Comparative Analysis and Optimization of Model-Method Combinations for Out-of-Distribution Detection in Medical Image Classification

GYUM CHO, University of Twente, The Netherlands

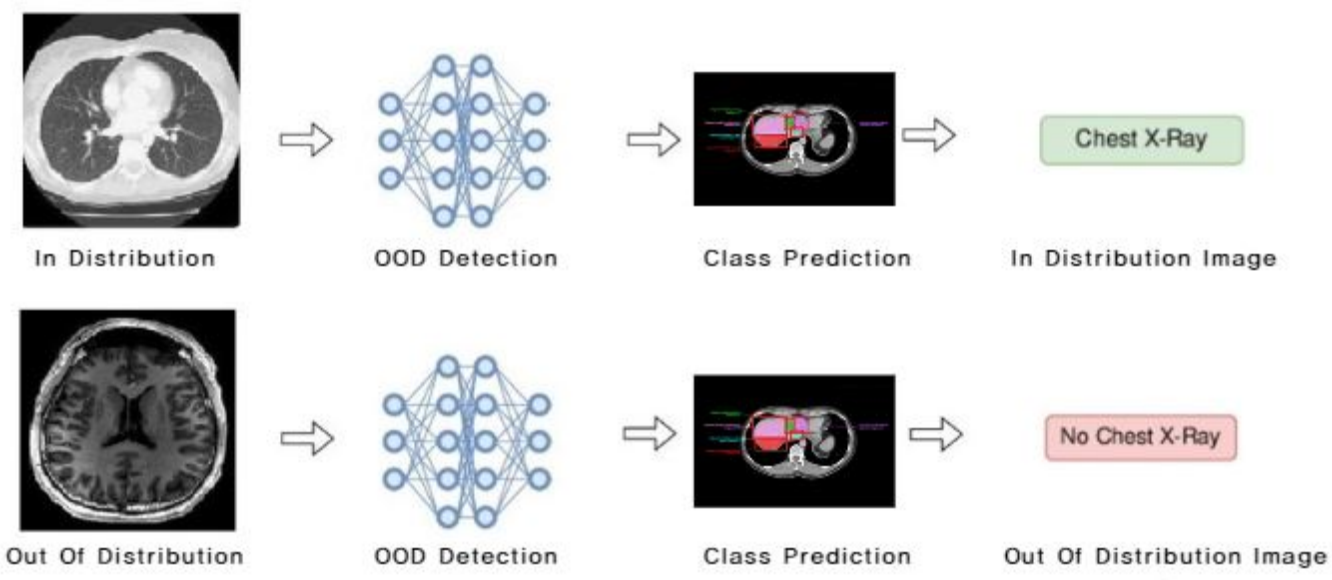

Fig. 1. General process of Out-Of-Distribution data detection.

The AI model is the logic circuit of the AI's behavior. Active research on AI models brought a remarkable enhancement to the performance of AI. AI is now replacing humans in various fields that were previously exclusive to humanity. Medical testing, scanning evaluation, and categorizing symptoms were sensitive tasks where human errors could lead to fatal situations. AI's high performance and low risk of missing critical factors have rapidly replaced this area. However, even with a high success rate, medical AI shows serious abnormal behavior when it faces new input that is far different from the training domain. This problem has been classified as an out-of-distribution (OOD) problem. It is one of the major problems related to the reliability of AI models in the medical field.

This paper analyzes the main factors of abnormal behavior when AI faces OOD problems. The external factors influencing this abnormal behavior will be analyzed and regulated with statistical observation through the training process. By incorporating evidence from experiments, this research demonstrates an approach to the OOD problem using three distinct pre-trained AI models. Each AI model will be trained with a medical MNIST dataset, and the statistical evaluation will prove which factors mainly affect the abnormal behavior of the AI. The result can prevent an OOD problem when AI classification models are used. Both patients and the collecting organization have consented to all medical images used for AI training.

