

Performance Evaluation of Transfer Learning Based Medical Image Classification Techniques for Disease Detection

Zeeshan Ahmad¹, *Member, IEEE*, Shudi Bao^{1,*}, *Member, IEEE*, and Meng Chen²

Abstract—Medical image classification plays an increasingly vital role in identifying various diseases by classifying medical images, such as X-rays, MRIs and CT scans, into different categories based on their features. In recent years, deep learning techniques have attracted significant attention in medical image classification. However, it is usually infeasible to train an entire large deep learning model from scratch. To address this issue, one of the solutions is the transfer learning (TL) technique, where a pre-trained model is reused for a new task. In this paper, we present a comprehensive analysis of TL techniques for medical image classification using deep convolutional neural networks. We evaluate six pre-trained models (AlexNet, VGG16, ResNet18, ResNet34, ResNet50, and InceptionV3) on a custom chest X-ray dataset for disease detection. The experimental results demonstrate that InceptionV3 consistently outperforms other models across all the standard metrics. The ResNet family shows progressively better performance with increasing depth, whereas VGG16 and AlexNet perform reasonably well but with lower accuracy. In addition, we also conduct uncertainty analysis and runtime comparison to assess the robustness and computational efficiency of these models. Our findings reveal that TL is beneficial in most cases, especially with limited data, but the extent of improvement depends on several factors such as model architecture, dataset size, and domain similarity between source and target tasks. Moreover, we demonstrate that with a well-trained feature extractor, only a lightweight feedforward model is enough to provide efficient prediction. As such, this study contributes to the understanding of TL in medical image classification, and provides insights for selecting appropriate models based on specific requirements.

I. INTRODUCTION

Medical imaging is a key research domain in modern healthcare systems. It enables a non-invasive visualization of anatomical structures and detailed description of various physiological processes. Various techniques, such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) are used for the diagnosis of numerous diseases followed by treatment planning and patient monitoring [1]. Generally, X-rays are used to detect fractures in body parts. Besides, they are also used to identify different diseases, such as pneumonia [2], lung cancer [3], emphysema, heart problems [4], multiple

fundus images detection [5], and covid detection [6]. Pneumonia is a condition where the air sacs in the lungs become filled with fluid or air, leading to breathing difficulties. In chest X-rays of individuals with pneumonia, white patches are often visible. Pulmonary nodules, on the other hand, are small spots that can develop in the lungs, and may be either malignant (cancerous) or benign (noncancerous). These nodules also appear as white spots in X-ray images and can be identified using low-dose CT scans.

In recent years, the application of deep learning (DL) models in healthcare has attracted great research interest, with the majority concentrated on medical imaging. It has been verified that neural networks [7] can generalize any nonlinear function since they follow the universal theorem [8]. Given a dataset, the neural network aims to find an efficient data-based surrogate model for predicting output features.

At present, there have been several comprehensive and in-depth studies conducted on the application of DL techniques in medical and healthcare domain. For example, Litjens et al. [9] reviewed over 300 articles on DL for medical image analysis, whereas Chowdhury et al. [10] examined state-of-the-art research on self-supervised learning in medicine. Other studies have concentrated on transfer learning (TL) with specific case studies, such as microorganism counting [11], cervical cytopathology [12], neuroimaging biomarkers of Alzheimer's disease [13], and magnetic resonance imaging of the brain. Kim et al. [14] provided a comprehensive review of TL techniques for medical image classification.

Despite the success of neural networks in recent years, most of the modelling techniques face issues, such as model interpretation [15], error convergence [16], and class imbalance [17]. Model interpretation is related to the layerwise parameter analysis. For a deep learning model, such as a convolutional neural network (CNN), the analysis becomes intractable due to its millions or even billions of parameters. In most cases, the training algorithm faces convergence issues which are mostly related to the optimization of a highly nonlinear objective function with possible local minima and saddle points. Class imbalance is another issue which comes from the data distribution. If a dataset has more data of one class "A", and scarce data for another class "B", then the model tends to provide more importance for prediction of class "A" compared to class "B".

Model performance for training a large neural network from scratch depends on the initial weight distribution. Since initially all the weights are randomly distributed, there always remains an uncertainty in the model performance.

