Assessing Transfer Learning Models for Medical Image Classification: A Comparative Study on Alzheimer's MRI, Chest CT-Scan, and Chest X-ray Images

Ryan Marcus Jeremy M. Lupague, Romie C. Mabborang, Alvin G. Bansil, Melinda M. Lupague

Abstract: ***Deep learning has revolutionized the field of neural network models, offering limitless applications in various domains. This study focuses on Transfer Learning (TL), a technique leveraging pre-trained deep learning models trained on large datasets for image classification tasks. Specifically, this research explores the effectiveness of various transfer learning models in three medical image datasets: Alzheimer's MRI images, Chest CT-Scan images, and Chest X-ray images. The main objective of this study is to assess and compare the performance of various TL models, including MobileNetV2, ResNet50, Xception, and InceptionV3, on the three medical image datasets. Additionally, a customized Convolutional Neural Network (CNN) model is developed to compare its performance against the pre-trained TL models. Each model was trained and evaluated on the three medical image datasets. The performance of the TL models was compared in terms of accuracy and training time. The results of this study revealed that ResNet50 consistently outperforms other TL models, demonstrating accurate predictions at the expense of longer training times. MobileNetV2 and InceptionV3 exhibit the fastest training times across all datasets, but they demonstrate poorer performance in certain datasets. The developed CNN model performs poorly in terms of accuracy and tends to overfit, indicating that creating a CNN model for medical image classification is not feasible in this study. The findings of this study offer valuable insights into the performance of TL models in medical image datasets. Researchers can utilize this information to make informed decisions when selecting TL models for medical imaging applications. Understanding the strengths and weaknesses of different TL models enhances the potential for accurate and efficient medical image classification. The insights gained from this study contribute to researchers' understanding of selecting transfer learning models for medical imaging applications, aiding in the advancement of medical image analysis and diagnosis.***

Keywords: ***Convolutional Neural Networks, Deep Learning, Medical Image Classification, Transfer Learning Models, ResNet50.***

Manuscript received on 13 August 2023 | Revised Manuscript received on 04 September 2023 | Manuscript Accepted on 15 September 2023 | Manuscript published on 30 September 2023.
***Correspondence Author(s)**

Ryan Marcus Jeremy M. Lupague, Department of Mathematics, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines. E-mail: rmjmlupague2019@plm.edu.ph, ORCID ID: 0009-0003-6648-8803

Romie C. Mabborang*, Department of Mathematics, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines. E-mail: rcmabborang@yahoo.com, ORCID ID: 0000-0002-3716-9673

Prof. Alvin G. Bansil, Department of Mathematics, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines. E-mail: agbansil@plm.edu.ph, ORCID ID: 0009-0002-0593-9773

Melinda M. Lupague, Department of Mathematics, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila, Philippines. E-mail: mmlupague@plm.edu.ph

I. INTRODUCTION

In recent years, there has been a growing interest in medical image classification in the field of healthcare and biomedical research. Accurately classifying medical images is crucial for diagnosing and treating various diseases. Transfer learning, a technique that utilizes pre-trained deep neural networks to extract relevant features from images, has shown promise in addressing the challenges associated with medical image classification. While previous studies have explored transfer learning in specific medical image modalities, there is a lack of comprehensive comparative analysis across multiple modalities and a limited understanding of the impact of different convolutional neural network (CNN) architectures and training dataset sizes on model performance. This research study aims to bridge this gap by conducting a comprehensive comparative analysis of transfer learning models for medical image classification, specifically focusing on Alzheimer's MRI, chest CT-scan, and chest X-ray images. The study aims to address several research questions, including the performance comparison of transfer learning models with a CNN model trained from scratch, the influence of different CNN architectures on transfer learning performance, and the impact of training dataset size on model performance. By addressing these research questions, this study significantly contributes to the existing literature by providing valuable insights into the optimal configuration for medical image classification tasks. It also enhances our understanding of the generalizability and robustness of transfer learning across different medical image datasets. Additionally, the findings of this study will have practical implications for the development of more accurate and efficient medical image classification systems, ultimately improving diagnostic accuracy and patient care in healthcare settings.

Retrieval Number: 100.1/ijrte.C78970912323
DOI: 10.35940/ijrte.C7897.0912323
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

In conclusion, this comprehensive comparative study aims to advance the field of medical image classification by evaluating transfer learning models, comparing them to CNN models trained from scratch, and exploring the influence of CNN architectures and training dataset sizes. The results of this study provide valuable insights and recommendations for optimizing the performance of medical image classification systems, leading to improved healthcare outcomes.

II. LITERATURE REVIEW

Transfer learning has revolutionized the field of medical image classification by enabling the utilization of pre-trained models to improve performance on specific tasks. This review synthesizes relevant literature on transfer learning in medical image classification, highlighting its effectiveness, previous comparative studies, and the gaps that the current research study aims to address. In a study published by [1], the paper highlighted the differences between the ImageNet dataset and the medical image datasets:

1) During the processing of medical image datasets, the initial stage of training deep learning models involves analyzing specific regions within an image and utilizing local texture variances to detect abnormalities. To illustrate this, the study utilized retinal fundus photographs as a case study, where a fundus camera captures an image of the eye's retina. In this specific instance, as illustrated in Figure 1, the existence of small crimson 'spots' indicates the presence of microaneurysms and diabetic retinopathy. This was utilized in the study by [2]. This varies from datasets containing natural images like ImageNet, where the images generally exhibit a clear overall subject, as depicted in Figure 2. This concept was employed by [3] in their work.

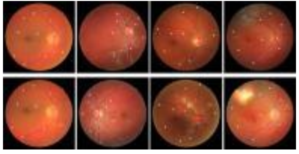

Figure 1. Illustrative Retinal Fundus Images from Medical Image Dataset

Figure 2. Illustrative Images from the Image Net Dataset

The study added that there is an open question on how much ImageNet feature reuse is helpful for these medical image classification problems.

2) The number of images in ImageNet, which has roughly one million plus images, is significantly larger than the number of images in any medical images dataset, which is ranging from several thousand images to a couple hundred thousand.

3) Lastly, medical images in a certain category have fewer classes than the standard ImageNet classification setup of 1000 classes. The standard ImageNet architecture has a lot of parameters in its layers of the convolutional network. With that, the design of the models pre-trained in the ImageNet architecture is likely to be suboptimal when integrated into the medical setting.

The study evaluated the performance of the deep learning models using standard architectures for natural images such as ImageNet and the non-standard but smaller and simpler models, on two large-scale medical imaging tasks, for which transfer learning is currently the norm. The study found the following:

1) In all those cases, the application of transfer learning did not significantly help the performance.

2) Smaller, simpler convolutional architectures performed comparably to standard Image NeT models.

3) ImageNet performance was not predictive of medical performance. These conclusions also hold in the very small data regime.

4) The use of pre-training the weights affected the hidden representations, but there was an issue with the model size, where the large, standard ImageNet models did not change significantly through the fine-tuning process.

5) There were feature-independent benefits to pretraining. Reusing only the scaling of the pre-trained weights but not the features can lead to large gains in convergence speed.

Transfer learning has been widely explored in the context of medical image classification tasks. [4] applied transfer learning to chest X-ray analysis, demonstrating the potential of deep learning algorithms in detecting abnormalities. They achieved performance comparable to expert radiologists by fine-tuning pre-trained models, such as Dense Net, on a large-scale dataset. However, their study focused solely on chest X-rays and did not compare different transfer learning models nor assess the impact of CNN architectures.

In the realm of medical image analysis, a comprehensive survey on transfer learning was undertaken, as documented by [5]. This survey placed a particular emphasis on the utilization of transfer learning within semi-supervised and multi-instance learning contexts. Notably, the authors shed light on the advantageous aspects of employing transfer learning to harness knowledge derived from natural image datasets, thereby enhancing histopathological image classification endeavors. Although this survey furnished valuable perspectives, it was observed that the specific research inquiries involving the comparison of transfer learning models, the assessment of their performance vis-à-vis CNN models developed from the ground up, and the exploration of the effects stemming from various CNN architectures were not directly addressed within its purview.

The significance of precise medical image segmentation using deep convolutional neural networks (CNNs) was explored by [6], who underscored the challenges inherent in this task. They also underscored the increasing achievements of deep neural models within this domain. In a divergent vein, the current investigation focuses on the assessment of transfer learning models within the realm of medical image classification.

This assessment is particularly targeted at Alzheimer's MRI, chest CT-Scan, and chest X-ray images.

While both studies share an interest in medical image analysis, they address different aspects and methodologies. [6] focuses on deep CNNs for segmentation, while the comparative study assesses transfer learning models for classification, emphasizing performance, efficiency, and potential improvements.

