BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 13, No. 1, March 2025
Research Article

A Comparative Analysis of Chest X-ray Examination with AI Enhancement Using XAI Techniques

Cem Ozkurt

***Abstract*— Chest X-ray analysis plays a vital role in diagnosing pneumonia, and recent advancements in Deep Learning (DL) methods have significantly improved the accuracy of automated diagnosis. This study explores the intersection of DL and explainable artificial intelligence (XAI) in the context of pneumonia diagnosis through chest X-rays. The dataset used in this study consists of 1,341 training images of healthy individuals and 3,875 images of pneumonia cases, with the test set comprising 234 healthy and 390 pneumonia cases. Additionally, the validation set includes 8 images for both categories. This diversity aims to enhance the model's ability to generalize across different scenarios. The Convolutional Neural Network (CNN) and Transfer Learning (TL) methods utilizing the ResNet50 model achieved accuracies of 95.23 and 96.67, respectively. Subsequently, the models were explained using XAI methods such as SHAP and Grad-CAM. The study concludes by highlighting the potential of DL and XAI to enhance the interpretability and reliability of pneumonia diagnoses through chest X-ray analysis, aiming to contribute to future research in this field.**

Index Terms— **Chest X-ray analysis, Convolutional Neural Networks (CNNs), Deep Learning (DL) methods, Explainable AI (XAI), Grad-CAM (Gradient-weighted Class Activation Mapping).**

I. Introduction

TODAY, ARTIFICIAL intelligence (AI) technologies play a crucial role in the diagnosis and treatment of critical health issues such as pneumonia [1]. However, concerns regarding the safety, transparency, and comprehensibility of these technologies in clinical applications are significant. The primary objective of this study is to evaluate the role of AI models in pneumonia diagnosis and treatment, with a particular emphasis on the potential utilization of Explainable Artificial Intelligence (XAI) methods [2].

With the widespread use of AI technologies in the healthcare sector, the understandability and traceability of decision-making processes are becoming increasingly important. In this context, understanding the role of explainable AI models in pneumonia diagnosis and treatment is of critical importance. This study aims to highlight the potential of XAI in ensuring the effective use of AI technologies in pneumonia diagnosis.

Cem Ozkurt, is with Department of Computer Engineering at Sakarya University of Applied Sciences is located in Sakarya, Turkey (cemozkurt@subu.edu.tr).

https://orcid.org/0000-0002-1251-7715

Manuscript received Marc 7, 2024; accepted Feb 25, 2025.
DOI: 10.17694/bajece.1448546

Deep learning (DL) methods, especially Convolutional Neural Networks (CNNs), have become the cornerstone of pneumonia diagnosis due to their remarkable ability to process complex visual data such as chest X-rays (CXR). CNNs are inspired by the human brain's visual processing mechanisms and are known for their proficiency in extracting hierarchical features from images. This adaptability allows CNNs to excel in medical image analysis, identifying subtle visual patterns critical to diagnosing pneumonia accurately [3]. Various layers within CNN architectures, including convolutional, pooling, and fully connected layers, contribute to this capability by analyzing features ranging from basic edges to more intricate structures indicative of disease. The use of CNNs in the healthcare field has extended far beyond pneumonia diagnosis, as they have proven their effectiveness in diverse domains, from medical imaging to wireless resource allocation [4,5].

Therefore, this article presents an analysis to evaluate the current status and future potential of AI models in pneumonia diagnosis and treatment, with particular emphasis on the role of CNNs and XAI techniques. A review of significant studies in the literature will be conducted to compile existing knowledge on how explainable AI methods can be utilized in pneumonia diagnosis and treatment to guide future research efforts. The findings of this study could contribute significantly to enhancing the effectiveness of AI technologies in pneumonia diagnosis and ensuring reliability in clinical applications.

Pneumonia is an inflammation of the lung parenchyma caused by infectious microorganisms and non-infective agents. It can affect all age groups but is particularly severe in fragile populations such as children and the elderly. Early and accurate detection of pneumonia is crucial to prevent fatal outcomes.Recent advancements in deep learning (DL) methods have significantly improved the accuracy of automated pneumonia diagnosis through chest X-rays (CXR).

Yang et al. proposed a deep learning approach that considers the background factors of lung X-ray images to improve pneumonia identification accuracy. Using VGG16, they achieved an accuracy of 95.6 and emphasized the importance of considering background factors in the diagnostic process while using Grad-CAM to highlight model explainability [6].

De Moura et al. utilized SHAP and Grad-CAM to differentiate chest X-ray images of COVID-19-based pneumonia from other lung patterns. Their approach achieved

an accuracy of 82 with the XGBoost model, underscoring the importance of explainable AI in distinguishing COVID-19 pneumonia from other types [7].

Ren et al. explored an interpretable approach combining deep learning with Bayesian Networks, achieving high performance in pneumonia detection using a dataset of 35,389 cases. This study emphasized the necessity of interpretability in AI models for clinical applications [8].

Zou et al. presented an ensemble AI explainability method combining SHAP and Grad-CAM++ to provide visual explanations for a deep learning prognostic model predicting the mortality risk of pneumonia patients. Their method showed high trust and localization effectiveness among radiologists, demonstrating the value of explainability in clinical decision-making [9].

Stephen et al. developed a convolutional neural network from scratch for pneumonia classification, achieving significant validation accuracy through data augmentation techniques. Their work addressed the challenges of reliability and interpretability in medical imagery by using a large, well-augmented dataset [10].

Alsharif et al. introduced PneumoniaNet, a novel CNN-based framework for automated detection and classification of pediatric pneumonia, achieving an accuracy of 99.7. Their model distinguished between viral, bacterial, and normal cases, demonstrating the potential of deep learning in improving diagnostic accuracy, especially in remote areas lacking expert radiologists [11].