This work was supported in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University (No.2023D15), Ningbo Clinical Research Center for Medical Imaging (No.2022LYKFYB01), and the Natural Science Foundation of Zhejiang, China (LGF22H120009). (*Corresponding author: Shudi Bao)

¹Zeeshan Ahmad and Shudi Bao are with the Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo 315200, People's Republic of China (email: sdbao@idt.eitech.edu.cn)

²Meng Chen is with the School of Cyber Science and Engineering, Ningbo University of Technology, Ningbo 315211, People's Republic of China (email: cm@nbut.edu.cn)

TL is a popular method that uses pre-trained models as a starting point for a new task, leveraging knowledge gained from one domain to improve performance in another domain [14]. This approach is particularly beneficial in medical imaging, where large labeled datasets are often scarce. By utilizing pre-trained models, typically from natural image datasets, e.g., ImageNet, TL can significantly reduce training time, computational costs, and the need for extensive labeled data in medical image classification tasks [11]–[13]. Several studies have demonstrated the effectiveness of TL in medical imaging, such as classification of skin and breast cancer images [14], and pneumonia detection in chest X-rays [18].

In this paper, we present a comprehensive case study of several state-of-the-art DL models for a custom medical imaging dataset. In particular, we have shown that rather than training a model from scratch, if we consider training a pre-trained model, the convergence of loss function is numerous times faster compared to the earlier approach. In addition, we also provide the uncertainty analysis which provides a bound on the loss value for certain training process, and runtime comparison of different models to provide background into the model's reproducibility. From this analysis, we will be able to find which type of deep learning architecture is useful for a particular type of dataset. This study focuses on models with convolutional feature extractors. While Vision Transformer (ViT) based models have shown efficiency in medical image analysis, they have fundamentally different architectures than convolutional models. Therefore, ViT based models are excluded from our comparative analysis.

The rest of the paper is organized as follows. Section II provides a detailed description of the methodology. In section III, we conduct experiments to evaluate the performance of various state-of-the-art TL based deep learning techniques. Finally, we make our conclusions and discuss future work in section IV.

II. METHODOLOGY

Neural networks are empirical models that are inspired by and resembles the idea of neural information processing of the nervous systems. The fundamental building block of a neural network is a series of interconnected nodes which constitute a layer. These nodes process the input data from the early layer nodes by applying weights and biases, and computes the final output based on learned features in the output layer.

Given the input matrix $X \in \mathbb{R}^{m,n}$, where m is the number of data points and n is the number of features, the output z_i^l of the l th layer for the i th data point is then given by

$$z_i^{(l)} = \sum_j w_{ij}^{(l)} a_j^{(l-1)} + b_i^{(l)} \quad (1)$$

where $w_{ij}^{(l)}$ and $b_i^{(l)}$ respectively denote the weight and bias associated with the i th neuron in the l th layer, and

$$a_i^{(l)} = \phi(z_i^{(l)}) \quad (2)$$

is the activations from the previous layer ($l - 1$). Here, ϕ denotes a nonlinear activation function that introduces non-

linearity into the model, and X can be considered as $a_j^{(0)}$, that is, the inputs to the first hidden layer. Without a non-linear activation function, the neural network is essentially a linear regression model, which performs a linear transformation from the input space to the output space.

The concept of TL arises from the ability of lower layers of a deep neural network to capture generic features of the data, such as horizontal and vertical patterns in images. The schematic diagram of a simple TL approach is shown in Fig. 1.

Let consider a DL model θ_{source} that is trained on a source task, minimizing the source loss L_S by updating the weights θ as shown in Eq. 3. After numerous iterations following the Eq. 5, we learn the weights which is now applied to the target learning task shown in Eq. 4. During training the model for target task, part of the pre-trained network remains fixed (frozen) and only deeper network layers are trained following Eq. 6. Here, f_{shared} denotes the pre-trained network's layers which are frozen, and only task specific layers $f_{specific}$ are trained.

$$\theta_{source} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N_S} \sum_{i=1}^{N_S} L_S(f_{source}(x_i^T; \theta), y_i^S) \quad (3)$$

$$\theta_{target} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N_T} \sum_{i=1}^{N_T} L_T(f_{target}(x_i^T; \theta), y_i^T) \quad (4)$$

$$\theta_{target}^{(k)} = \theta_{target}^{(k-1)} = \alpha \nabla_{\theta} L_T^{(k)} \quad (5)$$

$$f_{target}(x; \theta) = f_{specific}(f_{shared}(x; \theta_{shared}); \theta_{specific}) \quad (6)$$

