Explainable AI Based Fully Fine-Tuned Data-efficient Image Transformer (DeiT-B) Model for Multi Class Chest X-Ray Image Classification

Md Parvez Kabir[1]* Md Jahidul Islam Mozumdar[2] MD REZAUL[3]

Rasedul Islam[4]

Sourav Ghosh[5] and Md Toha Hayder[6]

[1] Daffodil International University, Savar, Bangladesh
[2] Daffodil International University, Savar, Bangladesh
[3] Daffodil International University, Savar, Bangladesh
[4] Daffodil International University, Savar, Bangladesh
[5] Daffodil International University, Savar, Bangladesh
[6] Daffodil International University, Savar, Bangladesh

[1] kabir15-5539@diu.edu.bd*
[2] mozumdar15-4790@diu.edu.bd
[3] rezaul15-5822@diu.edu.bd
[4] islam2305101540@diu.edu.bd
[5] ghosh2305101775@diu.edu.bd
[6] hayder02423100051011459@diu.edu.bd

Abstract. Chest X-ray imaging is utilized significantly in the diagnosis of respiratory disorders like COVID-19, viral pneumonia, and lung opacities. Deep learning has evolved computerized classification systems that are able to assist radiologists in making more accurate and rapid diagnoses. In this paper, we propose a fully fine-tuned Data-efficient Image Transformer (DeiT-B) model with Explainable AI (XAI) techniques, including LIME and Attention maps, for chest X-ray image classification. The method leverages DeiT-B's attention mechanism to focus on relevant regions of the X-ray images and provide visual explanations of its predictions. The model was trained and tested on 4,800 chest X-ray images from a Kaggle dataset. Experimental outcomes demonstrate that the model achieves a test accuracy of 95.21%, its weighted precision, recall, and F1-score values of 95.46%, 95.21%, and 95.26%, and its Cohen's Kappa value is 0.9361, superior to baseline CNN models such as VGG16, ResNet18, ResNet50, EfficientNetB3, DenseNet121, and even compared with transformer-based models such as ViT-B/16. XAI integration, in the sense of LIME and Attention maps, ensures interpretability and reliability of the model, thereby making the model suitable for real-world clinical application.

Keywords: Chest X-ray, Data-efficient Image Transformer, Explainable AI, LIME, Deep Learning

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_5

1 Introduction

Chest X-ray is a highly prevalent and low-cost diagnostic tool for identifying and monitoring pulmonary diseases such as COVID-19, viral pneumonia, and lung opacity [1]. X-ray image interpretation, if done correctly, can result in early diagnosis and early treatment, which can significantly improve patient outcomes. Manual interpretation by radiologists, however, is time-consuming, subjective, and prone to errors, particularly in epidemic outbreaks or high-volume clinical practice [2]. Advances in recent deep learning permit automatic image classification system construction to aid radiologists in clinical decision-making [3], [4]. Convolutional neural networks (CNNs) like VGG16, ResNet, and DenseNet have been able to deliver state-of-the-art performance in medical image analysis by learning hierarchical feature representation. While successful, CNNs fail to capture long-range dependencies and global contextual information in chest X-ray images, making them less interpretable and robust in general.Transformer-based architectures, such as Vision Transformers (ViT) and Data-efficient Image Transformers (DeiT), have recently reached state-of-the-art performance by leveraging self-attention mechanisms to model global dependencies in images [5], [6].Such models have great potential for medical imaging applications where small changes in tissue patterns can signify different diseases. In this study, we propose a fully fine-tuned DeiT-B model coupled with Explainable AI (XAI) techniques, i.e., LIME attention maps, for chest X-ray classification. Not only does the proposed technique achieve high classification accuracy, but also provides visual explanations depicting the regions influencing predictions, thereby improving interpretability and clinical reliability [7]. The model is trained and validated over 4,800 chest X-ray images in a Kaggle dataset and compared with the conventional CNN and transformer-based models for performance, demonstrating its practicality for real-world clinical application [8], [9].

