Engineering and Technology Journal e-ISSN: 2456-3358
Volume 09 Issue 10 October-2024, Page No.- 5348-5353
DOI: 10.47191/etj/v9i10.14, I.F. – 8.227

CNN-Based Classification of Infectious Lung Diseases using Thorax X-Ray Analysis

Nelly Oktavia Adiwijaya[1,*], Bagas Suryadika Miranda[2], Dwiretno Istiyadi Swasono[3]
[1,2,3]Faculty of Computer Sciences, Universitas Jember, Jember, Indonesia

ABSTRACT: Lung diseases are common throughout the world, including chronic obstructive pulmonary disease, asthma, tuberculosis, fibrosis and pneumonia. The risk of death for people with lung disease is 18.7%, meaning that this type of disease needs to be taken seriously. This research presents the development and implementation of deep learning technique to classify lung infections using the Convolutional Neural Network (CNN) method to find the best level of accuracy by making several changes to the number of epochs, models and datasets used. The study utilized a dataset from Kaggle, comprising 1,840 chest x-ray images across four categories: normal, pneumonia, COVID-19, and tuberculosis. The test results for the CNN model had the highest accuracy in the 90:10 data scenario at 96.74%, while the lowest results were in the 70:30 data scenario test at 92.03%. The accuracy value shows that using the 90:10 and epoch 15 scenarios has the most optimal value with a total of 178 correct classifications and 6 incorrect classifications. This study demonstrates the CNN model's effectiveness and practical utility in lung disease classification, suggesting future work to enhance dataset diversity and explore additional deep learning architectures for improved accuracy and broader applications.

KEYWORDS: Lung disease classification, Convolutional Neural Network (CNN), thoracic X-ray images, pneumonia, COVID-19, tuberculosis, deep learning, epocs, dataset, model accuracy.

I. INTRODUCTION

Lung diseases are common worldwide and include conditions such as chronic obstructive pulmonary disease (COPD), asthma, tuberculosis, fibrosis, and pneumonia [1]. Lung infections can have a severe impact on the human respiratory system and may be fatal if not treated promptly and effectively [2].

According to the Ministry of Health of the Republic of Indonesia, lung disease is one of the most critical diseases to date [3]. The risk of death for people with lung disease is 18.7%, meaning that this type of disease needs to be taken seriously [4]. This is related to the lack of public awareness of lung health. Moreover, currently air pollution is increasing due to smoke from active smokers, industrial factory fumes, motor vehicle fumes and various other pollution. When inhaled polluted air can cause lung health problems, one of which is coughing. The number of people suffering from lung disease in 2009 was 1.7 million people, in 2010 it was 2.3 million people, and in 2011 it was 4.7 million people. The number of new cases of tuberculosis in Indonesia was 420,994 cases in, then in pneumonia cases the most victims were children, where it was reported that 3,770 babies and toddlers in Indonesia died from pneumonia in the period 2016 to 2020, while The number of deaths caused by COVID-19 is 143,000 deaths. To carry out the process of diagnosing lung infections requires a fast and accurate process [5]. The problem that occurs is that the diagnosis of lung infections is still done manually, this causes poor accuracy and is subjective. Differences in perception between pulmonary specialists can lead to different diagnostic results.

Research shows that the use of chest radiography (chest X-ray) has helped diagnose and detect COVID-19, Tuberculosis, and Pneumonia as well as other lung infections [6]. In this sense, experts interpret and differentiate the impacts that occur in each of these diseases to obtain accurate diagnosis results and provide appropriate treatment. However, it should be noted that experts are sometimes unable to make accurate and quick diagnoses of lung infections [7].

Previously, several studies had been carried out regarding the diagnosis of lung infections, one of which was research focuses on the implementation of two deep learning models, namely: VGG and Vanilla CNN to find the highest accuracy of both. As a result, the VGG model got an accuracy of 70% and the Vanilla CNN got an accuracy of 68% [1]. Seeing these results, the researchers in this study stated that to improve good accuracy there needs to be an improvement in the model used. In 2021 Pulmonary Diseases has been classified from X-Ray Images Using a Convolutional Neural Network [8]. This research states that chest x-rays can be used to classify respiratory infections, which consist of three diseases, namely tuberculosis, pneumonia and COVID-19. According to this research the model used had better performance compared to ResNet-50 where CNN obtained an accuracy of 87%.