Varshni et al. evaluated the functionality of pre-trained CNN models for pneumonia detection, highlighting the effectiveness of using these models as feature extractors in conjunction with supervised classifiers. Their results indicated that pre-trained CNN models are highly beneficial for analyzing CXR images [12].

These studies collectively demonstrate the advancements and applications of deep learning and explainable AI in pneumonia diagnosis. The integration of SHAP and Grad-CAM provides comprehensive insights into model decision-making, which is critical for clinical acceptance and reliability. Our study builds upon these findings by employing a combination of ResNet50, SHAP, and Grad-CAM to enhance the interpretability and accuracy of pneumonia diagnosis models.

Table 1 includes a literature review table that outlines the dataset, methods (architectures and XAI techniques), and results of various related studies, providing a comprehensive overview of the field:

TABLE I
LITERATURE REVİEW TABLE

	SHAP	Grad-Cam	LIME	DL	TL	Chest Xray
[3]	-	✓	-	-	✓	-
[4]	✓	-	-	✓	-	✓
[5]	✓	-	-	✓	-	✓
[6]	✓	✓	✓	-	✓	-
[7]	-	-	-	✓	-	✓
[8]	-	-	-	✓	-	✓
[9]	✓	-	-	✓	✓	✓
Ours	✓	✓	-	✓	✓	✓

II. Materials And Methodology

A. DataSet

The dataset used in this study consists of chest X-ray images, focusing on the diagnosis of pneumonia. The test set includes 234 images of healthy individuals and 390 images of pneumonia cases. The training set comprises 1,341 images of healthy subjects and 3,875 images of pneumonia cases. Additionally, the validation set includes 8 images for both healthy and pneumonia cases. All images are in X-ray format, capturing various aspects of chest conditions. The diversity in the dataset aims to enhance the model's ability to generalize across different scenarios. Figure 1 provides a visual representation of selected images from the dataset. The composition of the dataset serves as a critical foundation for training, validating, and testing the models in subsequent phases of our methodology. Figure 1 illustrates selected examples from the dataset, providing a glimpse into the diversity of chest X-ray images used in this study. The dataset was obtained from https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.

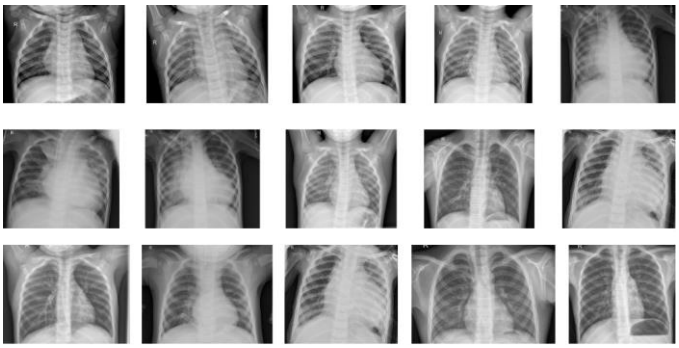

Fig.1. Pneumonia Chest X-ray Database dataset examples

B. DL Methods

Deep Learning (DL) is a subfield of machine learning known for its automatic learning capability, particularly in complex datasets. Emphasizing the inclusion of multi-layered neural networks and the capacity to learn from extensive datasets, deep learning highlights its ability to successfully accomplish complex tasks. Among these methods,CNNs stand out as a significant component extensively used in areas involving visual data, such as image recognition and classification. In Transfer Learning, the ResNet50 model has been used. Transfer Learning is another method within this framework, allowing the adaptation of learned general features for another task.

1) Convolutional Neural Networks (CNN)

CNNs form the core of our pneumonia diagnosis methodology, leveraging their ability to comprehend complex visual information [13,14,15]. Inspired by the visual processing mechanisms of the human brain, CNNs demonstrate remarkable proficiency in image analysis tasks [16,17]. In the context of chest X-ray analysis, CNNs excel at capturing intricate patterns and nuanced features crucial for accurate

diagnosis [3]. The architecture of our CNN model, illustrated in Figure 2, consists of multiple layers, each contributing to the network's capacity to understand and interpret hierarchical visual information [16]. The convolutional layers, essential for feature extraction, utilize filters to detect hierarchical features, progressing from basic edges and textures to more intricate structures indicative of pneumonia [15].

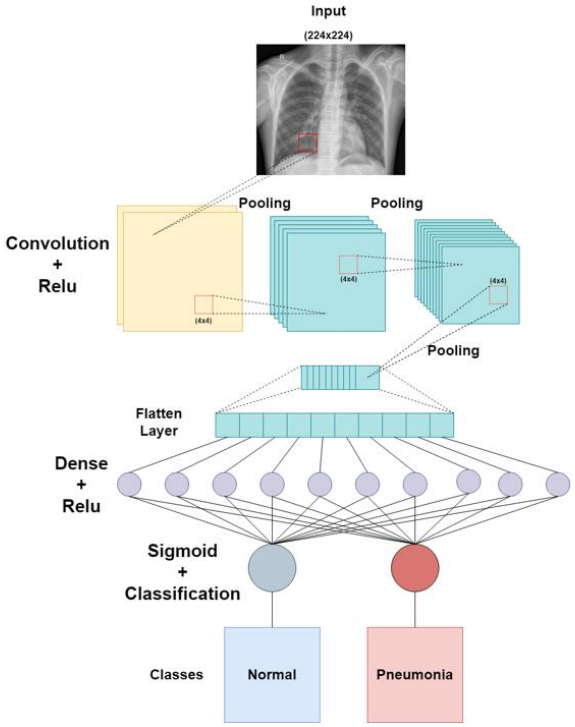

Fig.2. Used Convolutional Neural Network (CNN) Model

In this study, we adopt a CNN architecture with carefully designed layers, including convolutional layers for feature extraction, pooling layers for spatial down-sampling, and fully connected layers for decision-making [16,14]. These layers collectively contribute to the network's capability to comprehend chest X-ray images and make accurate diagnostic predictions. Our choice of leveraging CNNs is rooted in their proven efficacy in various domains, from medical image analysis [4,5] to wireless resource allocation [3] and beyond. The adaptability and versatility of CNNs make them a valuable tool in addressing the complexities associated with pneumonia diagnosis, where identifying subtle visual cues is paramount [35]. The inherent hierarchical feature learning capabilities of CNNs allow them to discern patterns in medical images, making them particularly well-suited for tasks requiring nuanced understanding, such as the diagnosis of pneumonia from chest X-ray images [4]. This adaptability is further enhanced by fine-tuning the pre-trained models on specific medical datasets, optimizing the network for the intricacies of pneumonia diagnosis [16,4].