In our study, we have the following major contributions:

- We propose a fully fine-tuned DeiT-B transformer model to classify chest X-rays with improved accuracy over baseline CNN and transformer models.
- We Utilize LIME-based explanation techniques to create visual attention maps, enabling model transparency and clinician interpretability.
- We provide a comparative performance evaluation with state-of-the-art CNN based models and transformer model, demonstrating the feasibility of the proposed framework for real-world medical imaging diagnosis.

2 Literature Review

Automated X-ray chest categorization has been extensively studied on the basis of its potential to assist radiologists, reduce diagnostic errors, and enhance disease detection time. The initial approaches largely employed CNN for feature extraction and classification. Rajpurkar et al. [10] proposed CheXNet, a DenseNet-based 121-layer deep neural network model, which was trained on a large database of chest X-rays

and identified pneumonia at a radiologist level. Although having high performance, the interpretability of CheXNet is limited and it requires large quantities of labeled training data. Wang et al. [11] introduced a multi-label CNN for eight thoracic disease classification under weakly supervised localization. Although effective, their approach is ineffective when handling small or class-imbalanced datasets and is unable to offer explanations of its predictions by design.There has been a recent shift towards transformer-based architectures since they are able to learn long-range dependencies and global context information. Dosovitskiy et al. [12] introduced the Vision Transformer (ViT), demonstrating that self-attention mechanisms can outperform CNNs on large vision recognition benchmarks. Sadly, ViT is trained with vast training sets so it is not directly translatable to medical imaging,which often have limited data. Touvron et al. [13] presented DeiT, a lightweight transformer applying knowledge distillation for an alleviation of dependence on large datasets, making it suitable for medical image classification. S. Aburass et al. [14] compared ViT with CNNs under medical image datasets and showed that the transformers outperform CNNs when maintaining global image patterns, particularly in disease classification multi-class.Explainable AI (XAI) has become increasingly important in medical application. Ribeiro et al. [15] introduced LIME, providing local explanations by approximating the model predictions with interpretable surrogate models. Selvaraju et al. [16] proposed Grad-CAM, generating heatmaps that represent areas that are more accountable for CNN predictions. Chattopadhyay et al. [17] generalized Grad-CAM to Grad-CAM++, representing various salient areas better in challenging images. Incorporation of XAI is essential since models such as CNNs and transformers are otherwise "black boxes" and therefore restrict clinical trust. A few studies tackled COVID-19 detection on chest X-rays directly. Apostolopoulos and Mpesiana [18] used transfer learning with CNNs for computerized COVID-19 detection, with high accuracy and no interpretability. Ozturk et al. [19] integrated CNNs with XAI, producing visual explanations in addition to classification. Although successful, their models did not take advantage of transformer architectures, which proved to be more capable of capturing global dependencies. To sum up, existing research demonstrates strong performance in chest X-ray classification but cites three main gaps: (1) poor interpretability of CNN models, (2) high data requirements of transformer models, and (3) inadequate integration of XAI for transformer-based medical images. Our work fills these gaps using a fully fine-tuned DeiT-B model and LIME attention maps to attain accurate and interpretable multi-class chest X-ray classification even on small datasets sizes.

3 Methodology

3.1 Dataset Description

In this study we uses the Kaggle COVID-19 Radiography Database [20], which contains 4,800 chest X-ray images distributed over four classes: lung opacity, viral pneumonia, COVID-19, and normal. The data were split into training(70%),

validation (15%), and testing (15%) sets for ensuring unbiased testing and proper generalization of the model [20]. The image per class depicts in Fig. 1.