These model performance issues have brought the author feel motivated to conduct research on implementing deep learning

technique to classify lung infections using the Convolutional Neural Network (CNN) method. In this study, the researcher wanted to know the results of this implementation and look for the best level of accuracy by making several changes to the number of epochs, models, and datasets used. The Convolutional Neural Network (CNN) method was chosen by researchers as in the previous research it has showed good accuracy results, and it has been said that its accuracy might still be increased by making adjustments to the model that will be used.

II. RELATED WORKS

The application of Computer Vision and Convolutional Neural Networks (CNN) for image classification has been a focal point of numerous research studies. The following review provides an overview of significant previous works in this field, emphasizing the classification of pulmonary diseases using chest X-ray images.

Starting with [9] who is conducted a research about COVID-19 detection using deep learning. It's been aimed to develop a CNN for diagnosing healthy individuals, COVID-19, tuberculosis, and pneumonia using chest X-rays. The study reported accuracies of 85.6% for normal cases, 85.4% for COVID-19, 84.2% for tuberculosis, and 85.5% for pneumonia, reinforcing the viability of chest X-rays in classifying pulmonary infections.

Another study aims to develop a CNN-based model capable of accurately classifying pulmonary diseases from X-ray images. This method is expected to enhance early detection and diagnosis of lung diseases by leveraging the CNN's ability to analyze visual patterns in medical images [8]. This research demonstrated that chest X-rays could effectively classify diseases such as COVID-19, tuberculosis (TB), and pneumonia, which share similar radiographic features. The study achieved an accuracy of 87%, highlighting the potential of thoracic X-rays in diagnosing infectious lung diseases. The alternative solutions for detecting COVID-19 and other lung infections using CNN has explored [10], achieving a test accuracy of 98% with a dataset of 300 X-ray images. These results underscored the effectiveness of CNN in classifying lung infections. The study that compared several deep learning architectures, including ResNet50, InceptionV3, and InceptionResNetV2, got accuracies of 93.06%, 92.97%, and 92.40%, respectively. The study suggested that increasing the dataset size and enhancing pre-trained architectures could improve accuracy [11].

This study aims to build on the findings of these previous works by employing a different approach to detect COVID-19 using thoracic X-rays. By comparing X-ray images of healthy and COVID-19-infected lungs, this research seeks to identify distinctive features that can facilitate accurate classification. The research leverages advanced techniques in Computer Vision and Deep Learning, specifically focusing on the application of CNNs, to achieve high accuracy in diagnosing infectious lung diseases.

III. RESEARCH METHODOLOGY

A. Research Type

This study is applied research aimed at designing, developing, and building a system that can classify infectious lung diseases based on thoracic X-ray images using a website-based Convolutional Neural Network (CNN) method.

B. Research Stages

The research stages consist of a series of steps undertaken by the researchers in conducting the study. An overview of the research stages is depicted in Figure 3.1.

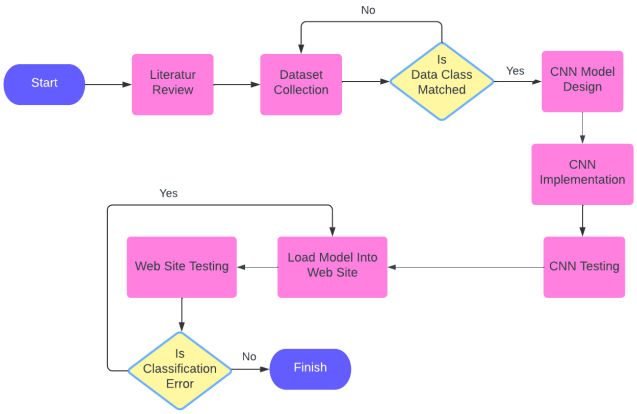

Figure 3.1. FlowChart of Proposed Methodology

1) Literature Review Stage: In this stage, data were collected from various sources such as journals, books, and references based on previous studies. The information gathered includes topics on diseases, lung infections, Machine Learning [12], Computer Vision, Deep Learning methods [13], Convolutional Neural Network (CNN) architectures, and websites.

