounding boxes due to lack of data available and distinction within the X-ray images. The results of a sample bounding box prediction is shown in blue on the image with green representing the true bounding box in Figure 9.

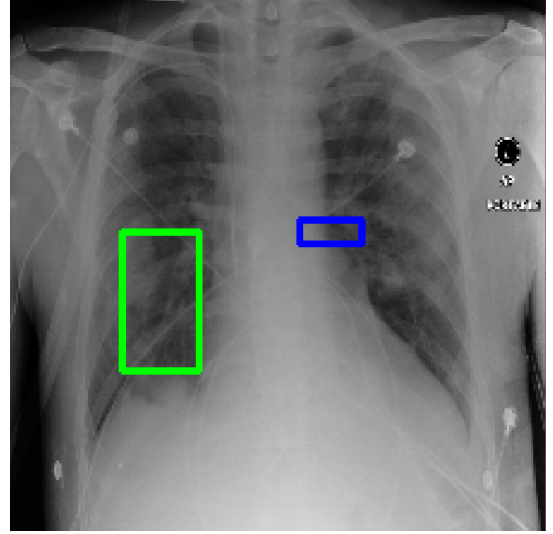


Fig. 9. Predicted bounding box shown in blue, actual bounding box shown in green.

IV. RELATED WORKS

In 2018, Ausawalaithong et al. used a DenseNet model with ImageNet weights and 121 layers to classify lung cancer through chest X-rays [14]. They trained their model on the ChestX-ray14 data set. They differed by training their model on 108 899 images with 4992 nodule images being the positives and deemed cancerous, but kept a 50/50 split of cancer non-cancer for the validation and testing set. They did not include masses as positives. 108 899 images were used for training, 2048 images were used for validation, and 532 images were used for testing. The models were trained on an unspecified number of epochs. On the data set, they achieved an accuracy of 84.02%, a specificity of 85.34%, and a recall value of 82.71% for nodule detection.

In this paper, X-ray scans were the primary medical imaging procedure used to identify masses/nodules. However, other imaging techniques, such as PET scans or MRI scans, can be used in order to obtain accurate results when applying machine learning models. In 2021, Idrahim et al. used a data set comprised of 33 676 X-ray and CT scans to 4 different image classification models, a VGG19-CNN, a ResNet152V2, a ResNet152V2 + Gated Recurrent Unit (GRU), and ResNet152V2 + Bidirectional GRU (Bi-GRU) to identify COVID-19, pneumonia, and lung cancer [15]. Of the 4 techniques, the VGG-19 model was deemed to have the most accurate results with an accuracy of 98.05%, a precision of 98.43%, a recall of 98.05% and a F1 score of 98.24%.

V. CONCLUSION

In this paper, 3 convolutional neural networks, an AlexNet, a ResNet50, and a VGG-16, were trained on chest X-ray images from the Chest-Xray8 data set. After applying data augmentation on an even split of images showing cancer and no cancer and splitting the data into training, validation, and testing sets, these models were unable to produce sufficiently high accuracy, precision, recall, and F1 scores for classifying chest X-ray images for lung cancer. Of the 3, VGG-16 performed the best with an accuracy of 73.26%, a precision of 0.73, a recall of 1.00, and an F1 score of 0.85. The bounding box regression model was also unable to produce bounding boxes that accurately mapped the area with masses or nodules.

A. Future Work

As evident in the DenseNet model produced by Ausawalaithong et al. on the same data set used in this paper, high scores in classifications are possible. This suggests a number of changes that could improve results: (1) Add more layers in the network - the DenseNet used 121 layers, while the ResNet50, the largest model evaluated in this paper, had 50 layers. (2) Add more images to validation - the DenseNet had 2048 images for validation while the 3 models had 44 images to validate from. Validation is not used to train the model, rather, it provides insight on how it performs while training. With a clearer output, researchers will have a better understanding on how to improve it. (3) Run for more epochs - more epochs should help the models converge in accuracy. As seen in the ResNet50 accuracy plot, the model has large spikes throughout training and had not plateaued in its learning yet. (4) Use a different optimizer, loss function, and/or decrease the learning rate - the AlexNet and VGG-16 models converged early on and do not suggest improvement through more epochs. Changing 1 or all 3 could improve on this.