Hossain et al. [7] found that lightweight CNN architectures were effective for developing pre-scanner mobile applications for Lyme disease. Notably, ResNet50 and EfficientNetB0 demonstrated strong diagnostic performance. These trained models facilitate transfer learning and support researchers and practitioners in building pre-scanners for Lyme disease. CNNs, particularly lightweight ones, show promise in enhancing Lyme disease diagnosis and management, offering valuable diagnostic support for non-expert practitioners. The study contributes by showcasing CNNs' potential and providing practical model selection recommendations for resource-limited settings.

Zhang et al. [8] demonstrated promising results in steel surface defect recognition using transfer learning, particularly TL-ResNet50. The study showcased TL-ResNet50's superiority over other neural network models, indicating its practical potential for defect identification. The combination of the Adam optimizer and a learning rate decay strategy facilitated efficient model convergence. Additionally, the study validated CNN models' effectiveness and reliability in recognizing various defect types through an interpretable algorithm. Future research can explore TL-ResNet50's applicability in other domains and investigate different interpretability techniques for deeper insights into model predictions. This study contributes to the advancement of intelligent defect detection methods through transfer learning and deep learning models.

In their empirical investigation, Kondaveeti and Edupuganti [9] conducted a thorough examination of skin lesion classification through the utilization of a customized ResNet50 model. The outcomes of their study yielded impressive results, revealing the model's proficiency in accurately identifying diverse categories of skin lesions. Additionally, the model demonstrated effective differentiation capabilities between malignant and benign lesions, as evidenced by its scores in weighted average precision and recall. These observations suggest that the proposed model holds the potential to function as a significant asset within clinical decision support systems, furnishing dermatologists with preliminary evaluations that can direct subsequent diagnostic protocols.

In their investigation, Shah et al. [10] compiled a varied collection of skin lesion images from multiple origins and applied five Transfer Learning models. The primary goal was to determine the most precise model for categorizing Monkeypox skin lesions. Their discoveries highlighted MobileNetv2 as the leading performer, showcasing a notable level of accuracy. This model shows potential for the automated identification of Monkeypox outbreaks in areas where PCR tests are not readily available.

In their study, Jaradat et al. [11] compared five pre-trained deep learning models for mpox detection, with MobileNetV2 showing the highest classification performance. The results demonstrate the model's potential for accurate and efficient diagnosis, surpassing previous approaches in mpox image classification.

Quian Xiang et al. [12] showcased the feasibility of employing lightweight neural networks, particularly the MobileNetV2 model, for fruit image classification in resource-constrained environments. Their method, combining transfer learning, accurately recognized fruit images while efficiently handling computational and storage constraints. Implementing this cost-effective approach in robotic fruit-picking systems can offer substantial advantages to fruit producers.

In their research, Yilun Qin et al. [13] proposed a novel approach to identify moving loads in a vehicle-bridge system. Their method utilized continuous wavelet transform to create a time-frequency image from dynamic responses, forming a database for load identification. Transfer learning with a pre-trained MobileNetV2 model allowed establishing the mapping relationship between structural response and specific moving loads. This approach enabled successful load identification by inputting bridge responses into the established relationship.

Numerical simulations were carried out to validate the proposed method's efficacy. The results showcased its exceptional accuracy in identifying vehicle weight and speed, along with its remarkable robustness. MobileNetV2 outperformed other deep convolutional neural network models in terms of identification accuracy, training time, memory utilization, and speed. Particularly, the method demonstrated swift and accurate load identification, making it highly suitable for practical applications requiring efficient and precise load analysis.

Li et al. [14] undertook an investigation into the realm of transfer learning, with a focus on the evaluation of six commonly utilized convolutional neural network models. These models had been pre-trained on the ImageNet dataset. Within this range of models, MobileNetV2 was selected for further refinement efforts.

The process involved the implementation of strategies such as data augmentation, careful selection of hyperparameters, methods to mitigate overfitting, and dynamic adjustments to the learning rate. These enhancements culminated in an advanced iteration of the MobileNetV2 model, which exhibited a noteworthy level of accuracy in classification. This accuracy was complemented by swift image detection capabilities and a compact model size.

The experimental results distinctly underscored the superior performance of MobileNetV2 across multiple dimensions, particularly in terms of testing accuracy and detection speed. These findings aligned cohesively with the specified criteria for quality assessment within the production framework of Oudemansiella raphanipes.

Patel and Chaware [15] introduced an innovative method for diabetic retinopathy detection and classification, leveraging transfer learning and fine-tuning the MobileNetV2 model.

By adapting the pre-trained model to diabetic retinopathy's unique characteristics, the approach achieved remarkable improvements in accuracy. Rigorous evaluation of the Kaggle diabetic retinopathy dataset validated its effectiveness in accurately detecting and classifying the condition. These findings offer promising advancements in automated computational mechanisms for diabetic retinopathy, contributing to improved detection and assessment methods, with potential implications for blindness management and prevention.

Synthesis

The study published by [1] highlighted the differences between the ImageNet dataset and medical image datasets. Medical image datasets often require the analysis of specific regions within an image and the utilization of local texture variances to detect abnormalities. This differs from natural image datasets like ImageNet, which typically possess a distinct global subject. The study also raised questions about the helpfulness of ImageNet feature reuse in medical image classification problems due to the variations in dataset sizes and class categories. Furthermore, research by Hossain et al. [7] demonstrated that even lightweight CNN architectures can be effective for developing pre-scanner mobile applications for Lyme disease. Their findings showcased the notable performance of the ResNet50 architecture and the EfficientNetB0 model in diagnosing Lyme disease. These trained models enabled transfer learning and the development of pre-scanners for Lyme disease, offering valuable diagnostic support and potential for early detection.

Transfer learning has also been explored in other medical image classification tasks. For example, [4] applied transfer learning to chest X-ray analysis, achieving performance comparable to expert radiologists. [10] and [11] focused on the detection of Monkeypox skin lesions, with MobileNetV2 consistently demonstrating high accuracy. Other studies utilized transfer learning for fruit image classification [12], vehicle-bridge interaction system identification [13], and Oudemansiella raphanipes dataset classification [14], showcasing the effectiveness and practical applications of transfer learning in different domains.

Compared to the existing literature, this research study significantly extends the current knowledge by conducting a comprehensive comparative analysis of transfer learning models for medical image classification. The study focuses on three distinct medical imaging modalities: Alzheimer's MRI, chest CT-scan, and chest X-ray images. By considering multiple modalities, this study addresses the variations and challenges specific to each modality, providing a more holistic assessment of transfer learning models. It narrows down the scope to TL models specifically for medical image classification, comparing their performance and exploring the impact of CNN architectures and the scalability of TL models under varying dataset sizes. Furthermore, this study goes beyond previous researches by directly comparing transfer learning models with a CNN model trained from scratch. It aims to evaluate classification accuracy and training time to assess the advantages and limitations of transfer learning in medical image classification tasks. Additionally, the study explores the impact of different CNN architectures on the performance of transfer learning models, allowing for an in-depth analysis of the influence of model architecture on transfer learning effectiveness. Lastly, this research investigates the effect of training dataset size on the performance of transfer learning models and the CNN trained from scratch. By varying the dataset size, the study aims to explore the scalability and robustness of transfer learning models under different data availability scenarios. The existing literature provides valuable insights into transfer learning models in medical image classification, demonstrating their potential in various domains and highlighting the need for further research and refinement. The present study contributes to this knowledge by conducting a comprehensive analysis, comparing TL models' performance, and offering insights for model selection in medical imaging tasks.

III. METHODS AND MATERIALS

A. Research Design

In this study, a comparative research design was adopted to assess and compare the performance of transfer learn- ing models for medical image classification. The research problems were addressed through a systematic evaluation of various transfer learning models and a CNN model trained from scratch. The study considered three medical imaging modalities: Alzheimer's MRI, chest CT-scan, and chest X-ray images.

B. Data Collection

The research study utilized publicly available datasets specific to each imaging modality. The Alzheimer's MRI dataset comprised a collection of MRI scans from patients with Alzheimer's disease and healthy controls. The chest CT-scan dataset consisted of CT scans of patients with various pulmonary conditions, and the chest X-ray dataset contained X- ray images of individuals with different respiratory conditions. The datasets were carefully curated to ensure data integrity and appropriate labels for classification tasks [7],[8], and [9]. The following are a few examples of the images contained within the medical datasets.