This study utilized a Convolutional Neural Network (CNN) model that was carefully structured by adjusting several hyperparameters. In the first two Conv2D layers, 32 and 64 filters were used, respectively, with a filter size of 3×3. Each Conv2D layer was employed to extract specialized features, while MaxPooling2D layers were applied to reduce spatial dimensions. For instance, the first MaxPooling2D layer reduced the output dimensions to 111×111×32. All these layers utilized the ReLU activation function, allowing the model to learn non-linear relationships. Additionally, the final stages of the model included fully connected (Dense) layers, which were adjusted to enable complex decision-making within the CNN. The Flatten layer transformed the output of the convolutional layers into a single vector, providing the necessary structure for classification. This model, which includes dense layers with a large number of parameters (e.g., 22,151,424 parameters in the dense1 layer), exhibits a strong learning capacity and delivers effective results in high-dimensional data processing.

A Convolutional Neural Network (CNN) model was trained on the dataset, and the layers of the CNN model are provided in Table 2.

TABLE II
CONVOLUTIONAL NEURAL NETWORK (CNN) LAYERS.

Layer(type)	Output Shape	Param
conv3d (Conv2D)	None, 222, 222, 32)	896
max_pooling2d (MaxPooling2D	(None, 111, 111, 32)	0
conv2d_1 (Conv2D)	(None, 109, 109, 64)	18496
max_pooling2d_1	(None, 54, 54, 64)	0
conv2d_2 (Conv2D)	(None, 52, 52, 128)	73856
max_pooling2d_2	(None, 26, 26, 128)	0
flatten (Flatten)	(None, 86528)	0
dense (Dense)	(None, 256)	22151 424
dense_1 (Dense)	(None, 1)	257

2) *Transfer Learning in Pneumonia Diagnosis*

Transfer Learning (TL) stands as a pivotal component in our methodology, facilitating the seamless adaptation of knowledge acquired from pre-training on a broader dataset to the specific task of pneumonia diagnosis. In our study, we adopted the ResNet50 model, pre-trained on a vast array of images encompassing diverse categories [18,19,20,21,22]. The weights obtained during the training of ResNet50 were then transferred to our pneumonia diagnosis model, serving as a foundational starting point. Transfer Learning is instrumental in addressing challenges associated with limited datasets specific to a particular medical domain [23,24,25,26,27].

By leveraging the learned features from ResNet50, our model gains a robust understanding of general image patterns and structures, significantly enhancing its ability to recognize relevant features in chest X-ray images. The process involves fine-tuning the pre-trained model on our pneumonia dataset, allowing the model to adapt its learned features to the intricacies of pneumonia diagnosis. This strategic integration not only accelerates the training process but also promotes better convergence and performance on our specific task. Transfer Learning, with its ability to transfer knowledge across domains, proves particularly beneficial in medical image analysis, where labeled datasets are often limited [18,19,20,21,22]. Our approach showcases the effectiveness of transferring pre-learned features, emphasizing the adaptability and enhanced performance that Transfer Learning brings to the realm of pneumonia diagnosis.

C. DataSet

In recent years, Explainable AI (XAI) has emerged as a crucial tool to address the interpretability challenges posed by complex machine learning models [27]. Particularly in critical domains like medical image analysis, where transparency is paramount, XAI aims to demystify the decision-making processes of these models. Our methodology places a strong emphasis on XAI principles, leveraging SHAP (SHapley Additive exPlanations) for enhanced interpretability [28]. Figure 3 provides a visual representation illustrating the working mechanism of explainable Artificial Intelligence (XAI) methods, showcasing the transparency and interpretability aspects integrated into our approach.

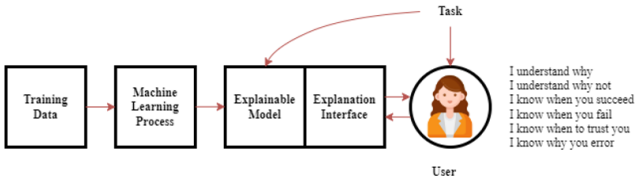

Fig.3. Provides a visual overview of explainable Artificial Intelligence (XAI) methods, illustrating their working mechanisms.

1) SHAP (SHapley Additive exPlanations)

SHAP, rooted in cooperative game theory, offers a principled way to allocate contributions among features for each prediction made by a model. In the context of pneumonia diagnosis, SHAP facilitates the identification and visualization of critical regions within chest X-ray images that heavily influence the model's classification decision. The SHAP value ϕ for feature prediction is mathematically defined as seen in Equation 1:

$$\phi_g^j(f) = \sum s \subseteq \{x^1, \dots, x^p\}\{x^j\} \frac{|S|!(p-|S|-1)!}{P!} \Big(g(f, S \cup \{x^i\}, \Omega) - g(f, S, \Omega) \Big) \quad (1)$$

This equation serves as a foundational tool in our methodology, enabling the systematic evaluation of each feature's impact on the model's predictions. Here, $\phi_g^j(f)$ represents the SHAP value for feature j in a given prediction. The summation across subsets S involves the consideration of different combinations of features, and the coefficients $\frac{|S|!(p-|S|-1)!}{P!}$ balance the combinatorial interactions. The terms $g(f, S \cup \{x^i\}, \Omega)$ and $g(f, S, \Omega)$ signify the model's predictions when including and excluding feature i, respectively.