Additional Key Words and Phrases: Deep learning, OOD, Neural network

1 INTRODUCTION

The charming AI assistant Jarvis from the movie Iron Man is an artificial intelligence secretary who supports Tony Stark. Jarvis can interact with humans with zero latency and provide creative suggestions to resolve complex tasks with Tony. With the evolution of technology, what seemed like pure science fiction is closer to reality. The radical evolution in technology gives AI the ability to perform complex tasks that were used to be exclusive to humans. Deep learning technology is a significant branch of machine learning, creating AI models that become a logic circuit for AI based on artificial neural networks. The neural networks process the input data with multiple layers. Each layer learns features of the input data and finds patterns to predict future output, similar to how the human brain processes information. For instance, convolutional neural networks (CNNs) are specialized deep learning models designed to process image data. It is widely used in image analysis tasks like image classification and motion evaluation.

The specialized characteristics of deep learning technology have proven helpful in medical fields where precision and reliability are essential in decision making. Deep learning models can analyze and learn from vast amounts of medical images and videos, creating robust AI models that help prevent medical errors and accurate

TScIT 41, July 5, 2024, Enschede, The Netherlands

classification aids in diagnosis, treatment planning, and patient management. However, ensuring the reliability of these models when they encounter unseen data or OOD data remains a significant challenge.

The OOD data is a terminology that explains the significant difference between the training data distribution. For example, a model trained on chest CT images might encounter breast MRI images as testing OOD data. If OOD data is not managed correctly, the model can make incorrect predictions according to training distribution, leading to negative consequences in the medical field. Recent research has focused on various OOD detection methods that support AI in distinguishing OOD data and categorizing them with different labels.

This thesis employs pre-trained deep learning models (ResNet50, MobileNetV2, and InceptionV3) for medical image classification tasks. Each AI model will be evaluated for their performance in conjunction with multiple OOD detection methods to find the most effective combination with higher accuracy.

The primary objective of this research is to evaluate various OOD detection methods to improve model interoperability and reliability. OOD methods such as Mahalanobis distance, ODIN, and Max Softmax will be compared. Furthermore, the XAI techniques of t-SNE will provide insight into model predictions. The XAI techniques will make the decision making process transparent and interpretable. Effective detection and managing OOD data can prevent incorrect diagnoses and ensure that AI models provide accurate and trustworthy results. The findings from this research will provide valuable insights for developing AI systems capable of supporting healthcare professionals in delivering high-quality patient care.

1.1 Research Questions

1.1.1 Question 1. Which combination of pre-trained deep learning model and OOD detection method provides the highest accuracy and reliability in medical image classification?

1.1.2 Question2. What optimization techniques can further improve this combination?

2 BACKGROUNDS AND RELATED WORKS

2.1 Backgrounds

Image Classification. Early image classification techniques relied on feature extraction and machine learning algorithms. However, with the improvement of deep learning, Convolutional Neural Networks (CNNs) have become the dominant method for image classification. CNN brought new insight to the AI field by hierarchical feature representations from raw pixel data, which led to unprecedented improvements in classification accuracy.

During the research about the history of image classification technique, Krizhevsky et al.[10] introduced AlexNet and CNNs in image classification by winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Rectifying linear units (ReLUs) for activation and dropout for regularization shows new standards for deep learning architectures. Simontan and Zisserman [16] proposed VGGNet, which explored the impact of network depth on accuracy and found that deeper networks with smaller convolutional filters could achieve better performance. Szegedy et al. [17] introduced GoogLeNet (Inception v1) and used a novel inception module to capture spatial hierarchies in images efficiently. Each research becomes a motivation to adapt ReLu for fine-tuning and reason to choose a pre-trained model for the thesis.

pre-trained model. The use of pre-trained models is a common practice in deep learning research. Pre-trained models offer several advantages, including faster training times and improved performance compared to scratch. Due to their prior learning of diverse features, pre-trained models possess strong generalization capabilities.