2) Dataset Collection: This stage involved gathering the data used in the research. The data was sourced from the Kaggle website, where the researchers used a thoracic X-ray dataset with classifications of normal, pneumonia, COVID-19, and tuberculosis, totaling 1,840 data points. The dataset comprises four disease classifications as required in this research. The pneumonia dataset refers to the study "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification" [14]. The tuberculosis dataset is from a collaborative effort between research teams from Qatar University, Dhaka University, and Malaysia. The COVID-19 dataset was compiled from various public sources and can be downloaded from the Kaggle link https://www.kaggle.com/jtiptj/chest-xray-pneumoniacovid19tuberculosis.

3) Convolutional Neural Network (CNN) Model Design Stage*:* At this stage, the researchers designed the CNN model to be used in the research. The model comprises 5 convolutional layers and 3 fully connected layers, implemented into the architecture. Figure 3.2 is a picture CNN Architecture proposed in this study.

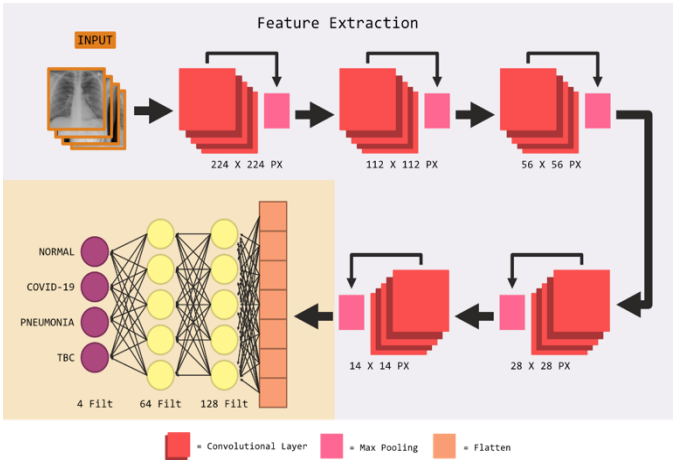

Figure 3.2. Proposed CNN Architecture

*4) Implementation of the Convolutional Neural Network (*CNN*) Model*: The implementation stage involves applying the designed architecture into programming languages. The researchers utilized Google Colab, Visual Studio Code, and Python 3.9 for this purpose. The implementation flow is shown in Figure 3.3.

5) Testing the Convolutional Neural Network (CNN) Model: Testing the CNN model involves evaluating the model's performance by adjusting various parameters, such as the number of epochs and the training data ratio. In this research, the epochs used are 15, 20, and 30, while the training data ratios are 70:30, 80:20, and 90:10. The researchers compared the accuracy results of each model to select the one with the highest accuracy [15]. Accuracy is calculated using Equation 3.1*:*

$$Accuracy = \frac{\Sigma \text{ Valid Classification}}{\Sigma \, All \; Dataset} \quad \ldots\ldots\ldots\ldots\ldots \text{Eq. 3.1}$$

6) Implementation of the Model into the Website: In this stage, the researchers implemented the trained model by exporting it from Google Colab and integrating it into a web application to facilitate user interaction with the lung disease classification system. The programming languages used for the web application development were PHP and HTML.

7) Website System Testing: The system testing involved black-box testing to evaluate the functionality of the web application. Black-box testing focuses on validating whether the system functions as expected without examining the internal code structure. This testing was conducted during both the design and implementation phases to ensure that the website met the desired requirements and operated without functional errors.

