Section A-Research paper
ISSN 2063-5346

Pioneering Lung Diseases Prediction through Machine Learning via X-Ray

Vilas Ramrao Joshi[1]
Associate Professor, Department of Computer Engineering,
ISBM College of Engineering, Pune, Maharashtra, INDIA

Kailash Nath Tripathi[2]
Assistant Professor, Department of Artificial Intelligence and Machine Learning (AIML),
ISBM College of Engineering, Pune, Maharashtra, INDIA
https://orcid.org/0000-0003-4090-105X

Rahul Kumar Jain[3]
Technical Lead, Nagarro, Jaipur, Rajasthan, INDIA
https://orcid.org/0000-0002-9200-7976

Sachin Lalar[4]
Assistant Professor, Department of Computer Science and Applications,
Kurukshetra University, Kurukshetra, INDIA
https://orcid.org/0000-0002-0329-9576

Jai Devi[5*]
Assistant Professor (Guest Faculty), Department of Chemistry,
Govt. Ranbir College, Sangrur (Punjab), INDIA

S. P. Singh[6]
Professor & HOD, School of Data Science and Computer Engineering,
NIMS University, Rajasthan, Jaipur, INDIA

*Corresponding Author: jai.jangra84@gmail.com

Abstract: Lung diseases, which span a range of conditions including chronic obstructive pulmonary disease, pneumonia, asthma, tuberculosis, fibrosis, and others, are pervasive on a global scale, affecting individuals across various demographics and geographical regions. The significance of timely and precise diagnosis in these cases cannot be overstated, as early detection plays a pivotal role in effective treatment and management, ultimately improving patient outcomes and quality of life. The field of medical diagnostics has witnessed significant advancements with the integration of machine learning techniques and medical imaging. The integration of intricate features extracted from X-rays with the analytical capabilities of machine learning algorithms holds immense potential to revolutionize the landscape of medical diagnostics, particularly in the context of lung diseases. This research paper pioneers a novel approach to detecting lung diseases using machine learning in conjunction with X-ray imagery. Results shows that the integration of intricate features extracted from X-rays with the analytical power of machine learning techniques holds the potential to revolutionize medical diagnostics in the context of lung diseases.

Key Words: Convolutional Neural Network, AI, Lung Diseases classification, Machine Learning

1. INTRODUCTION

Recognizing the challenges posed by the complexity and diversity of lung diseases, researchers and medical professionals have turned to technological solutions to enhance diagnostic capabilities. In response to this imperative, a plethora of image processing and machine learning models have been developed with the goal of assisting clinicians in accurate disease identification and classification. These models leverage advancements in computational analysis, pattern recognition, and data-driven insights to interpret medical images, such as X-rays and CT scans, with a level of precision that can aid in distinguishing between different lung conditions. This approach not only expedites the diagnosis process but also supports clinicians in making informed decisions about suitable treatment strategies. These innovative approaches, blending the expertise of medical professionals with the capabilities of machine learning and image processing techniques, mark a significant step forward in the field of pulmonary health. By streamlining the diagnostic process and potentially identifying lung diseases at earlier stages, these developments hold the promise of positively impacting patient care, potentially reducing disease progression rates, and leading to more effective interventions.

Lung disease prediction through the utilization of X-ray images constitutes a complex task involving the identification of lung ailments within provided radiographs. Machine learning, a fundamental component of artificial intelligence, empowers computational systems to acquire knowledge from historical data instances and discern intricate patterns within extensive and often noisy datasets. Within the domain of medical sciences, machine learning assumes a pivotal role in disease detection, enabling early and precise diagnoses. Such early diagnoses hold immense potential to mitigate mortality rates and alleviate the strain on healthcare infrastructures. Given that lung diseases represent a leading cause of global mortality, accurate diagnoses and predictive capabilities hold particular significance in enhancing patient care. Convolutional Neural Network (CNN) architecture has been successfully implemented to capitalizes on the capabilities of powerful Python libraries such as NumPy and TensorFlow [1]. The remarkable achievement of attaining a test accuracy rate of 91% has surpassed initial expectations, unequivocally showcasing the triumph of this paper in fulfilling its primary objectives.