He et al. [7] proposed ResNet (Residual Network), introducing residual connections that allow deep networks to train effectively. ResNet has demonstrated outstanding performance in various image classification tasks. Howard et al. [8] introduced MobileNet, a lightweight deep-learning model for real-time mobile and embedded mobile applications. MobileNet combines high accuracy with efficient computation, making it suitable for practical deployments. Szegedy et al. [18] presented the Inception model, which employs parallel convolutional filters of different sizes to capture features at multiple scales.

2.2 Related Works

The OOD problem is a significant challenge in deploying deep learning modes in real-world applications. Diverse methodologies have been evaluated to reduce the critical error of the AI model, and there are three major approaches to avoid the OOD problem. Understanding each research study will provide comprehensive knowledge of how these methods detect and mitigate OOD data.

Liang et al. [11]introduced ODIN (Out-of-Distribution detector for Neural networks) aims to improve the reliability of OOD detection in neural networks by employing temperature scaling and input pre-processing. The temperature scaling is used to make the network's predictions more separable by adjusting the softmax output of the neural network. Input pre-processing involves adding small perturbations to the input data to amplify the differences between in-distribution and OOD samples. The main experiment evaluated ODIN with the CIFAR-10 and CIFAR-100 datasets with DenseNet and WideResNet architectures. The researcher trained the models on in-distribution and tested them with both in-distribution and OOD samples. The measurement of false positive rate and true positive rate has been used to evaluate performance. As a result, ODIN significantly reduced the false positive rate compared to baseline methods. ODIN shows the highest improvement on CIFAR-10 with DenseNet, the FPR at a 95% TPR decreased from 34.7% to 4.3%. The research concluded that combining temperature scaling and input pre-processing enhances the OOD detecting performance.

Conformal prediction is evaluated in Shafer and Vovk [15] provides a framework for calculating predictions about data confidence. This methodology constructs prediction intervals or sets guaranteed to contain the output with a probability. Conformal predictors can indicate when a sample is likely OOD by producing huge prediction sets or low confidence scores. In research, conformal prediction consistently produced valid prediction sets with the desired coverage probability. In cases of OOD samples, the prediction sets were noticeably larger or indicated low confidence, effectively identifying

OOD data. Conformal prediction offers a robust statistical method for quantifying uncertainty in predictions.

D'Angelo and Henning [4] investigate how Bayesian Neural Networks (BNNs) incorporate Bayesian inference to estimate uncertainty in model predictions. BNNs use a probabilistic framework where weights are treated as distributions rather than fixed values. BNNs were trained on in-distribution data. The networks' uncertainty estimates were analyzed during testing to identify OOD samples. High uncertainty was indicative of OOD data. Gaussian processes and variational inference were used to approximate the posterior distribution of the network weights. BNNs demonstrated a high ability to flag OOD samples through increased predictive uncertainty. For example, when applied to MNIST, BNNs successfully identified OOD samples from the non-MNIST dataset by their high uncertainty scores.

3 METHODOLOGY

This thesis focuses on evaluating the effectiveness of the OOD detection method in combination with pre-trained deep learning models for medical image classification. Training, noise insertion, OOD detection design, model evaluation, optimization, and XAI techniques are chosen as major methods to test the performance of handling OOD data. The exquisite methodology award comprehensive insight.

3.1 Data preparation

Preparing the right type and format of the dataset was an essential step in testing the performance of different OOD methods. The Medical MNIST dataset [12] consists of 6 different categories of medical images used for the classification experiment task. AbdomenCT, BreastMRI, ChestCT, CXR, Hand, and HeadCT were the six categories of medical images. In this thesis, AbdomenCT, BreastMRI, and ChestCT are assigned as in-distribution data, and CXR, Hand, and HeadCT are assigned as OOD data. Each class consisted of 10,000 images related to a specific category with a size of 64 X 64 pixels. The dataset was collected by Hacettepe University at Medical School and evaluated for validity by Kaggle. Eighty percent of the dataset was used for training purposes, and twenty percent was used for validation testing. After separating training and validation datasets, true labels for the testing dataset were prepared to evaluate the model's performance by comparing the predicted and actual labels.