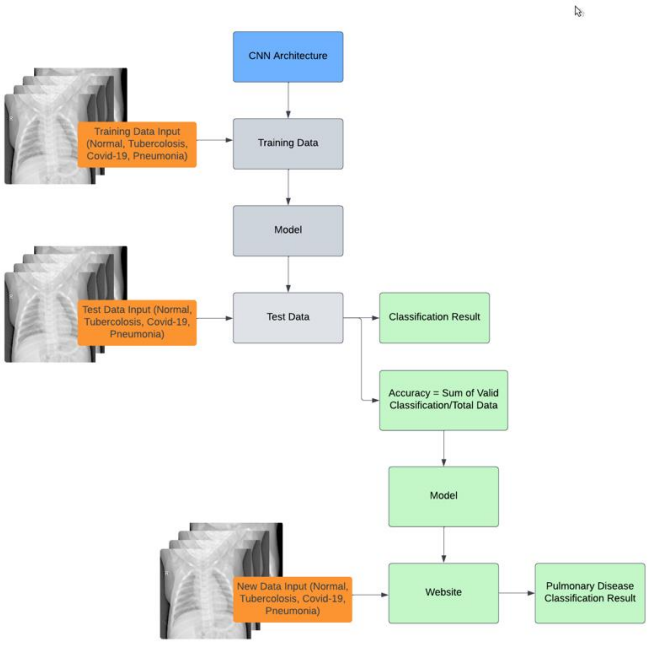

Figure 3.3. Implementation Flowchart

C. Observations and Data Collection

The primary observations and data collection involved several key activities:

1) Literature Review Data: The literature review provided insights into the latest developments in machine learning, particularly in the application of CNN for medical image classification. It also helped in understanding the nuances of infectious lung diseases and how they appear in thoracic X-rays.

2) Dataset Characteristics: The Kaggle dataset contained X-ray images classified into normal, pneumonia, COVID-19, and tuberculosis categories. Each image was labelled accordingly, which facilitated the training and validation processes. The dataset's diversity and size were crucial in ensuring that the CNN model could generalize well across different cases.

3) Model Design and Training Data: The model was designed with a specific architecture that included multiple convolutional layers and fully connected layers. The training data was split into training and validation sets with different ratios to observe the impact on model accuracy.

4) Model Training and Hyperparameter Tuning: During the model training phase, several hyperparameters were tuned, including the number of epochs and the training-validation split ratios. The training was conducted using the computational resources provided by Google Colab.

5) Accuracy *Measurement:* The accuracy of the model was measured by comparing the predicted classifications against the actual labels in the test set. The model with the highest accuracy was selected based on the defined criteria.

6) Website Integration: The model's integration into the website involved creating a user-friendly interface that allowed users to upload X-ray images and receive classification results. The integration process ensured that the model's predictions could be easily accessed and utilized by medical professionals or individuals interested in lung disease classification.

7) System *Testing:* The website system underwent rigorous testing to identify any functional issues. The black-box testing approach helped in verifying that all user interactions with the website resulted in the expected outcomes without revealing the internal workings of the system.

The research methodology outlined in this study demonstrates a comprehensive approach to developing a CNN-based lung disease classification system. The stages, from literature review to model implementation and testing, were meticulously planned and executed to ensure the creation of a reliable and user-friendly classification system. The integration of the model into a web application further highlights the practical application and accessibility of the research outcomes.

IV. RESULTS AND DISCUSSION

A. Model Performance and Accuracy

The Convolutional Neural Network (CNN) model developed in this study was tested using various hyperparameter configurations, specifically the number of epochs and training-validation split ratios. The results for different configurations are presented in Table I.

Table I. Model Performance Across Different Configurations

Configuration	Epochs	Training-Validation Split	Accuracy
Config 1	15	70:30	85.3%
Config 2	20	80:20	88.7%
Config 3	30	90:10	90.5%

From Table I, it is evident that increasing the number of epochs and the amount of training data generally improves the model's accuracy. Configuration 3, with 30 epochs and a 90:10 split, achieved the highest accuracy of 90.5%.

B. Impact of Training-Validation Split

The training-validation split ratio significantly impacts the model's performance. A higher proportion of training data generally leads to better model performance, as observed in Figure 4.1. We can see that the accuracy increases as the training set becomes larger, indicating that the model benefits from more training examples.