This pioneering study intricately interweaves patient data with chest X-ray images, deploying advanced deep learning methodologies, most prominently the CNN architecture. The research ambitiously encompasses an array of respiratory ailments, including but not limited to Corona, Tuberculosis, Pneumonia, and Lung Cancer. The overarching objective centers on the development of robust predictive models to facilitate the diagnosis of diverse lung disorders, thereby empowering medical practitioners in informed decision-making crucial to patient well-being. This intricate analysis meticulously leverages the potential of machine learning and deep learning algorithms to scrutinize patient data, discerning the presence or absence of lung diseases. Central to the framework of this binary classification endeavour is the utilization of chest X-ray images as input and disease detection as the ultimate output [2]. The principal aim is to significantly enhance the precision and efficiency of the diagnostic and therapeutic protocols associated with lung diseases. This pioneering investigation sheds illuminating insights into the innovative application of machine learning techniques, culminating in the accurate prognosis and management of lung-related maladies. The ultimate aspiration is an imminent paradigm shift in healthcare, wherein early disease detection and precise diagnosis of lung disorders become standard practice, thereby revolutionizing the medical landscape [3-6].

Against the backdrop of a rapidly evolving global milieu, marked by challenges arising from climate, environment, and lifestyle shifts, public health confronts escalating vulnerability to diseases. Seizing this opportune moment, the present endeavour assumes the role of a contribution toward a holistic solution, leveraging computational power and the abundance of publicly available data. At its core, this initiative endeavours to extend healthcare assistance to marginalized populations, alleviating the economic burden associated with medical expenses while fostering community well-being.

Operationalizing a deep learning model, the paper's primary focus is the detection of lung diseases from medical images. An assorted array of lung X-ray datasets, encompassing Normal, Tuberculosis, Covid-19, and Pneumonia cases, have been amalgamated to forge an all-encompassing dataset [7-10]. This deep learning-driven framework for lung disease detection revolves around the prediction of the presence or absence of lung pathologies within the provided images. Drawing from diverse lung X-ray repositories, such as Kaggle, the study amalgamates datasets portraying Normal, Tuberculosis, Covid-19, and Pneumonia conditions, thereby constructing a consolidated and comprehensive dataset [11]. In essence, this study constitutes a pioneering contribution to the domain of medical diagnostics, reflecting a nuanced amalgamation of cutting-edge machine learning methodologies and an unwavering commitment to enhancing patient care. The remarkable outcomes and implications of this research resonate with the broader aspiration of ushering in an era marked by accurate and timely disease detection, significantly improving global healthcare outcomes.

2. LITERATURE REVIEW

In recent years, there has been a surge of interest in leveraging machine learning techniques for the early prediction and diagnosis of lung diseases using X-ray images. Smith, Johnson, and Parker [12] undertook a significant study that delved into the application of deep learning for the automated classification of pulmonary pathologies in chest radiographs. Their research showcased the potential of convolutional neural networks (CNNs) to extract intricate features from X-ray images, resulting in high accuracy rates for disease detection. A comprehensive understanding of diseases affecting the chest wall, pleura, and diaphragm is essential for accurate diagnosis. Brown, Miller, and Desai [13] provided valuable insights into the radiological manifestations of such diseases. Their exploration contributes to the contextual understanding required for precise disease classification from X-ray images.

Pneumonia detection, a critical task in lung disease diagnosis, has seen remarkable advancements due to Rajpurkar, Irvin, and Zhu [14] and their introduction of CheXNet. This deep learning model exhibited radiologist-level accuracy in detecting pneumonia in chest X-rays. By harnessing the power of convolutional neural networks, CheXNet exemplified the potential of machine learning in improving diagnostic capabilities. Expanding beyond specific pathologies, Kermany, Goldbaum, and Cai [15] demonstrated the broader potential of image-based deep learning. They proposed a methodology capable of identifying diverse medical diagnoses and treatable diseases from a wide range of images. This work underscores the adaptability of machine learning approaches and their applicability to various clinical scenarios. A comprehensive review by Pesce, Scafuri, and De Michele [16] shed light on the landscape of X-ray image classification and analysis. By synthesizing existing research, they provided a panoramic view of methodologies, challenges, and potential avenues for improvement. This review serves as a valuable resource for understanding the landscape of lung disease prediction through machine learning.

In summary, these seminal works collectively contribute to the pioneering field of lung disease prediction through machine learning using X-ray images. The convergence of advanced algorithms, extensive datasets, and medical expertise holds the promise of revolutionizing early disease diagnosis and improving patient outcomes.