We explore different pre-trained models, such as ResNet [19], AlexNet [20], VGG [21] and InceptionV3 [22], in the comparative study. A majority of these networks considers residual network based models as the building block. Residual networks are efficient in solving vanishing gradient problem and are robust during learning process. Residual layers can be expressed as

$$H_i^{(l)}(x) = [\sum_j w_{ij}^{(l)} a_j^{(l-1)} + b_i^{(l)}] + H_j^{(l-k)}(x) \quad (7)$$

where $w_{ij}^{(l)}$ is the kernel matrix which converts the input space of j th dimension to an output space of the i th dimension, and H represents the output of a hidden layer. We compute $H_i^{(l)}$ for the i th node in the l th layer using Eq. 7. Since we are concatenating the output of a previous hidden layer to the output of the current hidden layer, the size of both layer's output should be the same. Hence, the length of index i should be same as the length of index j . According to this statement, the kernel matrix w_{ij} should be a square matrix. This process is called identity mapping, which provides a context of identity of previous layer to the next layer. Identity mapping improves the training process by mitigating vanishing gradient problems. Besides ResNet, we evaluate InceptionV3 model [22], which uses factorized convolution, batch normalization and label smoothing to improve the performance of a highly deep network.

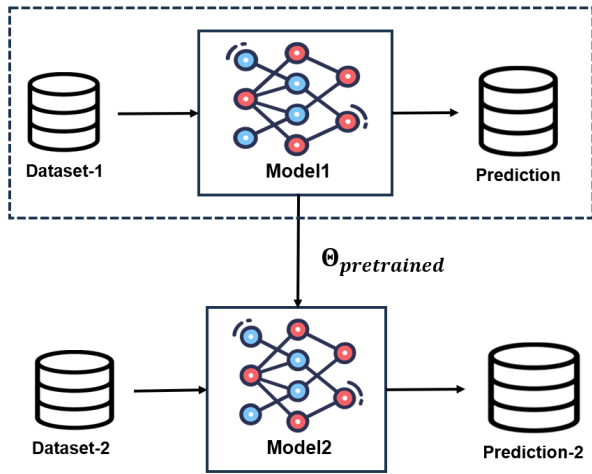


Fig. 1. Schematic diagram of transfer learning

III. RESULTS AND DISCUSSION

In this section, we conducted extensive experiments to evaluate the performance of most widely used TL models for medical image classification, including AlexNet, VGG16, different variants of ResNet model, and InceptionV3 model. We consider three variants of the ResNet model: ResNet-18, ResNet-34 and ResNet-50. Using these ResNet variants in the comparison helps to understand the performance variation of models with the depth of a neural network.

A. Exploratory Analysis on the Dataset

The dataset used to evaluate various TL models consists of chest X-ray images showing pneumonia, tuberculosis, and normal conditions, which is sourced from Kaggle's publicly available datasets [23]. We split the dataset into training, validation and testing sets after shuffling the data to provide generalized training. The number of data points in training, validation and testing sets are 13028, 761 and 1527, respectively. In the training set, we have maintained identical distribution among different classes. For example, normal, pneumonia, tuberculosis and unknown classes in the training set have 4667, 3633, 3573 and 1155 data points, respectively. The unknown class contains images where there are some anomalies, and the disease condition is not well represented. Rather than discarding these images from the dataset, we have included them as a separate class. This enhances the model's ability to distinguish anomalous conditions from both normal and known disease states. The original images in the dataset had a shape of (1272, 1040), which were resized to (224, 224) during modelling for computational efficiency. Additionally, we converted the dataset to grayscale before modelling, which has only one channel. Besides these primary transformations, Random Horizontal Flip, Random Rotation, and Normalization were also applied to provide diversity to the data distribution and increase the model robustness to external perturbations.

TABLE I

PERFORMANCE COMPARISON OF VARIOUS MODELS ON TRAINING SET

Model	Precision	Recall	Accuracy	F1 Score
AlexNet	0.931	0.938	0.926	0.93
VGG16	0.957	0.961	0.957	0.951
ResNet18	0.976	0.961	0.972	0.968
ResNet34	0.983	0.98	0.974	0.979
ResNet50	0.992	0.984	0.981	0.99
InceptionV3	0.994	0.989	0.99	0.991

TABLE II

PERFORMANCE COMPARISON OF VARIOUS MODELS ON TESTING SET

Model	Precision	Recall	Accuracy	F1 Score
AlexNet	0.922	0.918	0.926	0.927
VGG16	0.935	0.93	0.941	0.939
ResNet18	0.961	0.955	0.964	0.952
ResNet34	0.977	0.962	0.968	0.97
ResNet50	0.98	0.976	0.97	0.98
InceptionV3	0.987	0.981	0.989	0.985