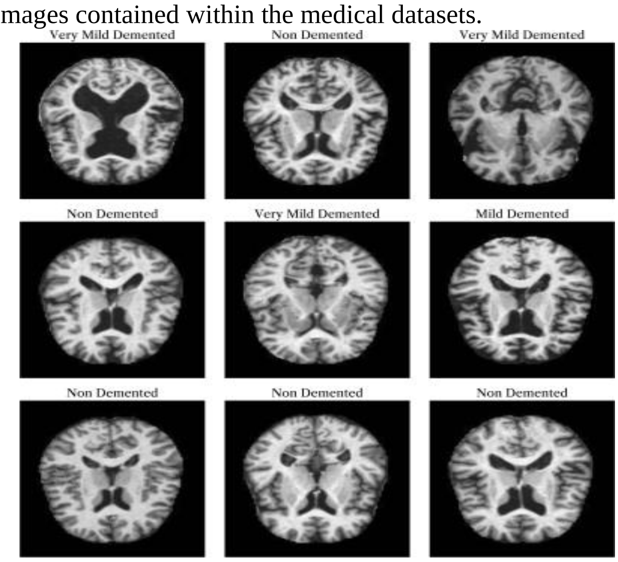

Figure 3. Images from Alzheimer's MRI Dataset

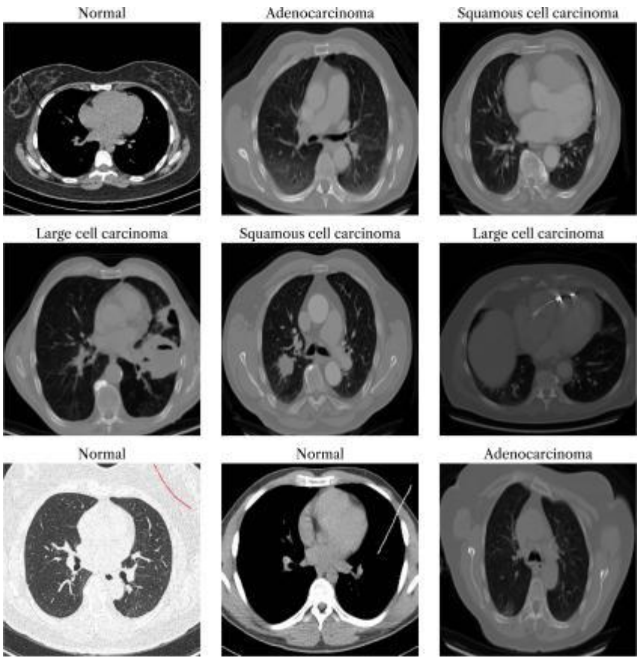

Figure 4. Images from Chest CT-Scan Dataset

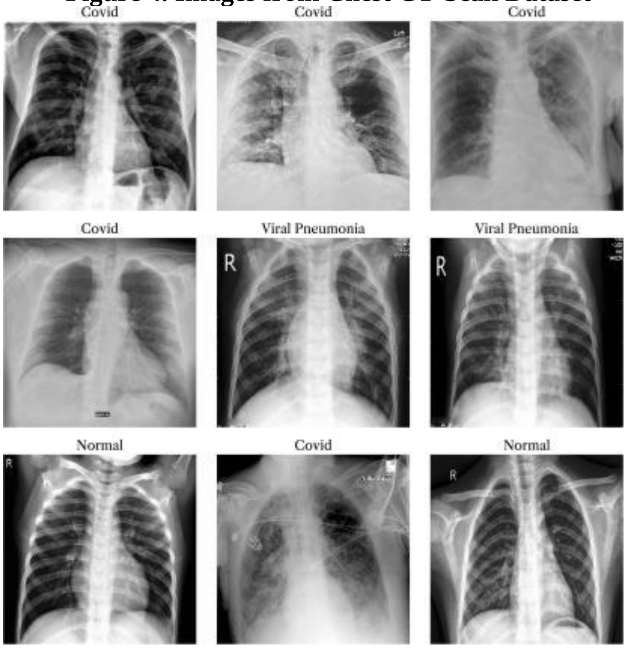

Figure 5. Images from Chest X-ray Dataset

C. Transfer Learning Models

Multiple transfer learning models were evaluated in this study, including popular architectures such as MobileNetV2, ResNet50, Xception, and InceptionV3. These models were pre-trained on large-scale general image datasets, such as ImageNet, and fine-tuned on the specific medical image datasets. The layers of the pre-trained models were modified and connected to task-specific output layers for classification.

D. CNN Trained from Scratch

To compare the performance of transfer learning models, CNN models were trained from scratch using the same architecture as the pre-trained models. The CNN models were initialized with random weights and optimized on the specific medical image datasets using appropriate training algorithms, such as stochastic gradient descent (SGD) or Adam.

E. Evaluation Metrics

The performance of the transfer learning models and the CNN trained from scratch was assessed using standard evaluation metrics for classification tasks. These metrics included accuracy, precision, recall, F1-score, and support. Additionally, the training time required for each model was recorded to compare the efficiency of transfer learning versus training from scratch.

IV. RESULTS AND DISCUSSION

This comparative study aims to understand the effect of transfer learning models in medical image classification tasks. Deep learning algorithms and transfer learning models were utilized on the Intel Core i7 8th generation machine with Windows 10 operating system. Python with deep learning packages such as TensorFlow, Keras, and Scikit-learn were used in implementing different transfer learning models.

In all the datasets used in this study, the performance of each model was measured by its accuracy and loss per epoch, which was set into 10 epochs as per researchers stated in the Methodology. The researchers analyzed all the model's performances and evaluated them to determine the best transfer learning model that can be utilized for medical image classification problems.

A. Alzheimer's MRI Dataset

Figures 6 to 10 present the training performance of each model in the Alzheimer's MRI Dataset. From Figures 6 to 10, it can be observed that ResNet50 exhibited the highest performance during training, whereas the own CNN model showed the poorest performance, with no improvement in training accuracy across epochs. Both InceptionV3 and the own CNN model demonstrate signs of overfitting, as the training accuracy continues to improve with each epoch while the validation accuracy remains relatively stagnant. It is important to note that the model was trained for a set number of 10 epochs, and further training for additional epochs would likely yield significant improvements.

Figures 11 to 15 display the Confusion Matrix of each model in the Alzheimer's MRI Dataset.

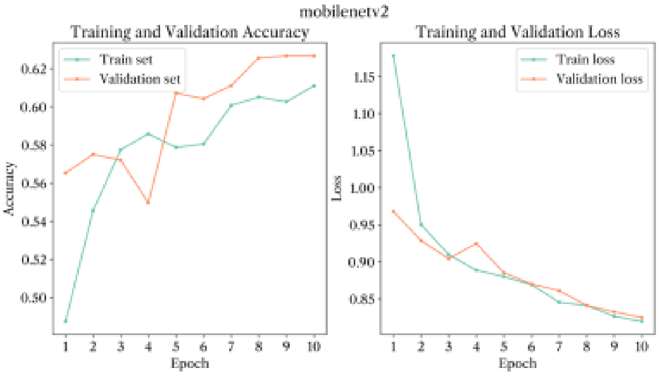

Figure 6. Training and Validation Accuracy per Epoch of MobileNetV2 in Alzheimer's MRI Dataset

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

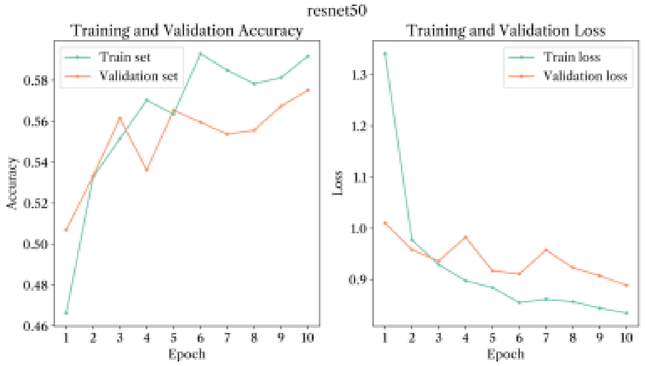

Figure 7. Training and Validation Accuracy per Epoch of ResNet50 in Alzheimer's MRI Dataset

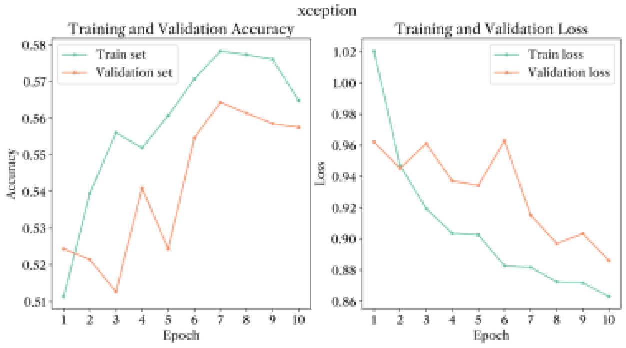

Figure 8. Training and Validation Accuracy per Epoch of Xception in Alzheimer's MRI Dataset