The dataset goes through pre-processing to change the shape of the data for fitting with different types of pre-trained models. To achieve the requirements of ResNet50, MobileNetV2, and InceptionV3, the pre-processing transformations are applied to resize the images to 299 X 299 pixels.

3.2 Model selection and training

Three pre-trained deep learning model have been chosen due to their proven effectiveness in image classification tasks. The pre-trained model has been fine-tuned to output predictions for six classes of medical images. During tuning and training, the base layers of the models are frozen to retain the pre-trained features. The final classification layers are fine-tuned to fit the six categories of the Medical MNIST dataset. The training process uses stochastic gradient descent (SGD) with a learning rate of 0.01 and a batch size of 16 to ensure efficient training while managing the limited GPU resources. The training was performed in the in-distribution categories (AbdomenCT, BreastMRI, ChestCT).

Resnet50 Model [3] is a deep residual network that addresses the problem of vanishing gradients in the deep network by skipping residual connection. The connection allows the network to learn effectively by skipping layers. Skipping layers helps in maintaining the gradient flow during backpropagation.

MobileNetv2 model [2] is designed to be lightweight and efficient to make the model suitable for deployment in a constrained resource environment. The model structure uses depthwise separable convolutions to reduce the number of parameters and computational cost. In the medical environment, computational resources can be limited for special purpose devices such as hearing signal detectors, embedded implants, and medical devices.

InceptionV3 model [1] employs a complex architecture with multiple parallel convolutional layers of different sizes that capture various levels of detail in the images. The model's architecture effectively captures intricate features in the images.

3.3 Data Augmentation

The noise is added to the training data to simulate challenging conditions for the OOD detection method and avoid overfitting problems. Gaussian noise with a specific noise level has been used to add noise to the dataset. The noisy data helps assess how well the models can perform when the input data is not perfectly clean.

3.4 OOD Detection

Three distinguishable OOD detection methods were chosen for this experiment. All of OOD detection methods have single objective to determine the OOD data inside of dataset.

The **Mahalanobis** method calculates the distance of the input data from the training data distribution in the feature space. A large Mahalanobis distance indicates a high possibility that the chosen data is likely OOD data. The Mahalanobis approach considers the correlations between data features that identify the image and measures how far an input is from the distribution center.

The **ODIN** method perturbs the input data with distance parameters and uses the resulting softmax scores to differentiate between in-distribution and OOD samples. ODIN uses a temperature scaling technique that adjusts the confidence of the probability output by the softmax layer. This process involves dividing the logits by a temperature parameter before applying the softmax function.

The **MaxSoftmax** method relies on the confidence of the softmax output. Low confidence scores suggest a high possibility of OOD data. Not exhibiting high confidence in any particular class can be evaluated and considered as OOD data. MaxSoftMax method is the most straightforward and effective approach to focus on the model's uncertainty in its predictions.

3.5 Evaluation

The evaluation phase analyzes the model's performance using statistical metrics, including accuracy rate, recall, F1 score, confusion

matrix, and the area under the receiver operating characteristic curve (AUROC). Accuracy measures the overall correctness of the model's prediction against the actual label of the data.

Accuracy measures the overall correctness of the model's predictions against the actual true label of the data. It is the ratio of correct predictions to the total number of predictions made. The accuracy ratio is mathematically represented as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

- **TP (True Positives)**: The number of correctly predicted positive samples.
- **TN (True Negatives)**: The number of correctly predicted negative samples.
- **FP (False Positives)**: The number of incorrectly predicted positive samples.
- **FN (False Negatives)**: The number of incorrectly predicted negative samples.

Accuracy provides a general sense of how often the model is correct but can be misleading if the dataset is imbalanced.

Recall, also known as sensitivity or true positive rate, measures the ability of the model to identify all relevant instances correctly. Mathematically represented as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

The recall value is crucial when missing a positive instance is more costly than incorrectly classifying a negative one.