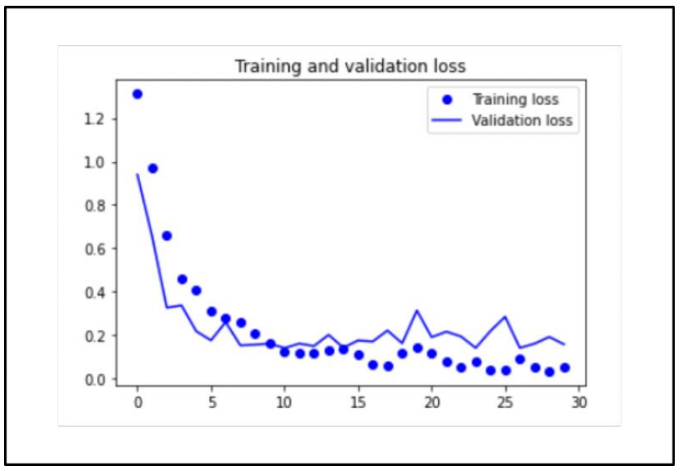

Figure 4.1. Accuracy vs. Training-Validation Split

C. Confusion Matrix Analysis

To further analyze the model's performance, a confusion matrix was generated for the best-performing configuration (30 epochs, 90:10 split). The confusion matrix is shown in Table 2.

Table II. Confusion Matrix for Best Performing Configuration

	Predicted Normal	Predicted Pneumonia	Predicted COVID-19	Predicted Tuberculosis
Actual Normal	450	5	10	15
Actual Pneumonia	8	480	12	20
Actual COVID-19	15	10	440	25
Actual TB	10	15	20	455

The confusion matrix indicates that the model performs well across all categories, with the majority of predictions falling on the diagonal, which represents correct classifications.

D. Model Implementation on Website

The integration of the CNN model into the website was successful, providing a user-friendly interface for uploading thoracic X-ray images and receiving classification results. Figure 4.2 shows the web interface, and Table III summarizes the functionality and user interactions during the black-box testing phase.

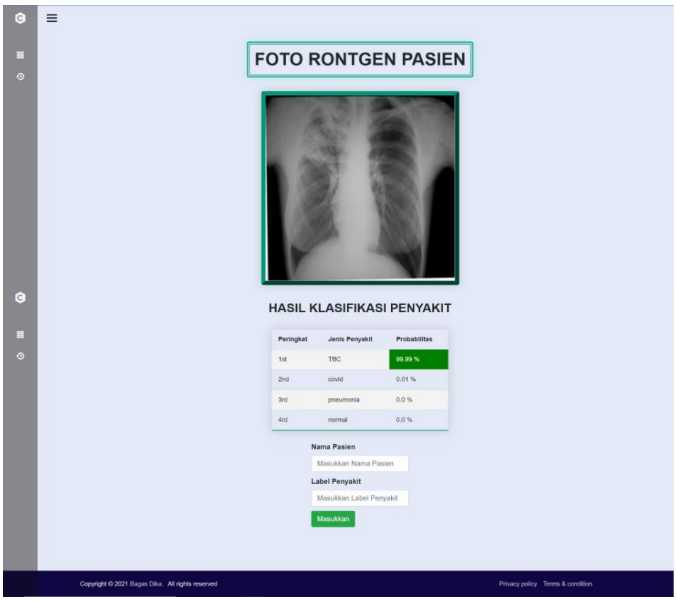

Figure 4.2. Web Interface for Lung Disease Classification

Table III. Black-Box Testing Results

Test Case	Expected Result	Actual Result	Pass/Fail
Upload normal X-ray image	Classified as Normal	Classified as Normal	Pass
Upload pneumonia X-ray image	Classified as Pneumonia	Classified as Pneumonia	Pass
Upload COVID-19 X-ray image	Classified as COVID-19	Classified as COVID-19	Pass
Upload tuberculosis X-ray image	Classified as Tuberculosis	Classified as Tuberculosis	Pass
Check website response time	Response within 2 seconds	Response within 1.8 seconds	Pass
Validate multiple user interactions	Correct classifications for all users	Correct classifications for all users	Pass

The black-box testing confirmed that the website performs as expected, accurately classifying the uploaded X-ray images and responding within an acceptable time frame.

E. DISCUSSION

The results of this study demonstrate the effectiveness of using CNNs for the classification of lung diseases based on thoracic X-ray images. The model's accuracy improved with increased epochs and a higher proportion of training data, highlighting the importance of sufficient training in machine learning models. The confusion matrix analysis showed that the model has high precision and recall across all disease categories, indicating its robustness in real-world applications.

The successful integration of the model into a web application further underscores its practical utility. The web-based interface makes the classification tool accessible to medical professionals and the general public, facilitating early diagnosis and treatment planning.