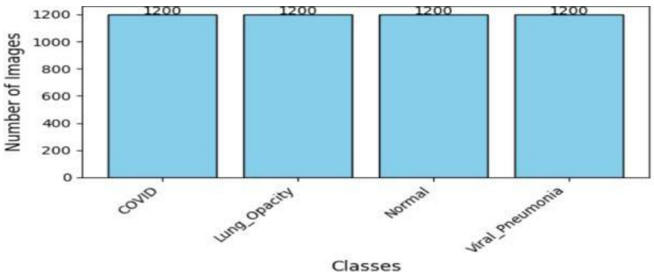

Fig. 1. Histogram (bar chart) of image count per class.

3.2 Preprocessing and Data Augmentation

All of the chest X-ray images were resized to 224×224 pixels order to meet the input requirement of the DeiT-B model. Pixel values were normalized to the ImageNet mean and standard deviation in order to normalize the input space. In order to augment generalization and preventing overfitting, all of the following data augmentation methods were applied to the training set: random resized cropping to 224×224 pixels, horizontal flipping with a fixed probability, and color jittering for changing the brightness, contrast, saturation, and hue and consequently to simulate varying imaging conditions. On the other hand, the validation and test sets were only resized and normalized for unbiased evaluation. Also, to address class imbalance, a weighted random sampler was incorporated during training so that all classes were equally weighted to contribute towards the learning process.

4 DeiT-B Model Architecture

The proposed framework uses the Data-efficient Image Transformer (DeiT-B) model that leverages self-attention to learn global dependencies among pixels in images [21]. The input chest X-ray image is divided into 16×16 patches, flattened, and projected into embeddings, and these are fed through a transformer encoder made up of multi-head self-attention layers and feed-forward networks [22]. A fully connected layer with softmax activation provides the class predicted probabilities.

Fine-tuning Strategy: DeiT-B is initialized with ImageNet pretrained weights. The entire network is fine-tuned over the chest X-ray dataset entirely in this study to achieve fine classification performance, although optionally layers may be frozen for the purpose of reducing trainable parameters.

4.1 Explainable AI Integration

For improving interpretability, the model incorporates LIME, which distorts the input image and inspects output differences to generate heatmaps that identify critical areas accountable for predictions [23]. These visual explanations may be overlaid on X-rays so that radiologists can better understand the reason behind the model's decision.
In addition to LIME, the internal self-attention mechanism of DeiT-B produces attention maps as visual representations of regions in the image extensively attended by the model when classifying [24]. This provides model-intrinsic interpretability, complementing LIME's model-agnostic explanations. Together, these two approaches ensure double interpretability and makes predictions reliable and clinically relevant.

4.2 Model Training Specifications

Optimizer: AdamW
Learning Rate: 3e-5
Weight Decay: 1e-2
Batch Size: 32
Number of Epochs: 15
Loss Function: Cross-Entropy Loss

4.3 Baseline Models for Comparison

To properly compare the performance of the proposed DeiT-B model, several renowned deep learning architectures were employed as baselines. Baselines for this task include convolutional neural networks (CNNs) as well as transformer-based models for comprehensive and impartial performance comparison.

- **VGG16:** 16 weight layer deep CNN which has extensive use in medicine imaging for its simplicity as well as high hierarchical feature extraction. It can possesses large parameter size, however, which typically makes computation cost better and training slower.
- **ResNet18:** A lightweight residual network which skip connections that alleviate vanishing gradients, making it suitable and stable for smaller datasets. Despite its shallow depth, it provides accurate baseline performance in medical classification tasks.
- **ResNet50:** A 50-layer deeper CNN network that can adapt more complex features using bottleneck blocks. Its deeper layers increase accuracy but enhance computational requirements over ResNet18.
- **EfficientNetB3:** Depth-wise, width-wise, and resolution-wise scaled CNN for increased efficiency. It is much efficient with lower parameters and thus perfect for resource-constrained settings.
- **DenseNet121:** Dense connected CNN where every layer is connected with all the previous layers, promoting gradient flow. It

supports feature reuse and is less prone to overfitting but not parameter-effective.

- **ViT-B/16:** A vision transformer that separates images into 16×16 patches and employs self-attention for acquiring global features. It can effectively learn long-distance dependencies but requires large-scale data for training optimally and need large traning time.