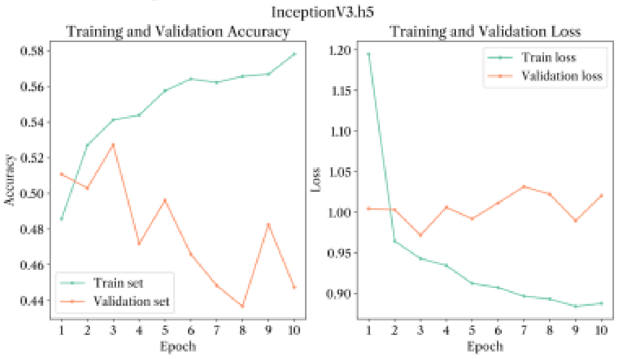

Figure 9. Training and Validation Accuracy per Epoch of InceptionV3 in Alzheimer's MRI Dataset

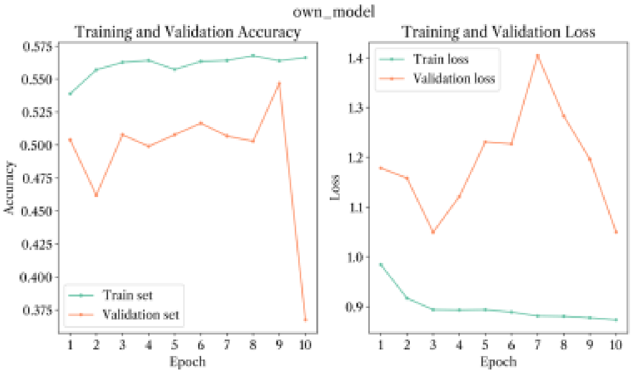

Figure 10. Training and Validation Accuracy per Epoch of Own CNN Model in Alzheimer's MRI Dataset

The subsequent matrices represent the Confusion Matrices derived from the Transfer Learning Models.

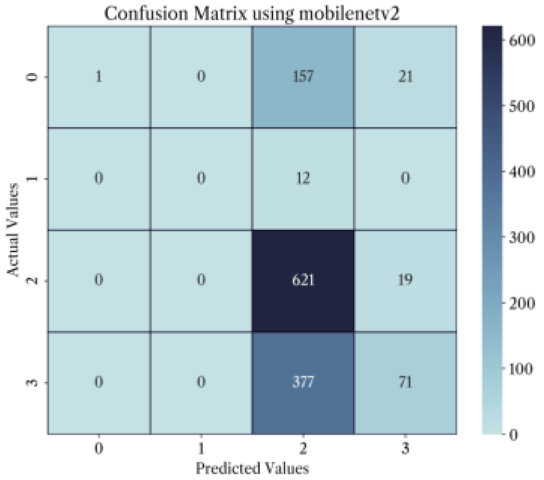

Figure 11. Confusion Matrix of MobileNetV2 in Alzheimer's MRI Dataset

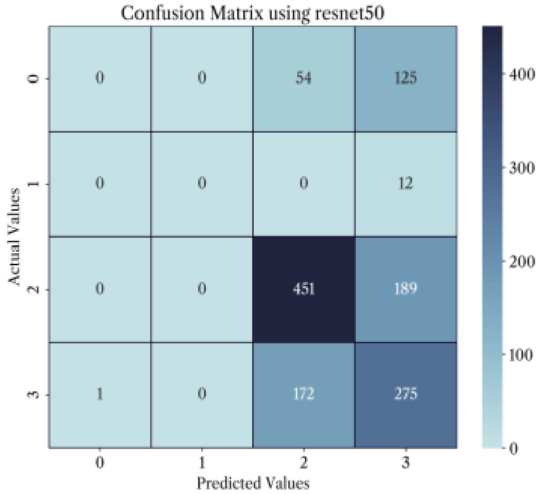

Figure 12. Confusion Matrix of ResNet50 in Alzheimer's MRI Dataset

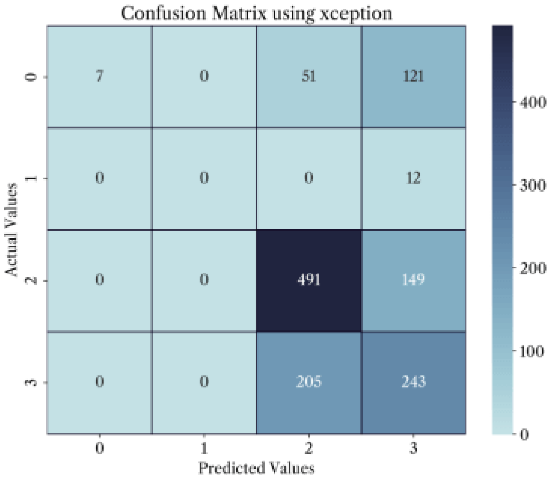

Figure 13. Confusion Matrix of Xception in Alzheimer's MRI Dataset

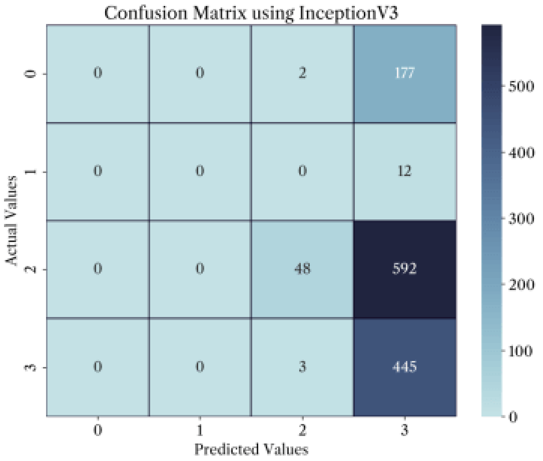

Figure 14. Confusion Matrix of InceptionV3 in Alzheimer's MRI Dataset

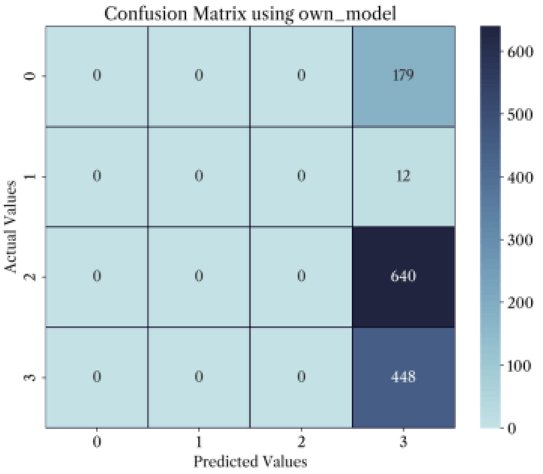

Figure 15. Confusion Matrix of Own CNN Model in Alzheimer's MRI Dataset

Using the Confusion Matrices, the study could easily visualize and assess that all models did not perform well with the use of transfer learning models. The models mostly captured the characteristics and features of the non-demented MRI images and not the other classes. However, transfer learning models performed better than the own CNN model, where the CNN model only got 37% accuracy in the test set.

Table 1. Train and Test Accuracy in Alzheimer's MRI Dataset

Transfer Learning Model	Training Time	Train Set Accuracy	Test Set Accuracy
MobileNetV2	00:24:50	0.629060	0.541830
ResNet50	01:16:43	0.584860	0.567631
Xception	01:17:07	0.594139	0.579359
InceptionV3	00:51:05	0.445421	0.385457
Own Model	01:54:02	0.370696	0.350274

Table I presents a comprehensive summary of the performance of different transfer learning models on the train set and test set, along with their respective training times. Among the evaluated models, Xception achieved the highest accuracy on the test set, reaching 57.9%. Following closely, ResNet50 attained an accuracy of 56.7%. Notably, the training duration for most models exceeded one hour, which can be attributed, in part, to the substantial number of images in the Alzheimer's MRI dataset. Specifically, the dataset comprised a total of 6400 images.

Among the models assessed, MobileNetV2 exhibited the shortest training time, taking only 24 minutes, while InceptionV3 required 51 minutes to train. In comparison, the best-performing model, Xception, necessitated 1 hour and 16 minutes to complete the training process. Consequently, although the transfer learning models yielded suboptimal results on the Alzheimer's MRI dataset, potential enhancements can be explored to improve their accuracy.