This integration provides valuable insights into the regions of chest X-ray images that contribute most to the final classification decision, ranging from basic edges and textures in the early layers to more complex structures indicative of pneumonia in the deeper layers.

By visualizing SHAP values, we gain insights into regions of X-ray images that are crucial for explaining the model's decision-making process. Our approach utilizes masking techniques, such as the "inpaint telea" method, to identify specific areas of interest, facilitating a comprehensive understanding of the model's interpretability [28,29].

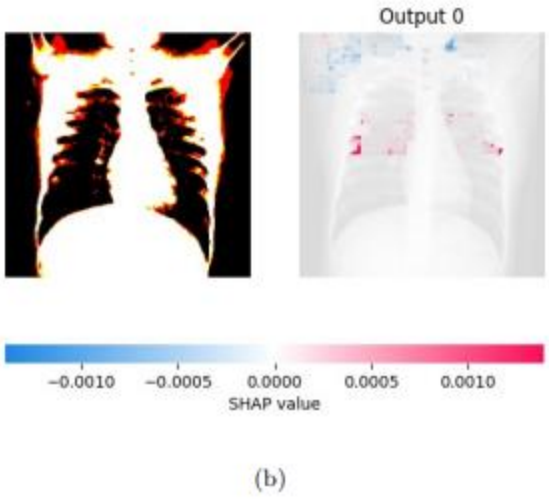

Fig.4. SHAP interpretation of randomly selected image from the dataset.

2) Grad - CAM (Gradient-weighted Class Activation Mapping)

In our exploration of Explainable AI (XAI) methodologies, we incorporated Grad-CAM (Gradient-weighted Class Activation Mapping) [20] alongside SHAP. This technique plays a pivotal role in unveiling the decision-making processes of CNNs, particularly in the realm of medical image analysis.

Grad-CAM provides valuable visual insights into the influential regions of an input image affecting the model's final classification decision. This is achieved through the computation of gradients ∇ of target class scores Y^c with respect to the feature maps A^k of the final convolutional layer.

The resulting gradient-weighted activation maps $L^c_{Grad-CAM}$ are obtained using global average pooling, as expressed in Equation 2:

$$L^c_{Grad-CAM} = \sum_{k=1}^{K} w_k^c \; . \; ReLU\left(\frac{\partial Y^c}{\partial A^k}\right) . \; F^k(x, y) \quad (2)$$

Here, w represents the importance weight associated with the feature maps. This weight is determined by summing the gradients with respect to the corresponding feature maps, normalized by Z as seen in Equation 3:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} . ReLU\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \quad (3)$$

Furthermore, the calculation of the gradient $\frac{\partial Y^c}{\partial A_{ij}^k}$ involves the weighted summation of gradient values across spatial dimensions, denoted by α_{kcij} as seen in Equation 4:

$$\frac{\partial Y^c}{\partial A_{ij}^k} = \frac{1}{Z}\sum_{i'}\sum_{j'} \alpha_{kcij}.ReLU\left(\frac{\partial Y^c}{\partial A_{i'j'k}}\right) \quad (4)$$

These mathematical formulations contribute to a comprehensive understanding of Grad-CAM's mechanism, elucidating the significance of each feature map in influencing the final decision. In our pneumonia diagnosis framework, Grad-CAM proved instrumental in uncovering the specific regions within chest X-ray images that played a critical role in the CNN's classification decisions.

By applying Grad-CAM to various layers of the network, ranging from top to mid layers, we gained insights into the hierarchical features learned by the model. The transparency introduced by Grad-CAM enhances the interpretability of our pneumonia diagnosis model, which is crucial in medical applications for fostering trust among healthcare practitioners.

This integration aligns with the current trend of leveraging XAI techniques to bridge the gap between complex model architectures and interpretability, promoting the responsible and ethical deployment of AI, especially in critical domains such as healthcare. Our work is inspired by prior research successfully applying XAI techniques in medical image analysis [30,31,32,33,34], emphasizing the significance of transparent and interpretable models in AI-based medical diagnosis systems.

III. Materials And Methodology

In the context of this study, the dataset contains chest X-ray images focusing on pneumonia diagnosis. This versatile research employs an advanced approach that combines the Transfer Learning model ResNet50 with SHAP (SHapley Additive exPlanations) and the CNN model with Grad-CAM (Gradient-weighted Class Activation Mapping).

The confusion matrix is a matrix used, particularly, to evaluate the performance of a classification model, focusing on comparing the model's predictions with the actual classes. This matrix includes True Positive (TP) and True Negative (TN) values, representing cases where the model accurately predicts Pneumonia classes, along with False Positive (FP) and False Negative (FN) predictions. Each cell represents a combination of the true class and the predicted class. This visual is used to understand which classes the model predicted correctly and in which cases it made errors. The resulting confusion matrix is presented in Figure 5.

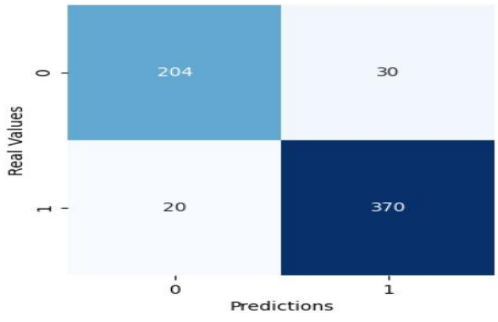

Fig.5. Confusion matrix results

• True Positive (TP): Represents the case where the model accurately predicts the positive class.

TP = Numerical Value

• True Negative (TN): Represents the case where the model accurately predicts the negative class.

TN = Numerical Value

• False Positive (FP): Represents the case where the model incorrectly predicts the positive class.

FP = Numerical Value

• False Negative (FN): Represents the case where the model incorrectly predicts the negative class.

FN = Numerical Value

Performance metrics such as precision, recall, and accuracy are obtained from the confusion matrix values.