B. Experiment Configuration

For training the models, we first froze the initial layers of the feature extractor and trained the final set of layers. Since we are using TL, it is assumed that initial layers are capable of extracting generic features. For training each model, we used Adam Optimizer [24] with a learning rate of 0.001, and a batch size of 32. All the models in the case study are convolutional models. The ResNet models have a common hyperparameter configuration. The kernel size for each residual model in the core feature extraction block is set to (3, 3), whereas the global average pooling layers use a kernel size of (7, 7). These architectures were pre-trained on ImageNet dataset and stored as ".pt" files in pytorch framework. Similarly, for Inception model building blocks, the kernel size and strides are set to (3, 3) and (2, 2), respectively. Inception model concatenates the output from multiscale transformation of the input and creates an output with different types of features. Additionally, we also analyzed the parameter pool size of each model. ResNet-18, ResNet-34, ResNet-50, and InceptionV3 model have 11.18M, 21.3M, 23.5M and 25.12M parameters, respectively. With such a huge number of parameters, they are highly computationally intensive and time consuming to train. TL freezes the pre-trained features extractor, significantly saving time and resource of intensive training. We used a global average pooling layer to aggregate features from the pre-trained model, which are then passed through a feedforward network (FFN) for classification. The number of hidden nodes follow the sequence 512 – 128 – 64 – 64 – 2. Since a standard FFN is used for all models, we fixed its structure to ensure consistency and avoid ambiguity in comparisons.

Python was used for code implementation. The models were trained on Google Colab using an NVIDIA T4 GPU system. We implemented the code using PyTorch libraries with additional analysis performed using NumPy, Pandas, Matplotlib, and other supporting libraries.

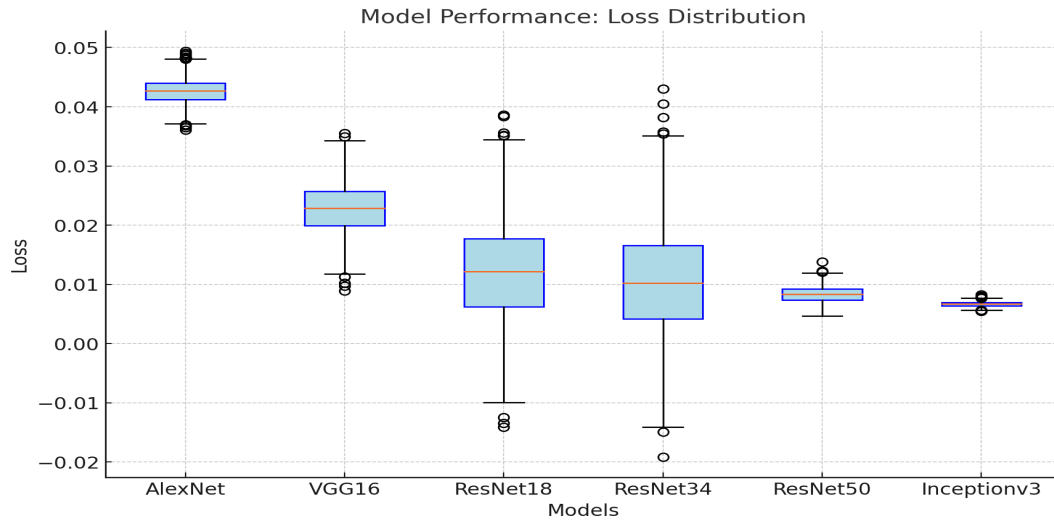


Fig. 2. Performance variations of various models

C. Evaluation Metrics

We evaluate the prediction performance of different models for classification of medical images using well established metrics: Precision, Recall, Accuracy, and F1 Score. These evaluation metrics are derived from the confusion matrix, consisting of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Each metric reflects a specific aspect of the model's predictive capability.

Precision measures the proportion of correctly identified positive cases among all the predicted positive cases. A higher precision value indicates fewer FP and is crucial for applications where minimizing incorrect positive predictions is essential, such as in disease diagnosis. Recall, on the other hand, quantifies the proportion of actual positive cases correctly identified by the model. It is vital in scenarios where missing positive instances (FN) can have significant consequences, such as failing to detect a critical medical condition. Accuracy provides an overall measure of the model's performance by calculating the proportion of correctly classified cases, both positive and negative, out of the total number of cases. While accuracy offers a comprehensive view, it may be less reliable in the presence of imbalanced datasets. The F1 Score serves as the harmonic mean of Precision and Recall, balancing the trade-off between these two metrics. This metric is particularly beneficial for datasets with imbalanced classes, as it ensures that neither precision nor recall is disproportionately favored.