All the models were trained and tested on the same train-test dataset split and data preprocessing pipeline for a fair comparison with the new DeiT-B model.

4.4 Experimental Setup

All of the experiments were performed under the Kaggle platform using NVIDIA Tesla P100 GPU. Our suggested DeiT-B model was finetuned from the chest X-ray dataset entirely, where all of the layers were trainable. Resampling was done from images to 224×224 pixels and normalized according to ImageNet mean and standard deviation, and randomization of the training set using resized cropped, horizontally flipped, and color jittered was performed for better generalization. During training, a weighted random sampler was used to address class imbalance. AdamW optimizer, 3e-5 learning rate, and 1e-2 weight decay were used for training the model, and learning rate scheduling was performed using CosineAnnealingLR. The objective for training was cross-entropy loss, and the model was trained for 15epochs, along with loss and accuracy on training,validation, and test sets at each epoch.

4.5 Evaluation Metrics

Accuracy:

Accuracy calculates the proportion of the number of correctly classified samples out of all samples. It is defined as:

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

Here, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Weighted Precision:

Precision calculates the proportion of the number of correctly predicted positive samples out of all positive samples predicted. Weighted precision takes into account class imbalance:

$$\text{Weighted Precision} = \sum_{i=1}^{C} \frac{n_i}{N} \cdot \frac{TP_i}{TP_i + FP_i} \quad (2)$$

Here, TP_i and FP_i are the true positives and false positives for class i, n_i is the number of samples in class i, and N is the total number of samples.

Weighted Recall:

Recall calculates the proportion of the number of correctly predicted positive samples out of all actual positives. Weighted recall balances the class contributions:

$$\text{Weighted Recall} = \sum_{i=1}^{C} \frac{n_i}{N} \cdot \frac{TP_i}{TP_i + FN_i} \quad (3)$$

Here, T^{P_i} and F^{N_i} are the true positives and false negatives for class i, n_i is the number of samples in class i, and N is the total number of samples.

Weighted F1-Score:

Weighted F1-Score is utilized to evaluate the performance of the model over all classes under consideration of class imbalance. It provides a balanced estimate of precision and recall to prevent unfair evaluation even for minority classes

$$\text{Weighted F1} - \text{Score} = \sum_{i=1}^{C} \frac{n_i}{N} \cdot \frac{2TP_i}{2TP_i + FP_i + FN_i} \quad (4)$$

Here, T^{P_i} and F^{P_i} are the true positives and false positives for class i, n_i is the number of samples in class i, and N is the total number of samples.

Cohen's Kappa:

Cohen's Kappa is used to measure agreement between predicted and true labels, adjusting for agreement by chance:

$$\text{Cohen's Kappa} = \kappa = \frac{\rho_0 - \rho_e}{1 - \rho_e} \quad (5)$$

5 Results and Discussion

The proposed DeiT-B model achieved 95.21% test accuracy using Equation (1) and precision, recall, and F1-score of 95.46% using Equation (2), 95.21% using Equation (3), and 95.26% using Equation (4), respectively, in weighted average Cohen's Kappa was found to be 0.9361 using Equation (5), indicating strong agreement between true and predicted labels. For examining model performance closely, a confusion matrix was generated, illustrating per-class accuracy report and misclassifications (Fig. 2). In addition to that, t-SNE plots of the learned features were also plotted to visually illustrate the separability between different types of classes (Fig. 3). Accuracy comparison at varying epochs is shown in Fig. 4, and Explainable AI (XAI) attention maps generated using LIME are shown in Fig. 5. Performance comparison of all the baseline models and DeiT-B is shown in Table I, illustrating the superiority of the proposed method.