Precision measures how many of the samples predicted as positive are actually positive. It expresses the ratio of true positives to the total positive predictions, as shown in Formula 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall measures how many of the true positives are detected. It expresses the ratio of true positives to the total number of positive examples, as shown in Formula 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Accuracy expresses the ratio of correctly predicted examples to the total number of examples. It is a metric that evaluates the overall model performance, as shown in Formula 7.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

F1-score balances precision and recall. This metric tends to minimize both false positives and false negatives, especially in balanced classification problems. Formula 8 illustrates the F1-score.

$$F! = 2\ x\ \frac{Precision\ x\ Recall}{Precision +\ Recall} \quad (8)$$

When the ResNet50 model, used as a Transfer Learning (TL) method, is applied to the same dataset, the F1 score, precision, recall, and support metrics are provided in Table 3.

TABLE III
CONFUSİON MATRIX RESULTS

	Accuracy	precision	Recall	F1-score	Support
NORMAL	0.92	0.91	0.87	0.89	234
PNEUMANIA	0.92	0.93	0.95	0.94	234

Subsequently, Grad-CAM was applied to visualize activation patterns within the convolutional layers of the CNN. This study clarified each step by emphasizing specific layers such as "conv2d2" for Grad-CAM Top and "conv2d1" for Grad-CAM Mid.

The integration of SHAP's analysis involved applying an image mask to the model's predictions, enabling a deeper understanding of the decision-making process with 15,000 evaluations. An example containing 15,000 evaluations is presented in Figure 6.

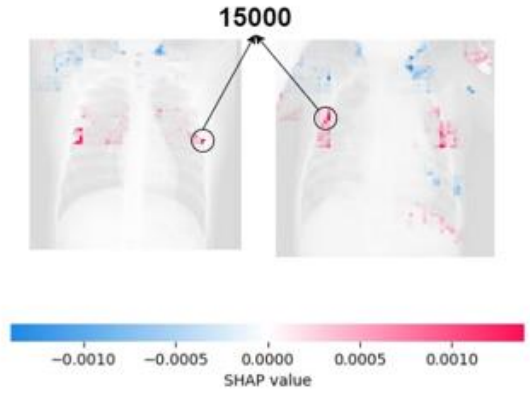

Fig.6. tems contributing to model misinterpretation in Shap analysis with 15,000 evaluations.

A notable issue encountered in this study is the model's susceptibility to misinterpretations. SHAP analysis highlights situations where the model may be misled by seemingly insignificant artifacts or misplaced objects in the image mask. This underscores the importance of meticulous preprocessing and awareness of potential errors in medical image analysis.

To illustrate this situation, we present a visual representation of misclassifications, showing instances where the model made correct predictions and errors in Figure 7.

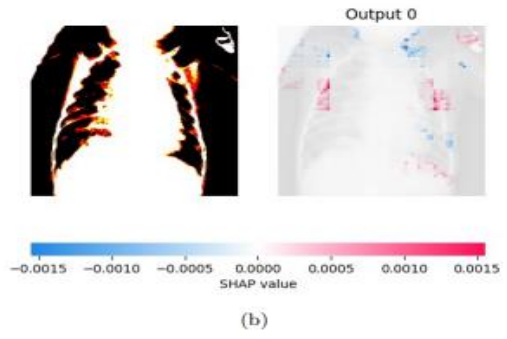

Fig.7. Misleading items causing model misinterpretation in Shap analysis.

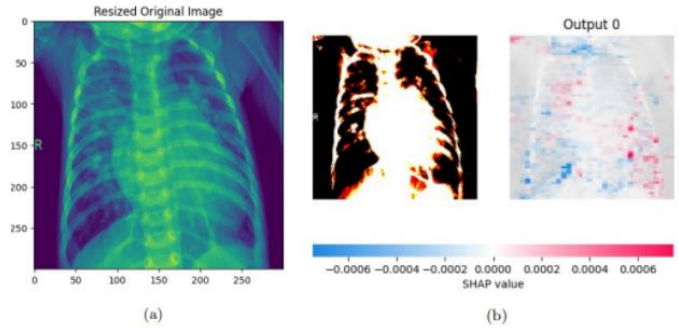

Fig.8. Comparison of (a) Original Image and (b) SHAP Output

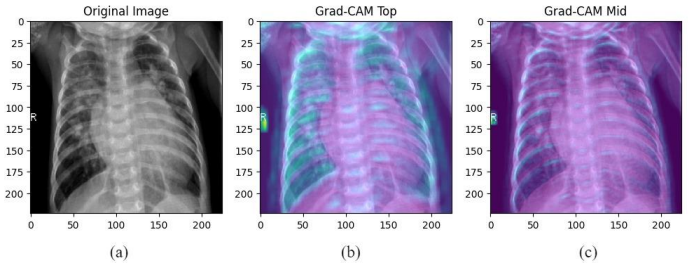

Fig.9. Grad-CAM Output

True Positive (TP): In the case of True Positive, where the model correctly identified instances of the disease, we initiated the analysis with the original image and progressed through subsequent stages. Figure 8 (a) shows our starting point with the original image. SHAP elucidated why the ResNet50 model classified this example as positive, highlighting the features supporting the positive decision. Figure 8 (b) presents the SHAP output, offering a detailed glimpse into the model's decision rationale. Grad-CAM, as illustrated in Figure 9, plays a crucial role in unraveling the rationale behind the Convolutional Neural Network (CNN) recognizing the provided example as positive. Figure 9 (a) serves as the anchor, showcasing the original image for our analysis. Figures 9 (b) and 9 (c) spotlight the two phases of Grad-CAM. Figure 9 (b) shows the completed Grad-CAM highlighting the top contributing regions, while Figure 9 (c) captures the ongoing process, emphasizing the mid regions. The integration of SHAP and Grad-CAM not only enriches our understanding of positive classification but also provides a unique perspective on the model's decision-making process.