D. Model Comparison Analysis

The results of training and testing performance of various TL models are presented in Tables I and II, respectively. It can be observed from Tables I and II that InceptionV3 outperforms other models across all the metrics, achieving highest Precision, Recall, Accuracy, and F1 Score during both the training and testing phases. Specifically, InceptionV3 achieves a testing precision of 0.987, recall of 0.981, accuracy of 0.989, and F1 Score of 0.985. This suggests that

InceptionV3 excels in maintaining a balance between minimizing false positives and false negatives, making it highly suitable for medical image classification task. ResNet50 also demonstrated strong performance, with a testing precision of 0.980, recall of 0.976, accuracy of 0.970, and F1 Score of 0.980, showing its effectiveness in accurately classifying medical images. Among remaining models, ResNet34 shows better performance than ResNet18, with higher values for precision, recall, accuracy, and F1 Score. ResNet34 achieves a testing precision of 0.977 and an F1 Score of 0.970, whereas ResNet18 attains slightly lower values (precision: 0.961, F1 Score: 0.952) during testing. These results indicate that deeper architectures, such as ResNet50 and InceptionV3, are more capable of capturing complex patterns in medical images compared to shallower architectures. VGG16 and AlexNet, though performing reasonably well, show lower accuracy. This difference highlights the rapid progress in DL model design and the benefits of more sophisticated architectures for medical image analysis.

E. Uncertainty Analysis

To assess the robustness and reliability of these comparative models, we conducted an uncertainty analysis during the validation phase. We considered multiple random splits of the model and repeated the same analysis numerous times for each realization of train-validation-test set to observe how the model is affected by the inherent variability of the neural network architecture. The results are illustrated in Fig. 2. InceptionV3 demonstrates the lowest mean loss (0.0067) and standard deviation of loss (0.0004), indicating not only superior performance but also high consistency and reliability. The ResNet models shows progressively lower mean loss and standard deviation as the depth increased, consistent with their performance metrics. Interestingly, while AlexNet and VGG16 have higher mean losses, they show relatively low standard deviations, suggesting consistent (albeit lower) performance across validation runs.

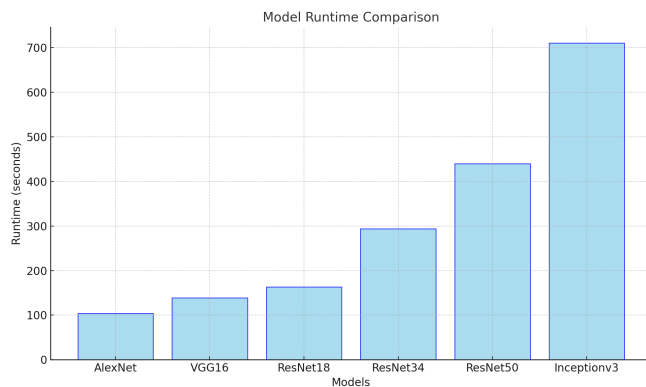


Fig. 3. Runtime comparison

F. Runtime Comparison

We also evaluate the computational efficiency of each model by measuring their validation runtime. The results are visualized in Fig. 3. The runtime analysis reveals a trade-off between model performance and complexity. InceptionV3, while providing the best performance, also requires the longest runtime (711 seconds). In contrast, simpler models like AlexNet and VGG16 are significantly faster but at the cost of lower accuracy. The ResNet family shows a clear correlation between depth and runtime, with ResNet50 taking nearly three times as long as ResNet18. This relationship highlights the need to balance model performance with available computational resources in practical applications.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we evaluate the performance of the most widely used TL models for medical image classification. TL offers numerous advantages in medical image analysis, including improved convergence, faster training, computational efficiency, generalization ability, and address class imbalance and data scarcity. Experimental results reveal that InceptionV3 offers the best overall performance in terms of accuracy, precision, recall, and F1 score. It also demonstrates the highest consistency and reliability as evidenced by its low uncertainty metrics. However, this superior performance comes at the cost of increased computational requirements. The ResNet family of models, particularly ResNet50, offers a good balance between performance and computational efficiency, making them suitable for a wide range of applications. AlexNet and VGG16, though less accurate, may still be viable options in scenarios where computational resources are limited and slightly lower accuracy is acceptable.