B. Chest CT-Scan Dataset

In the Chest CT-Scan Dataset, the performance of each model when training is shown in the following figures:

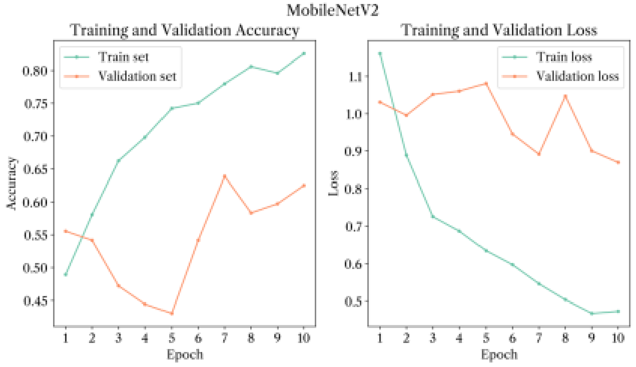

Figure 16. Training and Validation Accuracy per epoch of MobileNetV2 in CT-Scan Dataset

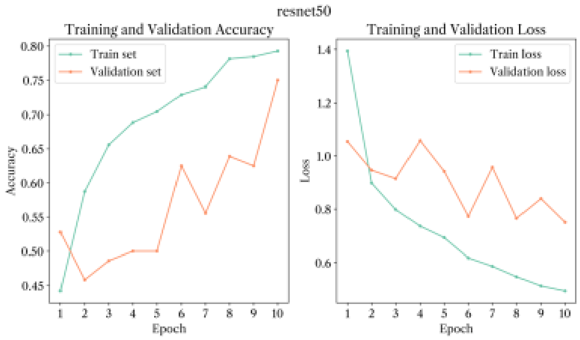

Figure 17. Training and Validation Accuracy per epoch of ResNet50 in CT-Scan Dataset

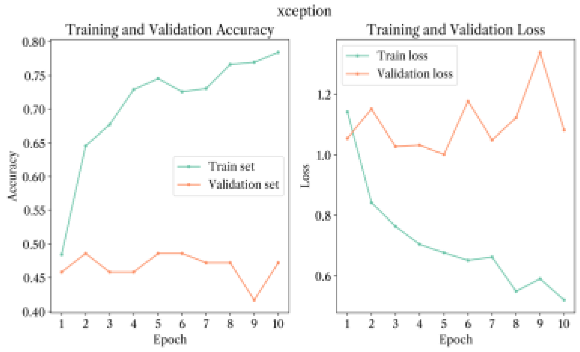

Figure 18. Training and Validation Accuracy per Epoch of Xception in CT-Scan Dataset

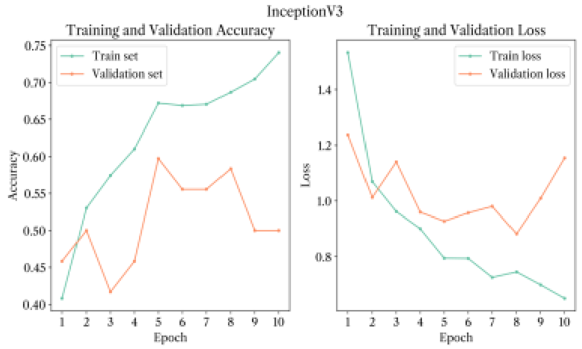

Figure 19. Training and Validation Accuracy per Epoch of InceptionV3 in CT- Scan Dataset

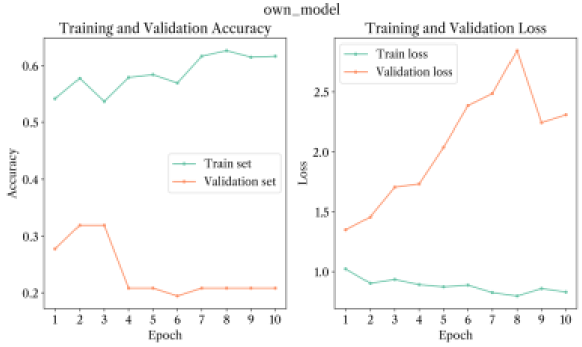

Figure 20. Training and Validation Accuracy per epoch of Own CNN model in CT-Scan Dataset

Based on Figures 16-20, ResNet50 and MobileNetV2 performed the best when training, while Xception and the own CNN model performed the worst, where the training accuracy has a large margin to the validation accuracy. Overfitting is also present in those models.

Figures 21-25 display the Confusion Matrices for each model in the Chest CT-Scan dataset. These matrices provide insight into how accurately each model predicted the classes. Notably, ResNet50 achieved an impressive 99% accuracy for the normal class and a 65% accuracy for squamous cell carcinoma.

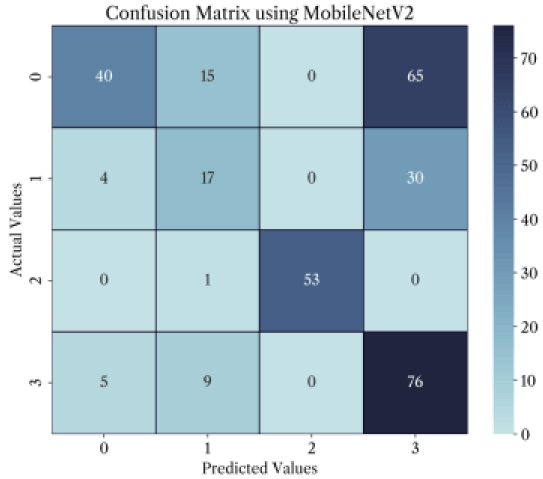

Figure 21. Confusion Matrix of MobileNetV2 in Chest CT-Scan Dataset

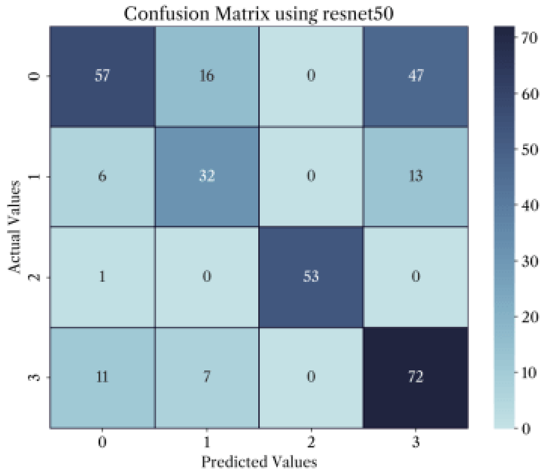

Figure 22. Confusion Matrix of ResNet50 in Chest CT-Scan Dataset

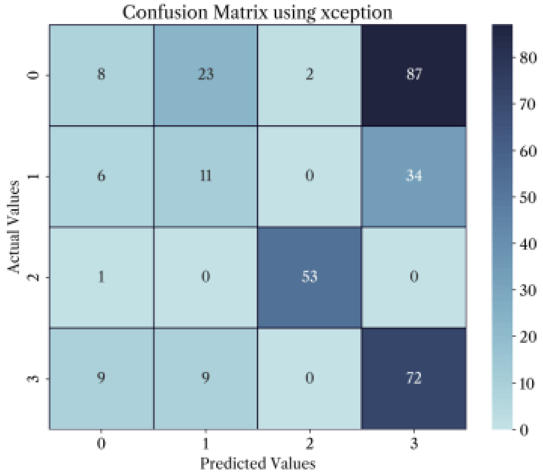

Figure 23. Confusion Matrix of Xception in Chest CT-Scan Dataset

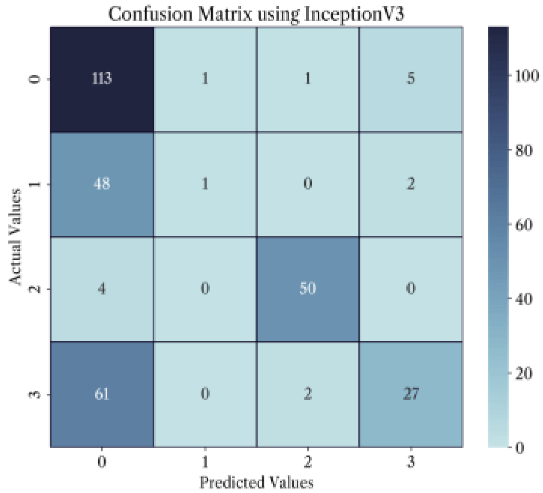

Figure 24. Confusion Matrix of InceptionV3 in Chest CT-Scan Dataset

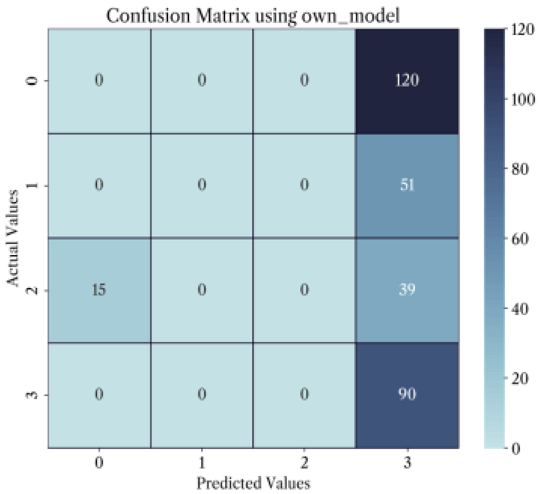

Figure 25. Confusion Matrix of Own CNN model in Chest CT-Scan Dataset

The Confusion Matrices show how many did each model predicted right per class. The ResNet50 in particular, predicted the normal class and the images that are classified as squamous cell carcinoma with 99% and 65% accuracy respectively.