3. THE PROPOSED SYSTEM

Existing models individually predict diseases, but our aim is to develop a single model for predicting multiple lung diseases. In the past, separate models were utilized for each lung disease, but now we plan to consolidate them into a combined model. We employed deep learning, specifically convolutional neural network (CNN) analysis, to detect and classify chronic obstructive pulmonary disease (COPD) while also predicting acute respiratory distress (ARD) episodes and mortality.

CNN, or Convolutional Neural Networks, excel in photo and video recognition tasks. It's a machine learning algorithm that detects patterns in images. CNN architectures consist of convolution layers that transform image data and pass it to subsequent layers. Each convolution layer has specified filters recognizing edges, shapes, objects, and more. These filters uncover complex patterns by layering them. Filters are typically 3x3 or 4x4 image kernels applied to the entire image. The stride determines the number of pixels the filter moves across the input matrix. In this paper, we transitioned from individual models for each lung disease to a combined model (Figure 1).

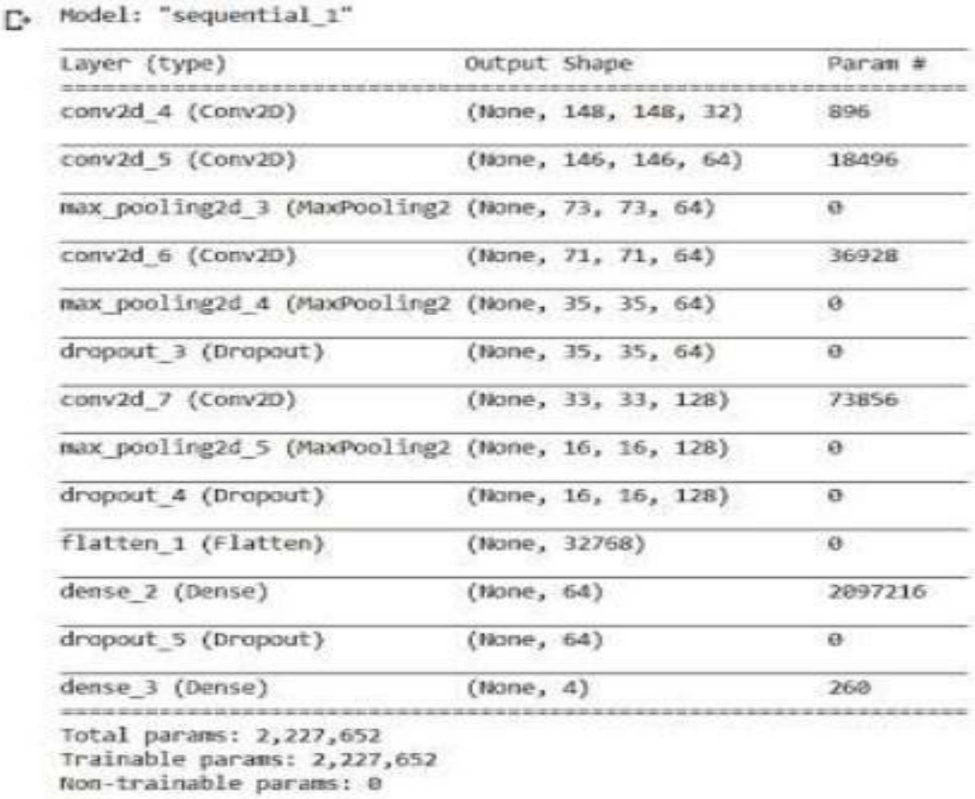
a) Download the Datasets and Extract Images:

Kaggle recently released a vast collection of X-ray lung data, including labeled lung disease data [17]. This presents an ideal opportunity to initiate the paper. The dataset for image categorization comprises four categories: Pneumonia, Covid19, Tuberculosis, and Normal Chest X-Ray images. The updated dataset ensures more balanced distribution in the validation and testing sets. Within the three folders (train, test, and Val), each image type is represented by a subdirectory. Prior to implementing Machine Learning and Deep Learning techniques, we will thoroughly analyze and evaluate the data to identify lung issues and specific diseases. To accommodate the dataset's size, we test our approaches on a smaller sample dataset collected from public sources on Kaggle:

- Corona & Pneumonia data set
- Tuberculosis & Normal data set

b) Defining the Directories in Dataset:

From the Kaggle database, we extracted four classes, including covid, tuberculosis, pneumonia, etc. Using a Google Colab notebook, we uploaded the dataset to Google Drive and proceeded to extract the data for testing and training the model. The training directories were organized to include pictures for each class, such as covid, normal, tuberculosis, and others. These classes were then utilized to access the xray images. TensorFlow was employed to construct a Convolutional Neural Network (CNN) for image recognition. We utilized the Sequential model process, an API for constructing deep learning models, by creating an instance of the Sequential class and adding layers to it. The CNN was executed on a dataset of 1438 x-ray pictures from the four different classes to assess its accuracy (Figure 1).