The F1 score is the harmonic mean of precision and recall. It is beneficial when the dataset is imbalanced. The F1 score is calculated mathematically

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

The F1 score combines precision and recall into a metric by taking harmonic mean. It offers a better measure of the model's performance where the class distribution is uneven.

The confusion matrix provides a detailed model performance breakdown by showing the actual versus predicted classification. It is a square matrix of size N X N, where N stands for a number of classes that include TP, TN, FP, and FN. The confusion matrix clearly identifies specific areas where the model may be under-performing.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) The ROC curve is a graphical representation of the model's diagnostic ability. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true positive rate is equivalent to recall, and the false positive rate is defined as

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5}$$

The Area Under the Curve (AUC) is a single scalar value that summarizes the model's performance across all threshold values. An AUC of 1 represents a perfect model, and a value near 0.5 indicates a model that performs no better than random chance.

3.6 OOD Detection Method Optimization

The optimization techniques are deployed to enhance the OOD detection capability in medical image classification tasks. The Mahalanobis distance based method is optimized through feature space reduction using principal component analysis (PCA). PCA is a statistical technique for dimensional reduction. The main objective of PCA is to simplify high-dimensional data without losing variance. The procedure starts with standardizing the data, ensuring each feature has a mean of zero and a standard deviation of one. Standardization transforms each feature by subtracting the mean and dividing it by the standard deviation. After the standardization, a covariance matrix calculation of the standardized data is performed to understand how the features vary concerning other features. This process reduces computational complexity and enhances the discriminative power of Mahalanobis by focusing on the most informative features.

3.7 Explainable AI (XAI)

Gaining insight into the model's decision making process is essential to evaluate why AI makes certain decisions. The t-distributed stochastic neighbor embedding algorithm works in order of feature extraction, transformation, and plotting. The t-SNE technique reduces the high-dimensional feature space into two-dimensional data. It allows for visual inspection of the feature clusters to visualize which factor was a major weight that affected the decision of the AI model. The transformed data is plotted in two-dimensional data points. t-SNE starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities representing similarities. Given a pair of points x_i and x_j, the similarity is given by:

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{6}$$

The low-dimensional counterparts y_i and y_j aim to maintain these similarities:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}} \tag{7}$$

The algorithm works by minimizing the divergence between two probability distributions. One represents pairwise similarities of the input objects in the high dimensional space and the other represents similarities of the corresponding low dimensional points.

4 EXPERIMENT PROCESS AND RESULTS

4.1 Kaggle environment setup and data preparation

Training deep learning models requires powerful GPU resources, which Kaggle [9] has provided in a virtual running environment. The primary tasks start from separating datasets into in-distribution and OOD datasets. The dataset was successfully settled, and the model was trained with an in-distribution dataset and tested with a

mixture of OOD datasets. This setup laid the foundation for training the models, adding noise, performing OOD detection, and evaluating the results. Each deep learning model requires a different format and shape. The image size was transformed and resized to 299 X 299 pixels to accommodate the requirements of the InceptionV3 model.

4.2 Model Setup

Three different pre-trained model were loaded and went through the fine-tuned process. The models were modified by updating their final layers to fit the six class medical MNIST dataset. Specifically, the final connected layers of each model that included ReLU activation and Softmax outputs were modified. The models' base layers stay frozen to retain the pre-trained weights.

4.3 First Training and Evaluation

The first evaluation shows how each pre-trained model treats the OOD problem without any specified method. The models were trained on the in-distribution dataset for ten epochs with a batch size 16. The training process uses the SGD optimizer and CrossEntropy-Loss. The average loss and accuracy of the models are represented in Table 1.

Model	Validation Loss	Accuracy
ResNet50 (Run 1)	1.3366	0.6986
MobileNetV2 (Run 1)	1.3410	0.6928
MobileNetV2 (Run 2)	1.3459	0.6893
InceptionV3 (Run 1)	1.3038	0.7390
InceptionV3 (Run 2)	1.3023	0.7454

Table 1. Validation Loss and Accuracy for Different Models After Training

As a result of the pure pre-trained model, the Inception model performed reasonably better than other models in natural status. However, this result contrasts after OOD detection.