These findings provide valuable insights for researchers in the field of medical image analysis, aiding model selection based on specific needs for accuracy, reliability, and computational efficiency. Future work should explore optimization techniques for complex models to achieve faster inference with minimal accuracy loss, enhancing their viability in resource-constrained environments.

REFERENCES

- [1] J. Zhang, Y. Xie, Q. Wu and Y. Xia, "Medical image classification using synergic deep learning," *Med. Image Anal.*, vol. 54, pp. 10–19, 2019.
- [2] S. Singh and B.K. Tripathi, "Pneumonia classification using quaternion deep learning," *Multimed. Tools Appl.*, vol. 81, pp. 1743–1764, 2022.
- [3] W.M. Salama, A. Shokry and M.H. Aly, "A generalized framework for lung Cancer classification based on deep generative models," *Multimed. Tools Appl.*, vol. 81, pp. 32705–32722, 2022.
- [4] M. Fradi, L. Khriji and M. Machhout, "Real-time arrhythmia heart disease detection system using CNN architecture based various optimizers-networks," *Multimed. Tools Appl.*, vol. 81, pp. 41711–41732, 2022.
- [5] G. Sun, X. Wang, L. Xu, C. Li, W. Wang, Z. Yi, H. Luo, Y. Su, J. Zheng, Z. Li, Z. Chen, H. Zheng and C. Chen, "Deep Learning for the detection of multiple fundus diseases using ultra-widefield images," *Ophthalmol. Ther.*, vol. 12, pp. 895–907, 2023.
- [6] M.M. Hasan, M.U. Islam, M.J. Sadeq, W.-K. Fung and J. Uddin, "Review on the evaluation and development of artificial intelligence for COVID-19 containment," *Sensors*, vol. 23, p. 527, 2023.
- [7] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
- [8] M.T. Augustine, "A survey on universal approximation theorems," *arXiv preprint arXiv:2407.12895*, 2024.
- [9] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J. A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [10] A. Chowdhury, J. Rosenthal, J. Waring, R. Umeton, "Applying self-supervised learning to medicine: review of the state of the art and medical implementations," *Informatics*, vol. 8, no.3, p. 59, 2021.
- [11] J. Zhang, C. Li, M.M. Rahaman, Y. Yao, P. Ma, J. Zhang, X. Zhao, T. Jiang and M. Grzegorzczek, "A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches," *Artif. Intell. Rev.*, vol. 55, pp. 2875–2944, 2022.
- [12] M.M. Rahaman, C. Li, X. Wu, Y. Yao, Z. Hu and T. Jiang, "A survey for cervical cytopathology image analysis using deep learning," *IEEE Access*, vol. 8, pp. 61687–61710, 2020.
- [13] D. Agarwal, G. Marques, I. de la Torre-Díez, M.A. Franco Martin, B. García Zapirain and F. Martín Rodríguez, "Transfer learning for alzheimer's disease through neuroimaging biomarkers: a systematic review," *Sensors*, vol. 21, no. 21, p. 7259, 2021.
- [14] H.E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M.E. Maros and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med. Imag.*, vol. 22, p. 69, 2022.
- [15] D.T. Huff, A.J. Weisman, and R. Jeraj, "Interpretation and visualization techniques for deep learning models in medical imaging," *Phys. Med. Biol.*, vol. 66, no. 4, p. 04TR01, 2021.
- [16] H.-M. Zhang, and B. Dong, "A review on deep learning in medical image reconstruction," *J. Oper. Res. Soc. China*, vol. 8, no. 2, pp. 311–340, 2020.
- [17] C. Mosquera, L. Ferrer, D.H. Milone, D. Luna and E. Ferrante, "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance," *Eur. Radiol.*, vol. 34, no. 12, pp. 7895–7903, 2024.
- [18] M.Z. Alam, T. Roy, H.M.N. Kawsar, I. Rimi, "Enhancing transfer learning for medical image classification with SMOTE: A comparative study," in *Proc. Int. Conf. Computer Inf. Tech.*, Cox's Bazar, Bangladesh, 2024.
- [19] K. He, X. Zhang, S. Ren and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 630–645.
- [20] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, Nevada, USA, 2012.
- [21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.
- [22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 2818–2826.
- [23] R.I. Majumder, "Pneumonia & tuberculosis with normal & non-X-ray [Data set]," *Kaggle*, 2024. doi:10.34740/KAGGLE/DSV/10269174.
- [24] Z. Zhang, "Improved adam optimizer for deep neural networks," in *Proc. IEEE/ACM 26th Int. Symp. Qual. Service*, Banff, AB, Canada, 2018, pp. 1–2.