TABLE II. Train and Test Accuracy in Chest Ct-Scan Dataset

Transfer Learning Model	Training Time	Train Set Accuracy	Test Set Accuracy
MobileNetV2	0:03:18	0.6803	0.5905
ResNet50	0:10:05	0.8091	0.6794
Xception	0:10:19	0.5693	0.4571
InceptionV3	0:06:15	0.6852	0.6063
Own Model	0:14:18	0.2529	0.2857

Table II shows the overall summary of the train set and test set scores with each model's time to train.

In the Chest CT-Scan dataset, ResNet50 performed the best with 67.9% accuracy, followed by InceptionV3 with 60.6%. Unlike in the Alzheimer's dataset, where it took the models more than 1 hour to train, the training in this dataset only takes from 3 minutes to 14 minutes. The reason for the time difference is the number of images that are being trained. The Chest CT-Scan dataset only contains a total of 1000 images that are classified into four classes. The model with the fastest training time is the MobileNetV2 again, with only 3 minutes, followed by the Inception V3, with 6 minutes. Overall, the transfer learning models performed just well in the Chest CT- Scan dataset. Other training factors can be changed to prevent the occurrence of overfitting.

C. Chest X-Ray Dataset

The figures presented below depict the training performances of the models in the Chest X-ray Dataset.

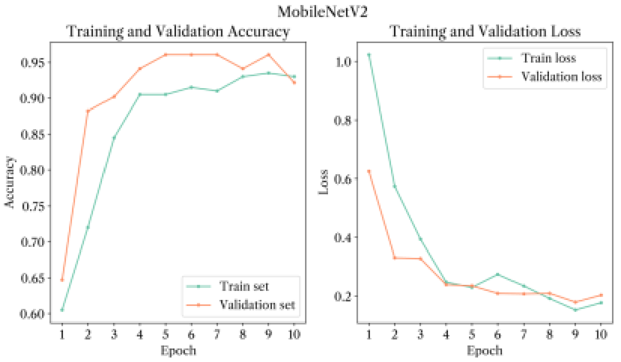

Figure 26. Training and Validation Accuracy per Epoch of MobileNetV2 in Chest X-ray Dataset

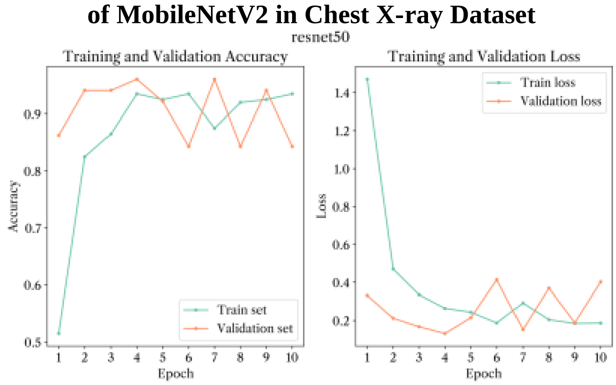

Figure 27. Training and Validation Accuracy per Epoch of ResNet50 in Chest X-ray Dataset

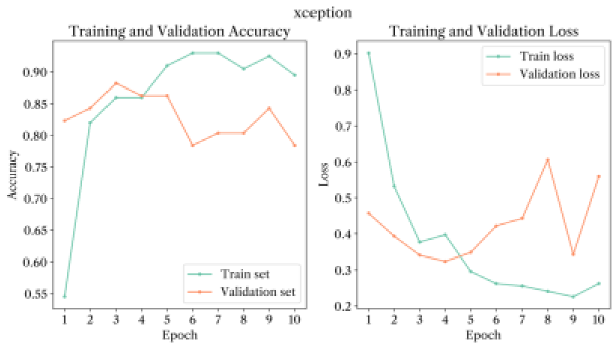

Figure 28. Training and Validation Accuracy per Epoch of Xception in Chest X-ray Dataset

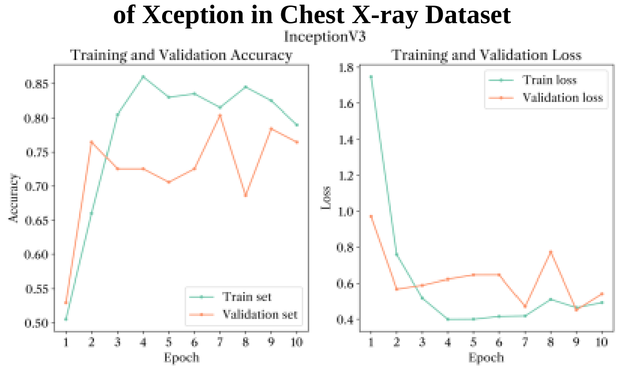

Figure 29. Training and Validation Accuracy per Epoch of InceptionV3 in Chest X-ray Dataset

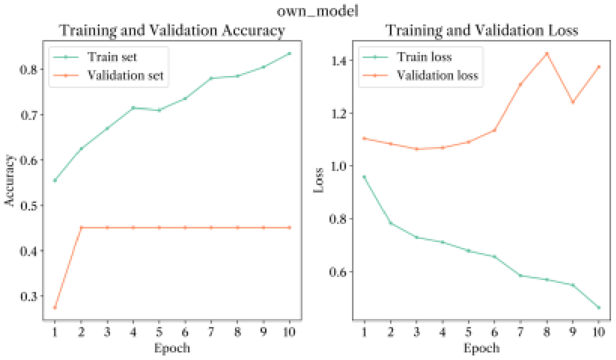

Figure 30. Training and Validation Accuracy per Epoch of Own CNN model in Chest X-ray Dataset

Figures 26 to 30 exhibit the notable performance of all Transfer learning models, displaying training and validation accuracies consistently surpassing 80% or higher. Conversely, the own CNN model exhibited inferior performance, displaying clear signs of overfitting as the validation accuracy remained stagnant from the second epoch onwards.

Figures 31 to 35 illustrate the Confusion Matrix of each model in the Chest X-ray dataset.