The same suggestions can be applied to the bounding box regression model, however this can also be attributed to the small size of the data set with exact positions of mass. More lung cancer chest X-ray data sets should be considered to supplement this. In addition, since Chest-Xray8 had the limitation of using auto-annotated labels, further steps include using radiologist-annotated datasets. Employing a transfer learning approach with these DCNNs may also significantly improve performance with a low number of epochs. Furthermore, novel approaches employing an ensemble methodology with multiple DCNNs may improve the accuracy of predictions for lung cancer and diagnosis using chest X-ray images [16].

REFERENCES

- [1] R. Nooreldeen and H. Bach, "Current and Future Development in Lung Cancer Diagnosis," *IJMS*, vol. 22, no. 16, p. 8661, Aug. 2021, doi: 10.3390/ijms22168661.
- [2] J. Ferlay et al., "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012," *Int. J. Cancer*, vol. 136, no. 5, pp. E359–E386, Mar. 2015, doi: 10.1002/ijc.29210.
- [3] D. R. Brenner et al., "Projected estimates of cancer in Canada in 2022," *CMAJ*, vol. 194, no. 17, pp. E601–E607, May 2022, doi: 10.1503/cmaj.212097.
- [4] S. C. Government Of Canada, "Impacts experienced by health care workers during the COVID-19 pandemic, by occupation, Canada, September to November 2021," Jun. 03, 2022. <https://www150.statcan.gc.ca/n1/daily-quotidien/220603/cg-a001-eng.htm> (accessed Mar. 13, 2023).
- [5] Fraser Institute, "Waiting Your Turn: Wait Times for Health Care in Canada, 2022 Report," 2022. [Online]. Available: <https://www.fraserinstitute.org/sites/default/files/waiting-your-turn-2022.pdf>
- [6] E. J. van Beek, S. Mirsadraee, and J. T. Murchison, "Lung cancer screening: Computed tomography or chest radiographs?," *World J Radiol*, vol. 7, no. 8, pp. 189–193, Aug. 2015, doi: 10.4329/wjr.v7.i8.189.
- [7] S. H. Bradley et al., "Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: a systematic review," *Br J Gen Pract*, vol. 69, no. 689, pp. e827–e835, Oct. 2019, doi: 10.3399/bjgp19X706853.
- [8] M. M. Jassim and M. M. Jaber, "Systematic review for lung cancer detection and lung nodule classification: Taxonomy, challenges, and recommendation future works," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 944–964, Jan. 2022, doi: 10.1515/jisys-2022-0062.
- [9] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3462–3471. doi: 10.1109/CVPR.2017.369.
- [10] B. Wang et al., "Automatic creation of annotations for chest radiographs based on the positional information extracted from radiographic image reports," *Computer Methods and Programs in Biomedicine*, vol. 209, p. 106331, Sep. 2021, doi: 10.1016/j.cmpb.2021.106331.
- [11] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification," *Remote Sensing*, vol. 9, no. 8, p. 848, Aug. 2017, doi: 10.3390/rs9080848.
- [12] S. Mukherjee, "The Annotated ResNet-50," *Medium*, Aug. 18, 2022. <https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758> (accessed Mar. 13, 2023).
- [13] M. Ferguson, R. Ak, Y.-T. T. Lee, and K. H. Law, "Automatic localization of casting defects with convolutional neural networks," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec. 2017, pp. 1726–1735. doi: 10.1109/BigData.2017.8258115.
- [14] W. Ausawalaithong, S. Marukatat, A. Thirach, and T. Wilaiprasitporn, "Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach," in *2018 11th Biomedical Engineering International Conference (BMEiCON)*, Nov. 2018, pp. 1–5. doi: 10.1109/BMEiCON.2018.8609997.
- [15] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Comput Biol Med*, vol. 132, p. 104348, May 2021, doi: 10.1016/j.combiomed.2021.104348.
- [16] L. Vogado, F. Araújo, P. S. Neto, J. Almeida, J. M. R. S. Tavares, and R. Veras, "A ensemble methodology for automatic classification of chest X-rays using deep learning," *Computers in Biology and Medicine*, vol. 145, p. 105442, Jun. 2022, doi: 10.1016/j.combiomed.2022.105442.