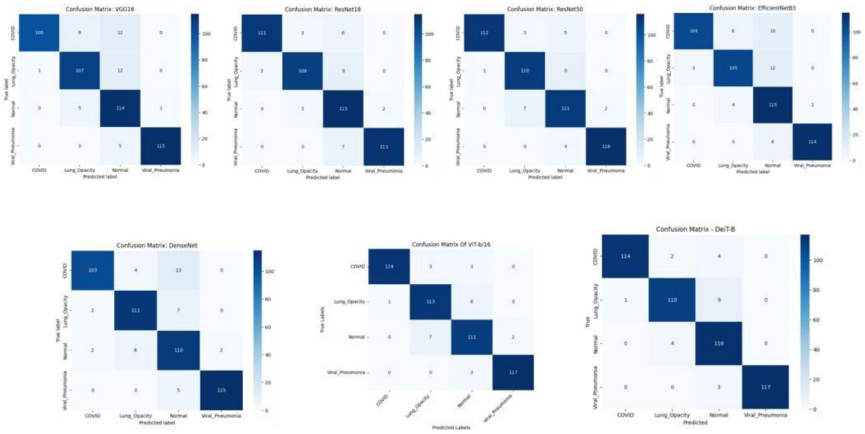

Fig. 2. Confusion matrices comparing DeiT-B with other deep learning models

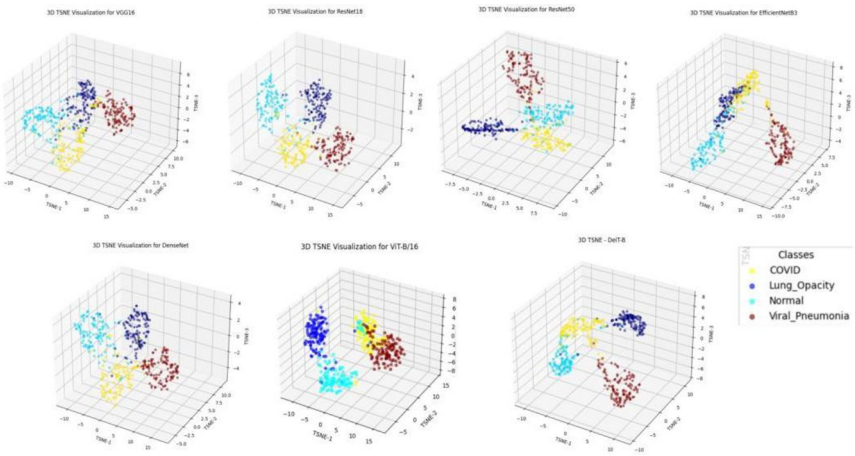

Fig. 3. t-SNE comparing DeiT-B with other deep learning models.

Table 1. Performance comparison of different deep learning models

Model	Train Acc (%)	Val Acc (%)	Test Acc (%)	Weighted Precision (%)	Weighted Recall (%)	Weighted F1 Score (%)	Cohen's Kappa (%)
VGG16	96.38	90.83	90.83	91.76	90.83	90.95	87.78

ResNet18	96.59	91.87	93.13	93.58	93.13	93.21	90.83
ResNet50	98.44	92.08	93.54	93.78	93.54	93.61	91.39
EfficientNetB3	90.29	88.75	91.25	92.01	91.25	91.37	88.33
DenseNet121	92.94	90.21	91.46	91.94	91.46	91.55	88.61
ViT-B/16	99.22	93.96	92.92	93.73	92.92	93.05	90.56
DeiT-B	**99.27**	**95.63**	**95.21**	**95.46**	**95.21**	**95.26**	**93.61**