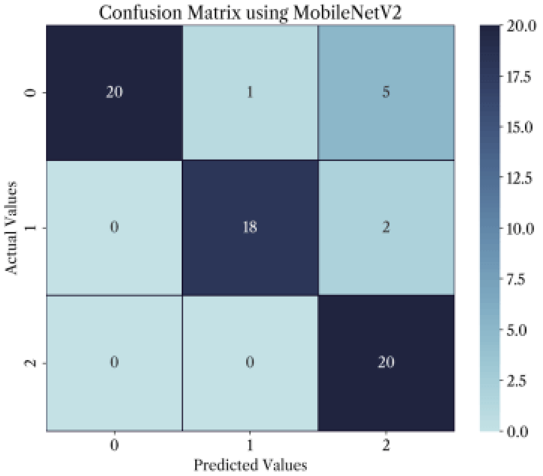

Figure 31. Confusion Matrix of MobileNetV2 in Chest X-ray Dataset

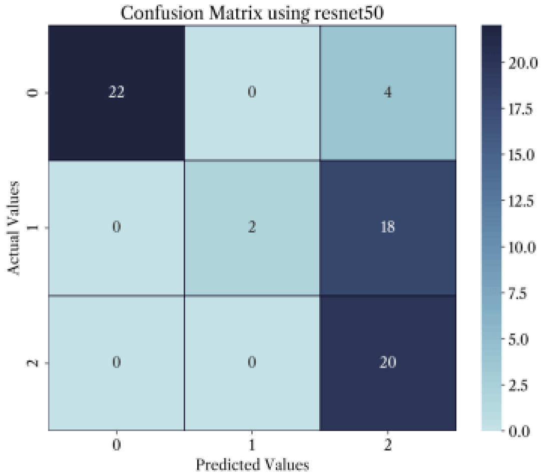

Figure 32. Confusion Matrix of ResNet50 in Chest X-ray Dataset

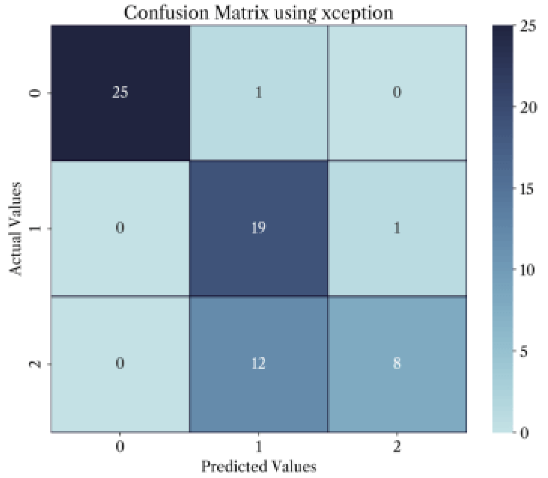

Figure 33. Confusion Matrix of Xception in Chest X-ray Dataset

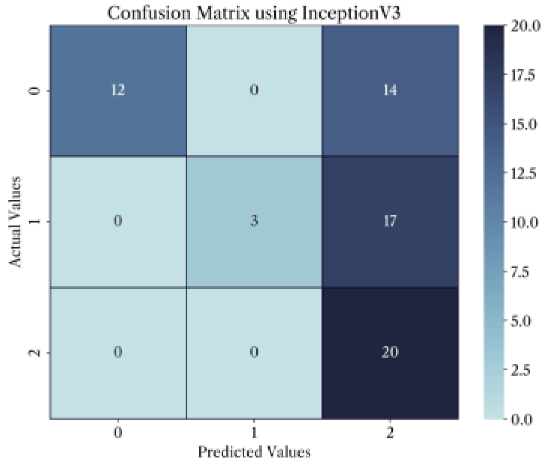

Figure 34. Confusion Matrix of InceptionV3 in Chest X-ray Dataset

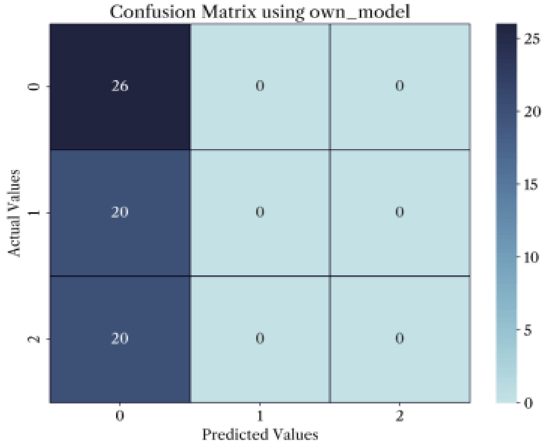

Figure 35. Confusion Matrix of Own CNN model in Chest X-ray Dataset

The Confusion Matrices demonstrated the number of predicted right per class. All transfer learning models have a 100% precision rate in the Covid class. In more straightforward terms, If any transfer learning model predicted an X-ray image to have Covid, it is 100% likely to be correct. MobileNetV2 performed the best in all classes.

Table III. Train and Test Accuracy in Covid X-Ray Dataset

Transfer Learning Model	Training Time	Train Set Accuracy	Test Set Accuracy
MobileNetV2	0:01:49	0.8400	0.8788
ResNet50	0:04:22	0.7650	0.6700
Xception	0:04:14	0.7800	0.7879
InceptionV3	0:02:45	0.6950	0.5303
Own Model	0:05:14	0.4400	0.3939

Table III shows the overall summary of the train set and test set scores with each model's time to train.

In table III, the best performing model is the MobileNetV2 with 87.9% accuracy, followed by Xception with 78.8%. The models only took a short amount of time to train since the Chest X-ray dataset only contains 317 images. The model with the fastest training time is the MobileNetV2, with only 1 minute and 49 seconds, followed by the InceptionV3, with 2 minutes and 45 seconds. The transfer learning models performed great in the Chest X-ray dataset, while the Own CNN model did poorly. The study evaluated the performance of different models in medical image classification using accuracy and loss over 10 epochs. In the Alzheimer's MRI dataset, ResNet50 performed the best, while the own CNN model showed poor performance with no improvement. Both InceptionV3 and the own CNN model exhibited signs of overfitting. Transfer learning models, particularly Xception, outperformed the own CNN model. In the Chest CT-scan dataset, ResNet50 and MobileNetV2 performed well, while Xception and the own CNN model struggled. ResNet50 achieved impressive accuracy for the normal class and squamous cell carcinoma. ResNet50 also performed the best in the overall Chest CT-scan dataset, followed by InceptionV3. In the Chest X-ray dataset, all transfer learning models showed notable performance, with MobileNetV2 performing the best. ResNet50 consistently performed well across all datasets, and MobileNetV2 had the fastest training speed. The study recommends prioritizing transfer learning models, especially ResNet50, for medical image classification tasks. The research contributes to existing literature by conducting a comprehensive analysis of transfer learning models and highlights the need for further refinement in prediction biases, training epochs, and model architectures.

V. CONCLUSION

The research study addresses the gap in the existing literature by conducting a comparative analysis of transfer learning models specifically for Alzheimer's MRI, Chest CT-scan, and Chest X-ray images. By evaluating different models, comparing them to a CNN trained from scratch, and investigating the impact of CNN architectures and training dataset sizes, this study aims to provide insights into selecting the most effective approach for medical image classification tasks. The findings provide significant insights into the performance and effectiveness of transfer learning models in the classification of medical image datasets. ResNet50 emerged as the most consistent performer, achieving accurate scores across all datasets. However, it requires more training time compared to other models, as ResNet50 trained for more than 1 hour on the Alzheimer's MRI dataset with 6400 images. Regarding training efficiency, MobileNetV2 demonstrated the fastest training speed, followed by InceptionV3. This held true across all three datasets, making them suitable for scenarios where training time is critical. On the other hand, the researchers' attempt to develop a custom CNN model from scratch yielded unsatisfactory results, consistently suffering from overfitting and limitations in generalization to unseen data. Therefore, the use of self-created CNN models is not recommended compared to transfer learning models.

Based on these findings, it is recommended to prioritize the use of transfer learning models, particularly ResNet50, for classification tasks involving similar datasets. ResNet50's consistency and accuracy make it a practical choice for researchers and practitioners. Additionally, the efficient training capabilities of MobileNetV2 and InceptionV3 make them suitable alternatives when time is a crucial consideration. In addition to the findings, an interesting trend was observed in the prediction behavior of the models across the Alzheimer's MRI dataset and the Chest CT-scan dataset. The models consistently showed a bias towards predicting the ”normal” class, raising concerns about their ability to effectively identify and classify images with significant medical implications. Further investigation and fine-tuning are necessary to address this bias and enhance the models' sensitivity toward detecting clinically relevant classes.

Furthermore, the choice of the number of training epochs emerged as a crucial factor in model performance. This study used ten epochs for training, which yielded acceptable results. However, increasing the number of epochs can further enhance model performance, considering the trade-off between performance improvement and increased training time. In conclusion, this study emphasizes the importance of selecting appropriate pre-trained models for specific classification tasks. The findings provide valuable insights for researchers and practitioners, enabling them to make informed decisions about model choices. The study also sheds light on the performance and behavior of transfer learning models in medical image classification tasks, highlighting the need for further investigation into addressing prediction biases, optimizing training epochs, and fine-tuning model architectures. These considerations are crucial for enhancing the models' diagnostic capabilities and their effective integration into real-world medical applications.

IMPLICATIONS

- Prioritizing Transfer Learning Models: The findings of this study highlight the superior performance and consistency of transfer learning models, particularly ResNet50, in the classification of medical image datasets. Researchers and practitioners should consider utilizing these pre-trained models as a practical choice for similar classification tasks due to their remarkable accuracy and consistency across datasets.
- Training Efficiency: The study reveals that MobileNetV2 and InceptionV3 exhibit faster training speeds compared to other models. These models can be valuable alternatives when time is a crucial consideration, allowing for efficient training without compromising performance.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

- Caution Against Self-Created CNN Models: The researchers attempt to develop a custom CNN model from scratch yielded unsatisfactory results due to consistent overfitting. As a result, it is cautioned against using self-created CNN models, as they did not demonstrate competitive performance compared to transfer learning models. Researchers should prioritize utilizing pre-trained models for improved classification accuracy.
- Addressing Prediction Biases: The analysis of the re-results revealed a bias in the prediction behavior of the models towards the ”normal” class, regardless of other classes associated with specific diagnoses or diseases. This raises concerns about the models' ability to effectively identify and classify images with significant medical implications. Further investigation and fine-tuning are necessary to address this bias and enhance the models' sensitivity toward detecting clinically relevant classes.
- Optimizing Training Epochs: The choice of the number of training epochs emerged as another crucial factor in model performance. While the study used ten epochs for training, increasing the number of epochs can further enhance model performance. However, the trade-off between performance improvement and increased training time should be carefully considered in practical implementation.

This study emphasizes the importance of selecting appropriate pre-trained models for specific medical image classification tasks. The findings provide valuable insights into the performance and feasibility of different models, guiding researchers and practitioners in making informed decisions. Further investigation is needed to address prediction biases, optimize training epochs, and fine-tune model architectures, ultimately enhancing the diagnostic capabilities of these models in real-world medical applications.