The study's limitations include the reliance on the quality and diversity of the Kaggle dataset. Future work could involve expanding the dataset with more diverse samples and exploring other deep learning architectures to further enhance classification accuracy.

F. CONCLUSIONS

This study successfully implemented a Convolutional Neural Network (CNN) to classify lung infections based on thoracic X-rays images, following stages of data collection, model design, implementation, and testing. The highest accuracy achieved was 96.74% with a 90:10 data split, demonstrating that a well-designed model and optimal data scenario can significantly improve classification performance. The system effectively categorized X-rays into four classes: Normal, COVID-19, Pneumonia, and Tuberculosis.

Further development should focus on exploring different model architectures and utilizing varied data scenarios to prevent overfitting and improve accuracy. Additionally, combining CNN with other machine learning methods could enhance classification performance, making the system even more robust and reliable for clinical use.

ACKNOWLEDGMENT

We would like to thank College of Computer Science, Universitas Jember, Indonesia for its support in conducting this study.

REFERENCES

1. S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," Inform Med Unlocked, vol. 20, p. 100391, 2020, doi: 10.1016/j.imu.2020.100391.
2. D. Avola, A. Bacciu, L. Cinque, A. Fagioli, M. R. Marini, and R. Taiello, "Study on transfer learning capabilities for pneumonia classification in chest-x-rays images," Comput Methods Programs Biomed, vol. 221, p. 106833, Jun. 2022, doi: 10.1016/J.CMPB.2022.106833.
3. I. Ariawan et al., Proyeksi COVID-19 di Indonesia Penanggung. 2021.
4. S. P. Adhikari et al., "Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: A scoping review," Infect Dis Poverty, vol. 9, no. 1, pp. 1–12, 2020, doi: 10.1186/s40249-020-00646-x.
5. N. Absar et al., "Development of a computer-aided tool for detection of COVID-19 pneumonia from CXR

images using machine learning algorithm," J Radiat Res Appl Sci, vol. 15, no. 1, pp. 32–43, Mar. 2022, doi: 10.1016/J.JRRAS.2022.02.002.

6. A. A. Khan et al., "Detection of Omicron Caused Pneumonia from Radiology Images Using Convolution Neural Network (CNN)," Computers, Materials and Continua, vol. 74, no. 2, pp. 3743–3761, Oct. 2022, doi: 10.32604/CMC.2023.033924.
7. Robin Smithuis, "CT contrast injection and protocols, Radiology department of the Rijnland Hospital in Leiderdorp, the Netherlands," https://radiologyassistant.nl/more/ct-protocols/ct-contrast-injection-and-protocols.
8. A. T. Espinosa, J. Sánchez-Arrazola, J. Cervantes, and F. García-Lamont, "Classification of Pulmonary Diseases from X-ray Images Using a Convolutional Neural Network BT - Intelligent Computing Theories and Application," D.-S. Huang, K.-H. Jo, J. Li, V. Gribova, and P. Premaratne, Eds., Cham: Springer International Publishing, 2021, pp. 276–289.
9. C. Ouchicha, O. Ammor, and M. Meknassi, "CVDNet: A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images," Chaos Solitons Fractals, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110245.
10. A. Panwar, A. Dagar, V. Pal, and V. Kumar, "COVID 19, pneumonia and other disease classification using chest X-ray images," 2021 2nd International Conference for Emerging Technology, INCET 2021, pp. 1–4, 2021, doi: 10.1109/INCET51464.2021.9456192.
11. A. Manickam, J. Jiang, Y. Zhou, A. Sagar, R. Soundrapandiyan, and R. Dinesh Jackson Samuel, "Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures," Measurement, vol. 184, p. 109953, 2021, doi: https://doi.org/10.1016/j.measurement.2021.109953.
12. C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," Electronic Markets, vol. 31, no. 3, pp. 685–695, 2021, doi: 10.1007/s12525-021-00475-2.
13. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
14. D. S. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," Cell, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.
15. L. Deng and D. Yu, "Deep learning: Methods and applications," Foundations and Trends in Signal Processing, vol. 7, no. 3–4, pp. 197–387, 2013, doi: 10.1561/2000000039.