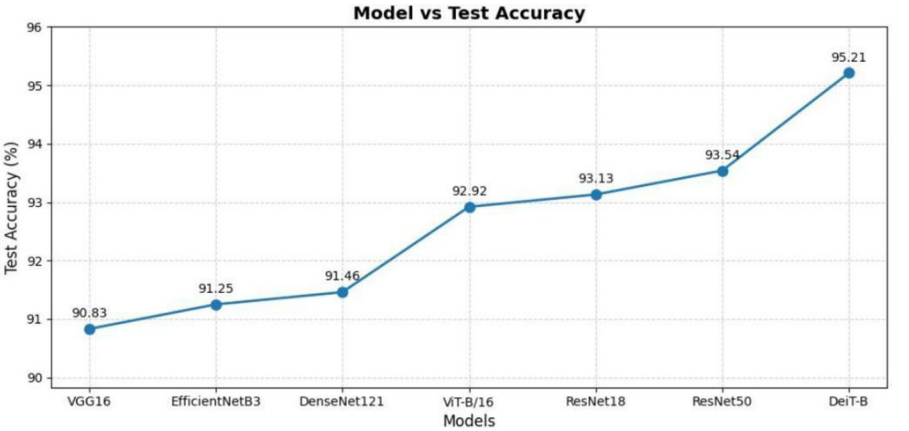

Fig. 4. Accuracy comparison with different deep learning models

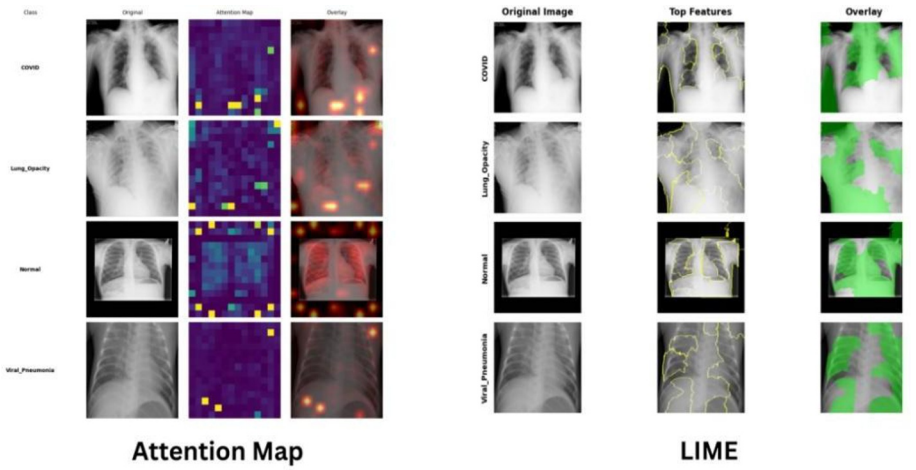

Fig. 5. DeiT-B model Attention map and LIME

5.1 Qualitative Analysis of XAI

Fig. 5 illustrates the explainability results obtained with the DeiT-B model by using Attention Maps (left) and LIME (right) across different classes of chest X-rays. Attention Maps highlight the internal focus of the model by underlining the most discriminative regions inside the thoracic cavity. Indeed, for COVID-19 and Pneumonia classes, the model consistently pays attention to clinically relevant areas corresponding to peripheral opacities and dense lung infiltrations that are typical radiological patterns for such conditions. Instead, Normal samples show uniformly low-activation regions corresponding to an absence of abnormal findings. Complementing this, the visualizations through LIME provide a localized and interpretable explanation by marking the superpixels which had the most influence on the model's decision. The highlighted regions through LIME correlate with the anatomical abnormalities shown in the Attention Maps, indicating coherence between the model's global and local interpretability cues. Overall, the combined visualization supports the clinical relevance of the model's decision process and strengthens the qualitative interpretability of the predictions.

5.2 Limitations

The major limitation of the present study is the dependence on a single dataset, the COVID-19 Radiography Database, which might raise problems of generalizability across various X-ray machines and demographics. Moreover, the in-built XAI framework provides interpretability through visualizations; however, this work lacks large-scale qualitative clinical validation by expert radiologists in order to establish its clinical precision.

5.3 Future work

Future work will validate the proposed model through diverse, multi-center data for its robustness in clinical practice. Moreover, a thorough clinical user study with radiologists will be performed to assess practical utility as well as trust-building capability in diagnostic workflows.