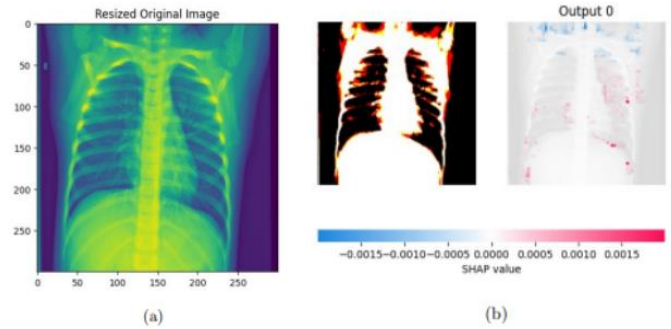

Fig.10. Comparison of (a) Original Image and (b) SHAP Output

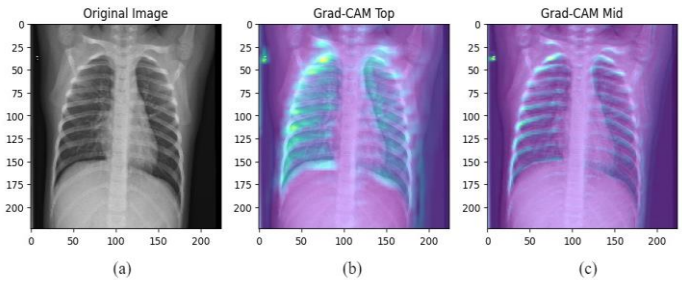

Fig.11. Grad-CAM Output

True Negative (TN): In the case of True Negative, where the model correctly identified a healthy state, our analysis began with the original image, setting the stage for subsequent examinations. Figure 10 (a) depicts the initial step with the original image, followed by Figure 10 (b), presenting the SHAP output that sheds light on the reasons behind the negative classification. Grad-CAM, featured in Figure 11, was employed to unravel the Convolutional Neural Network's (CNN) reasoning behind identifying this instance as negative, visually emphasizing the contributing regions. Figure 11 (a) presents the original image, forming the basis for our analysis. Figures 11 (b) and 11c shed light on the two distinctive phases of Grad-CAM. In Figure 11 (b), the completed Grad-CAM showcases the highlighted top contributing regions, while Figure 11 (c) captures the ongoing process, accentuating the mid regions. The strategic fusion of SHAP and Grad-CAM not only enhances our understanding of negative classifications but also introduces a fresh dimension to the interpretability of the model.

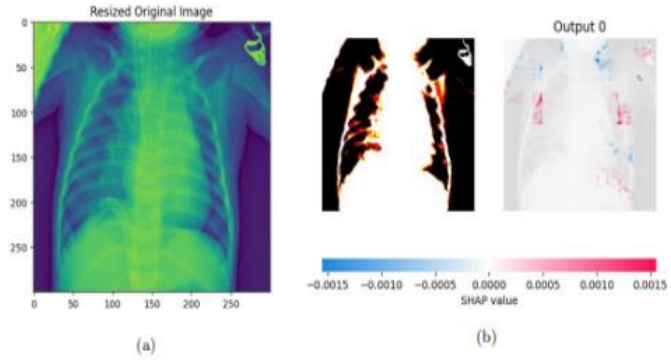

Fig.12. Comparison of (a) Original Image and (b) SHAP Output

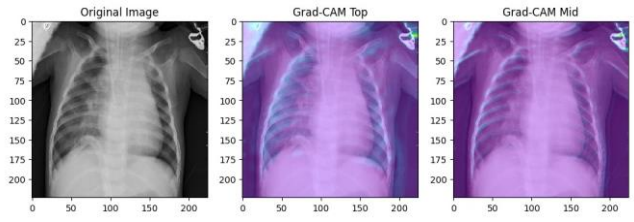

Fig.13. Grad-CAM Output

False Positive (FP): For False Positive instances where the model inaccurately predicted disease presence, our investigation began with the original image, offering insights into the misclassification. Figure 12 (a) captures the starting point with the original image, followed by Figure 12 (b), showcasing the SHAP output that explains the false positive classification by the ResNet50 model. Grad-CAM, showcased in Figure 13, extensively explored the reasons behind the Convolutional Neural Network (CNN) model's erroneous positive identification, visually spotlighting the regions responsible for the misclassification. Figure 13 (a) exhibits the original image, forming the basis for our analysis. Figures 13 (b) and 13 (c) illustrate the two significant phases of Grad-CAM. In Figure 13 (b), the completed Grad-CAM reveals the highlighted regions contributing to the false positive identification, while Figure 13 (c) captures the ongoing process, emphasizing mid regions that played a role in the misclassification. The strategic amalgamation of SHAP and Grad-CAM not only exposes the misclassification patterns of the model but also introduces innovation in comprehending false positives.