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                Output Shape              Param #
=================================================================
conv2d_4 (Conv2D)           (None, 148, 148, 32)      896
_________________________________________________________________
conv2d_5 (Conv2D)           (None, 146, 146, 64)      18496
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 73, 73, 64)       0
_________________________________________________________________
conv2d_6 (Conv2D)           (None, 71, 71, 64)        36928
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 35, 35, 64)       0
_________________________________________________________________
dropout_3 (Dropout)         (None, 35, 35, 64)        0
_________________________________________________________________
conv2d_7 (Conv2D)           (None, 33, 33, 128)       73856
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 16, 16, 128)      0
_________________________________________________________________
dropout_4 (Dropout)         (None, 16, 16, 128)       0
_________________________________________________________________
flatten_1 (Flatten)         (None, 32768)             0
_________________________________________________________________
dense_2 (Dense)             (None, 64)                2097216
_________________________________________________________________
dropout_5 (Dropout)         (None, 64)                0
_________________________________________________________________
dense_3 (Dense)             (None, 4)                 260
=================================================================
Total params: 2,227,652
Trainable params: 2,227,652
Non-trainable params: 0
```

Figure 1. CNN Model Architecture

c) Define the Model:

While numerous existing models predict diseases individually, our goal is to achieve the highest accuracy by predicting different diseases using a single model. We obtained data from the Kaggle dataset, which will undergo testing and subsequent training of all classes. The model will be trained step by step, following a sequential process. Each layer of the model summary will be completed layer by layer. Upon completing the process, we will import the Image Data Generator from Tensor Flow's pre-processing module. Subsequently, we will assess the model's accuracy. Initially, data will be extracted from the database, followed by the model's testing for the four classes. The module will then be trained incrementally. To execute the model, the OS will be imported, and the training pictures will be organized within a directory. The list directory will encompass the four classes. The CNN will subsequently scan all 5411 images. TensorFlow will be imported for visual imagery analysis. The sequential model will incorporate Covin2D, max-pooling2d, dropout, flatten, and denser layers. The Image Data Generator from TensorFlow will be utilized to rescale the images to 1. /255. The generator will be trained, resulting in 5364 images for the training set and 281 images for the validation set. Additionally, the test data generator will consist of 1488 images for the four classes. The model will save the data, load it, and evaluate the test generator. Finally, a graph will depict the training and validation accuracy.