6 Conclusion

In this study, we have employed a fine-tuned full Data-efficient Image Transformer (DeiT-B) model with Explainable AI (LIME, Attention maps) in this paper for chest X-ray classification. Experimental results on 4,800 chest X-ray images show that our approach is better with a test accuracy of 95.21%, superb precision, recall, F1-score, and Cohen's Kappa of 0.9361, as compared to baseline CNN and transformer models. The use of LIME attention maps provides visual explanations that increase model

interpretability and clinical validity. The performances demonstrate that the proposed DeiT-B with XAI framework is stable and efficient, and it is a promising solution for interpretable, automatic, and accurate diagnosis of chest X-rays in clinics. The extension of this method to multi-modal imaging and real-time clinic implementation is the direction for future research.

7 References

[1] World Health Organization. Chest X-ray imaging in respiratory diseases. WHO, Geneva, Switzerland (2020).
[2] Delrue, L., Gosselin, R., Ilsen, B., Van Landeghem, A., de Mey, J., Duyck, P.: Difficulties in the interpretation of chest radiography. Medical Radiology, 27–49 (2010). https://doi.org/10.1007/978-3-540-79942-9_2
[3] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
[4] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM 60(6), 84–90 (2012).
[5] Panfilov, E., Saarakkala, S., Nieminen, M.T., Tiulpin, A.: Predicting knee osteoarthritis progression from structural MRI using deep learning. In: Proc. ISBI, 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761458
[6] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention. arXiv preprint arXiv:2012.12877 (2021).
[7] Wang, Y., Wang, J., Zhang, H., Song, J.: Bridging prediction and decision: Advances and challenges in data-driven optimization. Nexus 2(1), 100057 (2025). https://doi.org/10.1016/j.ynexs.2025.100057
[8] Kaggle: Chest X-ray Dataset (2021). https://www.kaggle.com/datasets
[9] Takahashi, S., et al.: Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. Journal of Medical Systems 48(1) (2024). https://doi.org/10.1007/s10916-024-02105-8
[10] Rajpurkar, P., et al.: CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017).
[11] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks. In: Proc. CVPR (2017). https://doi.org/10.1109/CVPR.2017.369
[12] Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
[13] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention. arXiv preprint arXiv:2012.12877 (2021). (Duplicate of [6], keep only if cited twice)
[14] Aburass, S., Dorgham, O., Al Shaqsi, J., Abu Rumman, M., Al-Kadi, O.: Vision transformers in medical imaging: A comprehensive review. Journal of Imaging Informatics in Medicine (2025). https://doi.org/10.1007/s10278-025-01481-y
[15] Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. arXiv preprint arXiv:1602.04938 (2016).
[16] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations via gradient-based localization. International Journal of Computer Vision 128(2), 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7

[17] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized gradient-based visual explanations. In: Proc. WACV (2018). https://doi.org/10.1109/WACV.2018.00097
[18] Apostolopoulos, I.D., Mpesiana, T.A.: COVID-19 detection from X-ray images using transfer learning. Physical and Engineering Sciences in Medicine (2020). https://doi.org/10.1007/s13246-020-00865-4
[19] Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Acharya, U.R.: Automated detection of COVID-19 using deep neural networks. Computers in Biology and Medicine 121 (2020). https://doi.org/10.1016/j.compbiomed.2020.103792
[20] Rahman, T.: COVID-19 Radiography Database [Dataset]. Kaggle (2021). https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database
[21] Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. In: ICLR (2021). https://arxiv.org/abs/2010.11929
(Duplicate of [12], normally remove unless cited separately)
[22] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention. arXiv preprint arXiv:2012.12877 (2021). (Duplicate of [6], [13])
[23] Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. arXiv preprint arXiv:1602.04938 (2016). (Duplicate of [15])
[24] Selvaraju, R.R., et al.: Grad-CAM: Visual explanations. In: Proc. ICCV, 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74