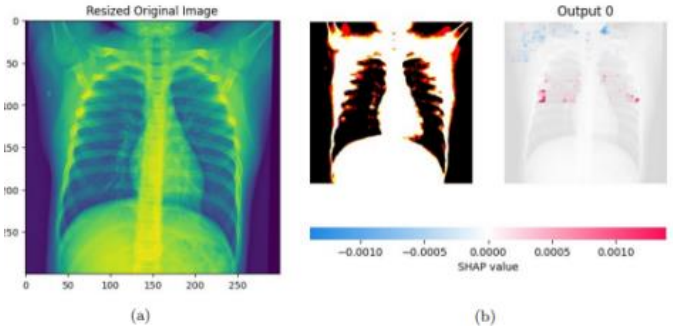

Fig.14. Comparison of (a) Original Image and (b) SHAP Output

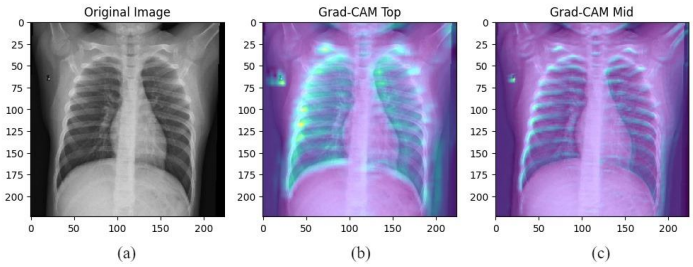

Fig.15. Grad-CAM Output

False Negative (FN): In cases of False Negative, where the model wrongly predicted a healthy state, our scrutiny began with the original image, providing insights into the misclassification. Figure 14 (a) marks our starting point with the original image, succeeded by Figure 14 (b), showcasing the SHAP output that explains the false-negative classification by the ResNet50 model. Grad-CAM, depicted in Figure 15, delved deeper into understanding the reasons behind the Convolutional Neural Network (CNN) model's erroneous negative identification, visually representing the regions accountable for the misjudgment. Figure 15 (a) showcases the original image as the foundation for our analysis. Figures 15 (b) and 15 (c) delineate the two key phases of Grad-CAM. In Figure 15 (b), the completed Grad-CAM unveils the highlighted regions contributing to the false negative identification, while Figure 15 (c) captures the ongoing process, emphasizing mid regions that

played a role in the misjudgment. The collaborative synergy of SHAP and Grad-CAM not only brings to light the model's misjudgments but also introduces a fresh perspective on comprehending false negatives.

The results obtained in this study are noteworthy when compared to other studies in the literature, revealing both similarities and differences. For instance, Yang et al. used the VGG16 model, considering background factors in pneumonia diagnosis, and achieved an accuracy of 95.6 [6]. Their results, particularly in terms of model explainability using Grad-CAM, align with our ResNet50-based model. However, our model incorporates an additional layer of explainability through SHAP, allowing for a deeper understanding of the decision-making process. This added explainability supports the model's reliability in clinical applications.

In the study conducted by De Moura et al. on COVID-19 pneumonia, SHAP and Grad-CAM were utilized with the XGBoost model, resulting in an accuracy of 82 [7]. In comparison, the integration of SHAP and Grad-CAM in our model achieved higher accuracy, underscoring the robustness of ResNet50 as a transfer learning model in medical image analysis tasks.

Moreover, Zou et al.'s work, which combined Grad-CAM++ and SHAP to predict the mortality risk in pneumonia patients [9], highlighted the critical role of explainability in clinical decision-making. Similarly, in our study, the integration of SHAP and Grad-CAM significantly enhanced the explainability of the model, thereby strengthening its reliability for clinical use.

Finally, while Alsharif et al. achieved a remarkable accuracy of 99.7 with their CNN-based PneumoniaNet framework for pediatric pneumonia diagnosis [11], our results are comparably strong. Although their model demonstrated high accuracy, the addition of explainability techniques such as SHAP and Grad-CAM in our model offers a clearer understanding of the decision-making process. These explainability methods not only improve the reliability of deep learning models but also enhance their acceptance in clinical settings by providing greater transparency.

IV. Conclusions and the Scope for Future Work

Chest X-ray analysis plays a vital role in pneumonia diagnosis, and recent advancements in Deep Learning (DL) methods have significantly increased the accuracy of automated diagnosis. This article explores the intersection of DL and explainable artificial intelligence (XAI) in the context of pneumonia diagnosis through chest X-rays. Using the ResNet50 model from CNNs and Transfer Learning (TL) methods, an accuracy of 95.23 was achieved.

In this study, we found that the combination of ResNet50, SHAP, and Grad-CAM provides a robust methodology for interpreting and explaining pneumonia diagnosis model decisions. SHAP's ability to individually evaluate the contribution of each input to the model output allows medical professionals to better understand why the model made a particular decision and assess the reliability of that decision. Grad-CAM, on the other hand, visually shows which regions the model considers, but it may sometimes be insufficient for a deep understanding of the decision-making process.

Our results show that when SHAP and Grad-CAM are used together, they can enhance interpretability in medical imaging analyses. Future studies should aim to further explore the synergy between these two methods to develop and improve methodologies for medical imaging analyses. The combination of SHAP's quantitative explanations with Grad-CAM's visual explanations can provide medical professionals with a more holistic and reliable interpretation. Such integration could enhance the reliability of machine learning models in clinical applications and contribute to the development of more effective decision support systems.

Funding Information
No funding

Data Availability
The Kaggle link of dataset is "https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia".

Conflict of Interest
The author declare that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Considerations
This research adheres to ethical principles and guidelines in conducting a comparative analysis of Explainable Artificial Intelligence (XAI) techniques, specifically SHAP (SHapley Additive exPlanations) and Grad-CAM (Gradient-weighted Class Activation Mapping), on pneumonia X-ray dataset.

Declarations
Ethics Approval Not Applicable
Competing interests The author declare no competing interests.

References

[1] H. Sharma, et al., "Feature Extraction and Classification of Chest X-Ray Images Using CNN to Detect Pneumonia," in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2020, pp. 227–231.

[2] S.-H. Lo and Y. Yin, "A novel interaction-based methodology towards explainable AI with better understanding of Pneumonia Chest X-ray Images," *Discover Artificial Intelligence*, vol. 1, no. 1, p. 16, 2021.

[3] M. Eisen and A. Ribeiro, "Optimal Wireless Resource Allocation With Random Edge Graph Neural Networks," *ARXIV-EESS.SP*, 2019.

[4] M. Rahimzadeh and A. Attar, "A Modified Deep Convolutional Neural Network for Detecting COVID-19 and Pneumonia from Chest X-ray Images Based on The Concatenation of Xception and ResNet50V2," *ARXIV-EESS.IV*, 2020.

[5] G. F. Elsayed, B. Wohlberg, and S. Jastrzębski, "Deep Double Descent: Where Bigger Models and More Data Hurt," *ARXIV-EESS.ST*, 2020.