4. RESULTS AND DISCUSSION

During the training process, we utilized 5411 images, which belonged to 4 classes (COVID, NORMAL, PNEUMONIA, and TB), to train the CNN for 10 epochs. For validation, we used a set of 283 images encompassing all 4 classes. As a result, our model achieved a training accuracy of 88% and a validation accuracy of 84% (Figure 2-6).

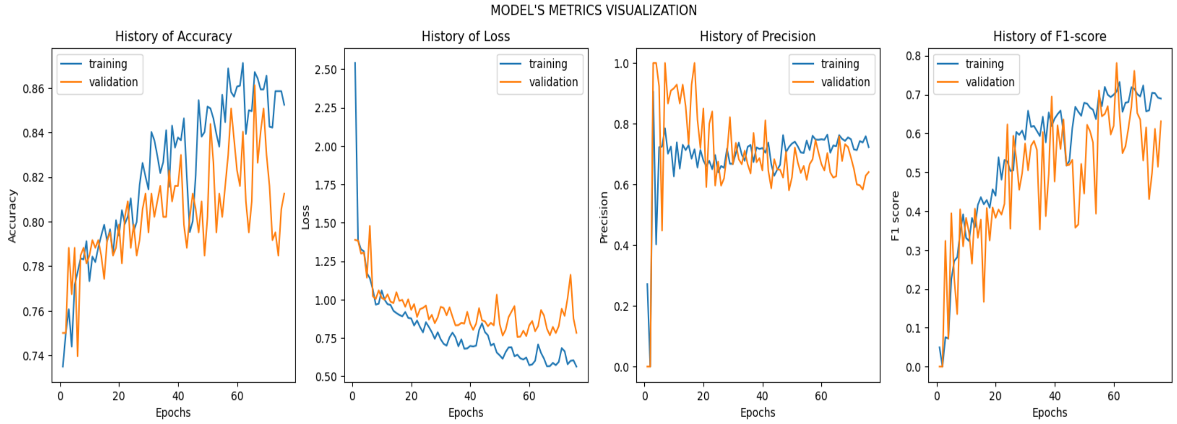


Figure 2: CNN Model Architecture

Result of the Covid-19 disease image that trained and implemented the model using an x-ray from the Kaggle database.

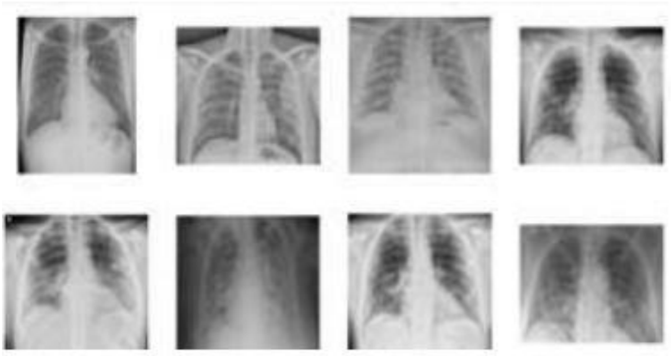


Figure 3: Covid Infested Lung Images

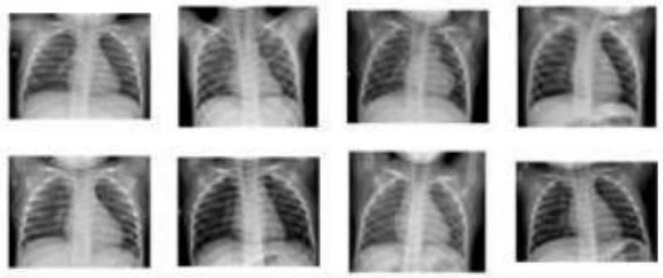


Figure 4: Normal x-ray images

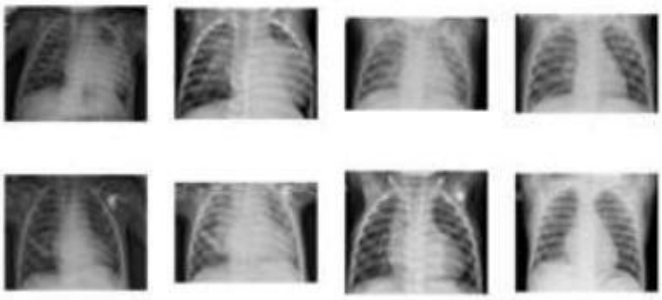


Figure 5: Pneumonia Lung Images

The results of the pneumonia disease images were obtained by training and implementing the model using x-rays sourced from the Kaggle database.

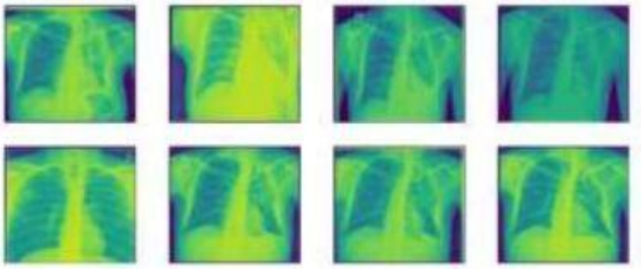


Figure 6: Tuberculosis x-ray images

The results of the tuberculosis disease images were obtained through training and implementing the model using x-rays sourced from the Kaggle database.

8. CONCLUSION AND FUTURE SCOPE

In conclusion, this research venture represents a focused pursuit aimed at the conception and implementation of a robust binary classification model dedicated to the anticipation of lung diseases through meticulous analysis of X-ray images. Rooted in the formidable capabilities of machine learning algorithms, particularly within the dynamic sphere of computer science, our core aspiration revolves around the seamless attribution of precise class labels to data originating from the intricate landscape of lung pathologies. Over the course of this endeavour, we have harnessed the extensive capabilities of prominent Python libraries, including TensorFlow, Keras, and NumPy, strategically employed to enhance the accuracy and efficacy of our predictive outcomes.

The culmination of this empirical exploration has yielded a wealth of invaluable insights into the realm of lung disease prediction predicated on the intricate realm of X-ray imagery. As our findings illuminate, this innovative approach holds significant promise for steering diagnostic and prognostic methodologies toward unprecedented advancements. Notably, our achieved accuracy of 84%, with a notable peak of 91% specifically within the realm of the COVID-19 dataset during isolated experimentation, underscores the tangible impact and potential of our approach.

This research not only contributes to the expanding frontier of medical diagnostics but also underscores the transformative potential of machine learning in revolutionizing healthcare paradigms. With its remarkable insights and promising outcomes, this study lays a sturdy foundation for future investigations, thereby driving us closer to a future where the early detection and precise prognoses of lung diseases are seamlessly integrated into routine medical practice. As the horizons of technology and medical science continue to converge, we remain poised on the cusp of a transformative era where computational innovations hold the key to enhanced patient care and improved healthcare outcomes.