DECLARATION STATEMENT

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Yes, The data and materials connected to our research are publicly accessible. They guide us in our quest for knowledge. We provide information about specific datasets: The Alzheimer's dataset contains images categorized into four classes, managed by [16]. The Chest CT-Scan images Dataset, curated by [17], offers valuable insights
Authors Contributions	All authors having equal contribution for this

REFERENCES

1. M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” Advances in neural information processing systems, vol. 32, 2019.
2. C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. Argyros, “Fire: Fundus image registration dataset,” Journal for Modeling in Opthalmology, Jan. 2017. https://www.maio-journal.com/index.php/MAIO/article/view/42/90https://doi.org/10.35119/maio.v1i4.42
3. C. Chen, Y. Ren, and C.-C. Kuo, “Global-attributes assisted outdoor scene geometric labeling,” in Feb. 2016, pp. 93–120, ISBN: 978-981-10-0629-6. DOI: 10 . 1007 / 978-981-10-0631-9 5.
4. P. Rajpurkar, J. Irvin, R. L. Ball, et al., “Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists,” PLoS medicine, vol. 15, no. 11, e1002686, 2018. https://doi.org/10.1371/journal.pmed.1002686
5. V. Cheplygina, M. de Bruijne, and J. P. Pluim, “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” Medical image analysis, vol. 54, pp. 280–296, 2019. https://doi.org/10.1016/j.media.2019.03.009
6. P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, and W. Enbeyle, “Deep neural networks for medical image segmentation,” Journal of Healthcare Engineering, vol. 2022, 2022. https://doi.org/10.1155/2022/9580991
7. Sk Imran Hossain, Jocelyn de Goër de Herve, Md Shahriar Hassan, Delphine Martineau, Evelina Petrosyan, Violaine Corbin, Jean Beytout, Isabelle Lebert, Jonas Durand, Irene Carravieri, Annick Brun-Jacob, Pascale Frey-Klett, Elisabeth Baux, Céline Cazorla, Carole Eldin, Yves Hansmann, Solene Patrat-Delon, Thierry Prazuck, Alice Raffetin, Pierre Tattevin, Gwenaël Vourc'h, Olivier Lesens, Engelbert Mephu Nguifo, Exploring convolutional neural networks with transfer learning for diagnosing Lyme disease from skin lesion images, Computer Methods and Programs in Biomedicine, Volume 215, 2022,106624, ISSN 01692607,https://doi.org/10.1016/j.cmpb.2022.106624. https://www.sciencedirect.com/science/article/pii/S016260722000098) https://doi.org/10.1016/j.cmpb.2022.106624
8. Zhang L, Bian Y, Jiang P, Zhang F. A Transfer Residual Neural Network Based on ResNet-50 for Detection of Steel Surface Defects. Applied Sciences. 2023;13(9):5260. https://doi.org/10.3390/app13095260.
9. H. K. Kondaveeti and P. Edupuganti, "Skin Cancer Classification using Transfer Learning," 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI), Buldhana, India, 2020, pp. 1-4, doi: 10.1109/ICATMRI51801.2020.9398388. https://doi.org/10.1109/ICATMRI51801.2020.9398388
10. A. Shah, "Monkeypox Skin Lesion Classification Using Transfer Learning Approach," 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2022, pp. 1-5, doi: 10.1109/IBSSC56953.2022.10037374. https://doi.org/10.1109/IBSSC56953.2022.10037374
11. Jaradat AS, Al Mamlook RE, Almakayeel N, Alharbe N, Almuflih AS, Nasayreh A, Gharaibeh H, Gharaibeh M, Gharaibeh A, Bzizi H. Automated Monkeypox Skin Lesion Detection Using Deep Learning and Transfer Learning Techniques. Int J Environ Res Public Health. 2023 Mar 1;20(5):4422. doi: 10.3390/ijerph20054422. PMID: 36901430; PMCID: PMC10001976. https://doi.org/10.3390/ijerph20054422
12. Qian Xiang, Xiaodan Wang, Rui Li, Guoling Zhang, Jie Lai, Qingshuang Hu, CSAE '19: Proceedings of the 3rd International Conference on Computer Science and Application EngineeringOctober 2019Article No.: 121Pages 1–7https://doi.org/10.1145/3331453.3361658. https://doi.org/10.1145/3331453.3361658
13. Qin, Y.; Tang, Q.; Xin, J.; Yang, C.; Zhang, Z.; Yang, X. A Rapid Identification Technique of Moving Loads Based on MobileNetV2 and Transfer Learning. Buildings 2023, 13, 572. https://doi.org/10.3390/buildings13020572.
14. Li T, Huang H, Peng Y, Zhou H, Hu H, Liu M. Quality Grading Algorithm of Oudemansiella raphanipes Based on Transfer Learning and MobileNetV2. Horticulturae. 2022; 8(12):1119. https://doi.org/10.3390/horticulturae8121119.

15. R. Patel and A. Chaware, "Transfer Learning with Fine-Tuned MobileNetV2 for Diabetic Retinopathy," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154014. https://doi.org/10.1109/INCET49848.2020.9154014
16. S. Dubey, Alzheimer's Dataset (4 class of Images), https: //www.kaggle.com/datasets/tourist55/alzheimers-dataset--4-class-of-images, 2020.
17. M. Hany, Chest CT-Scan images Dataset), https://www.kaggle.com/datasets/mohamedhanyyy/ches-ctscan-images9. P. Raikote, Covid-19 Image Dataset, https://www.kaggle.com / datasets / pranavraikokte / covid19image-dataset?datasetId=627146&sortBy=voteCount, 2020.

AUTHORS PROFILE

Ryan Marcus Jeremy M. Lupague completed his BS Math degree at Pamantasan ng Lungsod ng Maynila (PLM) with top honors (Magna Cum Laude), distinguishing himself as the top student in his class. His mind is a playground for programming prowess, statistical sorcery, and the mystique of machine and deep learning. With sleeves rolled up, he conjures cutting-edge models that breathe life into data, yielding insights and wisdom for decision-making. As if that weren't enough, Ryan donned the mantle of President (in the academic realm of 2022-2023) and earlier, the cape of Vice President (in the chronicles of 2021-2022) for the PLM Mathematical Society, showcasing his knack for leadership. He was also supported in his academic journey by a DOST-SEI scholarship, which acted as a guiding light. Venture into his world, and you'll see him deeply involved in Data Science, Machine Learning, Deep Learning, Natural Language Processing (NLP), and the creative side of Computer Vision.

Romie C. Mabborang is a dedicated educator with a profound commitment to mathematics. Currently serving as the Chair of the Mathematics Department at the Pamantasan ng Lungsod ng Maynila (University of the City of Manila), His journey has been marked by a passion for learning. Starting with a degree in Civil Engineering, his academic pursuits led him to achieve a Master of Science in Mathematics and has made significant strides toward a Doctor of Philosophy in Mathematics, having completed his coursework. His thirst for knowledge knows no bounds. With a heart firmly rooted in the classroom, he has spent considerable years in academia, molding the minds of math majors, engineering students, graduate students, and others from different fields. Beyond his role as an educator, his research pursuits span a wide spectrum, including Cryptography, Mathematical Modeling, Mathematics Education, Pattern Recognition, Machine Learning, Ensemble Learning, Algebra, Analysis, Combinatorics, Statistics, Time Series Analysis, and Multivariate Analysis.

Prof. Alvin G. Bansil is a Mathematics professor at Pamantasan ng Lungsod ng Maynila (University of the City of Manila). With a foundation rooted in a BS Chemistry degree and a reputable status as a registered Chemist, his intellectual journey took a twist as he earned his Master of Science in Mathematics Education. He stands at the threshold of his Ph.D. odyssey, vigorously penning his dissertation in pursuit of knowledge in Mathematics Education. His name reflects a strong dedication that lights up his journey and the determination that powers his efforts. Yet, it's his enchanting sense of humor that casts a spell, turning every encounter into a vivacious escapade. With him, dull moments are but a distant memory. Delving into his intellectual terrain unveils a fascination for Mathematics Education, where he breathes life into learning. His exploratory spirit extends to the realms of Statistics, unearthing insights from data's depths. Simple shapes open up to him, showing beautiful designs that enchant his love for math. In numerical analysis, he deciphers the language of numbers, unraveling the tales they tell. Amidst it all, Prof. Bansil crafts learning models that bridge the gap between theory and application.

Melinda M. Lupague, Within the educational domain of PLM in Manila, Philippines, resides an accomplished associate professorial lecturer. Notably, she also shoulders the responsibility of Vice-President within the MSP-NCR Chapter, showcasing her commitment to academia. Her intellectual curiosity takes flight across an array of captivating fields, encompassing the realms of data analytics, intricate modeling, foundational linear algebra, abstract algebra's complexities, the interwoven patterns of graph theory, and the frontiers of reinforcement machine learning. This constellation of interests attests to her profound engagement with diverse facets of knowledge and innovation.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.