[6] Y. Yang, G. Mei, and F. Piccialli, "A Deep Learning Approach Considering Image Background for Pneumonia Identification Using Explainable AI (XAI)," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. PP, 2022, doi: 10.1109/TCBB.2022.3190265.

[7] L. V. de Moura, C. Mattjie, C. M. Dartora, R. C. Barros, and A. M. Marques da Silva, "Explainable Machine Learning for COVID-19 Pneumonia Classification With Texture-Based Features Extraction in Chest Radiography," *Frontiers in Digital Health*, vol. 3, 2022, doi: 10.3389/fdgth.2021.662343.

[8] H. Ren, et al., "Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models With Multisource Data," *IEEE Access*, vol. 9, pp. 95872–95883, 2021, doi: 10.1109/ACCESS.2021.3094025.

[9] L. Zou, et al., "Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory

infections," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 2, pp. 242–254, 2022, doi: 10.1109/TAI.2022.3154871.
[10] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *Journal of Healthcare Engineering*, vol. 2019, Article ID 4180949, 2019, doi: 10.1155/2019/4180949.
[11] R. Alsharif, et al., "PneumoniaNet: Automated detection and classification of pediatric pneumonia using chest X-ray images and CNN approach," *Electronics*, vol. 10, no. 23, p. 2949, 2021, doi: 10.3390/electronics10232949.
[12] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, and A. Mittal, "Pneumonia detection using CNN based feature extraction," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1–7, doi: 10.1109/ICECCT.2019.8869364.
[13] W. Zhang, et al., "Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network," *ARXIV-EESS.IV*, 2019.
[14] D. Valsesia, et al., "Deep Graph-Convolutional Image Denoising," *ARXIV-EESS.IV*, 2019.
[15] M. Gil-Martín, J. Montero, and R. San-Segundo, "Parkinson's Disease Detection from Drawing Movements Using Convolutional Neural Networks," *ELECTRONICS*, 2019.
[16] H. Gao, et al., "PhyGeoNet: Physics-Informed Geometry-Adaptive Convolutional Neural Networks For Solving Parameterized Steady-State PDEs On Irregular Domain," *ARXIV-EESS.IV*, 2020.
[17] F. Eitel, K. Ritter, and Alzheimer's Disease Neuroimaging Initiative (ADNI), "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings*, 2019, pp. 3–11.
[18] I. D. Apostolopoulos and T. Bessiana, "Covid-19: Automatic Detection From X-Ray Images Utilizing Transfer Learning With Convolutional Neural Networks," *ARXIV-EESS.IV*, 2020.
[19] H. S. Maghdid, et al., "Diagnosing COVID-19 Pneumonia From X-Ray And CT Images Using Deep Learning And Transfer Learning Algorithms," *ARXIV-EESS.IV*, 2020.
[20] N. E. M. Khalifa, et al., "Detection of Coronavirus (COVID-19) Associated Pneumonia Based on Generative Adversarial Networks and A Fine-Tuned Deep Transfer Learning Model Using Chest X-ray Dataset," *ARXIV*, 2020.
[21] T. Rahman, et al., "Transfer Learning With Deep Convolutional Neural Network (CNN) For Pneumonia Detection Using Chest X-ray," *ARXIV-EESS.IV*, 2020.
[22] P. R. A. S. Bassi and R. Attux, "A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays," *ARXIV-EESS.IV*, 2020.
[23] Z. Zhou, et al., "Models Genesis: Generic Autodidactic Models For 3D Medical Image Analysis," *ARXIV-EESS.IV*, 2019.
[24] Y.-A. Chung and J. Glass, "Generative Pre-Training For Speech With Autoregressive Predictive Coding," *ARXIV-EESS.AS*, 2019.
[25] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep Neural Network Models For Computational Histopathology: A Survey," *ARXIV-EESS.IV*, 2019.
[26] Z. Zhao, et al., "Applications of Unsupervised Deep Transfer Learning to Intelligent Fault Diagnosis: A Survey," *ARXIV-EESS.SP*, 2019.
[27] M. Goyal, et al., "Artificial Intelligence-Based Image Classification For Diagnosis Of Skin Cancer: Challenges And Opportunities," *ARXIV-EESS.IV*, 2019.
[28] S. M. Lundberg, et al., "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017.
[29] SHAP Contributors, "SHAP (SHapley Additive exPlanations) Documentation," 2020. [Online]. Available: https://shap.readthedocs.io/en/latest/
[30] S. Basu, S. Mitra, and N. Saha, "Deep Learning For Screening COVID-19 Using Chest X-Ray Images," *ARXIV-EESS.IV*, 2020.
[31] C. Xia, et al., "Vision Based Defects Detection for Keyhole TIG Welding Using Deep Learning with Visual Explanation," *Journal of Manufacturing Processes*, 2020.
[32] M. R. Karim, et al., "DeepCOVIDExplainer: Explainable COVID-19 Diagnosis Based On Chest X-ray Images," *ARXIV-EESS.IV*, 2020.
[33] S. Vijayarangan, et al., "Interpreting Deep Neural Networks For Single-Lead ECG Arrhythmia Classification," *ARXIV-EESS.SP*, 2020.
[34] M. Kim, et al., "Medinoid: Computer-Aided Diagnosis and Localization of Glaucoma Using Deep Learning," *Applied Sciences*, 2019.
[35] I. Elbouknify, et al., "CT-xCOV: A CT-scan Based Explainable Framework for COVID-19 Diagnosis," *ARXIV-EESS.IV*, 2023.

BIOGRAPHIES

Cem Ozkurt, received his B.Sc. in Industrial Engineering from Sakarya University in 2002. He completed two M.Sc. degrees in Industrial Engineering and Curriculum & Instruction in 2016, and earned his Ph.D. in Industrial Engineering in 2020. Since 2018, he has held academic positions at Sakarya University of Applied Sciences, currently serving in the Department of Computer Engineering. His research focuses on digital transformation, artificial intelligence in industry, and Industry 4.0.