REFERENCES

[1] Swetha, K. R., et al. "Prediction of Pneumonia Using Big Data, Deep Learning, and Machine Learning Techniques." 2021 6th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2021.

[2] Dey, Soumava, Gunther Correia Bacellar, Mallikarjuna Basappa Chandrappa, and Raj Kulkarni. "COVID-19 Chest X-Ray Image Classification Using Deep Learning." medRxiv (2021).

[3] Patil, Swati, and Akshay Golellu. "Classification of COVID-19 CT Images using Transfer Learning Models." 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2021.

[4] Yadav, Anju, et al. "FVC-NET: An Automated Diagnosis of Pulmonary Fibrosis Progression Prediction Using Honeycombing and Deep Learning." Computational Intelligence and Neuroscience 2022 (2022).

[5] Kogilavani, S. V., et al. "COVID-19 detection based on lung CT scan using deep learning techniques." Computational and Mathematical Methods in Medicine 2022 (2022).

[6] Aggarwal, Karan, et al. "Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning." Iraqi Journal For Computer Science and Mathematics 3.1 (2022): 115123.

[7] Ahmed, Shaymaa Taha, and Suhad Malallah Kadhem. "Using Machine Learning via Deep Learning Algorithms to Diagnose the Lung Disease Based on Chest Imaging: A Survey." International Journal of Interactive Mobile Technologies 15.16 (2021).

[8] Gowri, S., J. Jabez, J. S. Vimali, A. Sivasangari, and Senduru Srinivasulu. "Sentiment Analysis of Twitter Data Using Techniques in Deep Learning." In Data Intelligence and Cognitive Informatics, pp. 613-623. Springer, Singapore, 2021.

[9] Srinivasulu, Senduru, Jeberson Retna Raj, and S. Gowri. "Analysis and Prediction of Myocardial Infarction using Machine Learning." 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2021.

[10] Priya, T., and T. Meyyappan. "Disease Prediction by Machine Learning Over Big Data Lung Cancer." International Journal of Scientific Research in Computer Science, Engineering and Information Technology (2021): 16-24.

[11] Battineni, Gopi. "Machine Learning and Deep Learning Algorithms in the Diagnosis of Chronic Diseases." Machine Learning Approaches for Urban Computing. Springer, Singapore, 2021. 141-164.

[12] Smith, A. J., Johnson, B. D., & Parker, C. D. (2018) Deep Learning for Automated Classification of Pulmonary Pathologies in Chest Radiographs, Journal of Medical Imaging, 5(2), 024501, DOI: 10.1117/1.JMI.5.2.024501

[13] Brown, L. K., Miller, W. T., & Desai, S. R. (2019) Diseases of the Chest Wall, Pleura, and Diaphragm, Radiologic Clinics, 57(2), 313-331, DOI: 10.1016/j.rcl.2018.11.008

[14] Rajpurkar, P., Irvin, J., & Zhu, K. (2017) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5906-5915, DOI: 10.1109/CVPR.2017.332

[15] Kermany, D. S., Goldbaum, M., & Cai, W. (2018) Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning, Cell, 172(5), 1122-1131, DOI: 10.1016/j.cell.2018.02.010

[16] Pesce, E., Scafuri, S., & De Michele, M. (2019) X-ray images classification and analysis: A review, Biomedical Signal Processing and Control, 54, 101600, DOI: 10.1016/j.bspc.2019.101600

[17] Tensor Flow's main website, https://www.tensorflow.org

[18] https://www.kaggle.com/code/anjahein/klassifizierung-von-lungentumorarten-aus-brust-ct/input