4.4 OOD Detection and Visualization

Three distinguished OOD detection methods were utilized in this phase. Mahalanobis, ODIN, and MaxSoftMax methods were performed in order. The OOD scores generated by these methods were evaluated using precision, recall, and F1-score

4.4.1 ResNet50 Model. In Figure 2, The Mahalanobis method shows the most accurate performance with high TP and TN results. The true positive rate shows that the Mahalanobis method with the Resnet50 model is a powerful combination for detecting OOD data. The high precision value also shows the reliability of the method. The high performance of the Mahalanobis method with ResNet50 is based on the feature extraction capability of the ResNet50 model. A deep convolutional neural network has the unique ability to learn complex and hierarchical feature representations. The skip connections help train deeper networks by mitigating the vanishing gradient problem. Hence, it was possible to capture intricate patterns and dependencies in the data.

MaxSoftMax method shows relatively poor performance out of the three different models in Figure 3. The method assumes that the model should produce high confidence scores for in-distribution

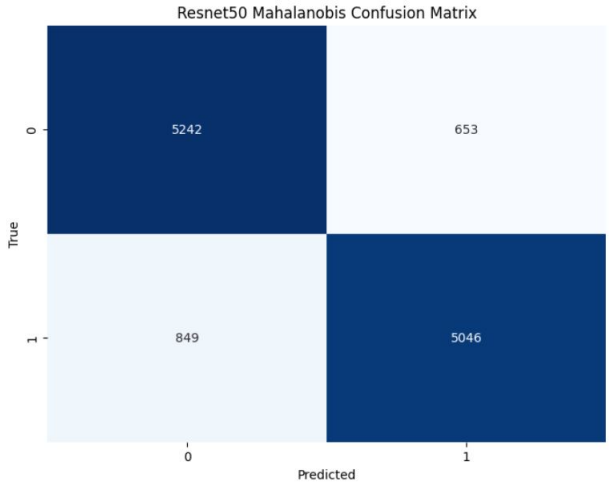

Fig. 2. Confusion matrix of Resnet50 - Mahalanobis.

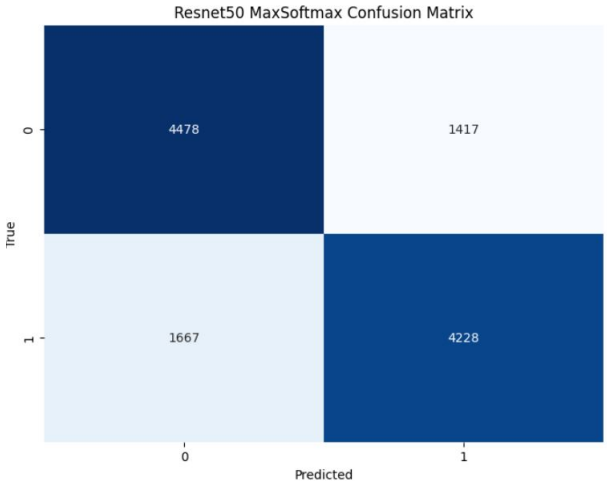

Fig. 3. Confusion matrix of Resnet50 - MaxSoftMax.

data and lower confidence scores for OOD data. The detection process is sensitive about setting a threshold for these confidence scores. However, even with the AUROC score, it was not easy to get the best threshold value. MaxSoftMax evaluates a given input and produces a probability distribution over the possible classes using the softmax function in the final layer. The maximum value of this distribution represents the model's confidence. However, in the case of the medical MNIST image, where the difference between in-distribution and OOD is slight, it is hard to determine the maximum value. For example, breast MRI and AbdonmenCT have a similar percentage of white areas in the human body. The major difference between the two classes is the image shape, but the MaxSoftMax method makes it difficult to get a different image shape. Therefore, determining the maximum value among the datasets took much work. For this reason, the threshold value could not fit for the best performance.

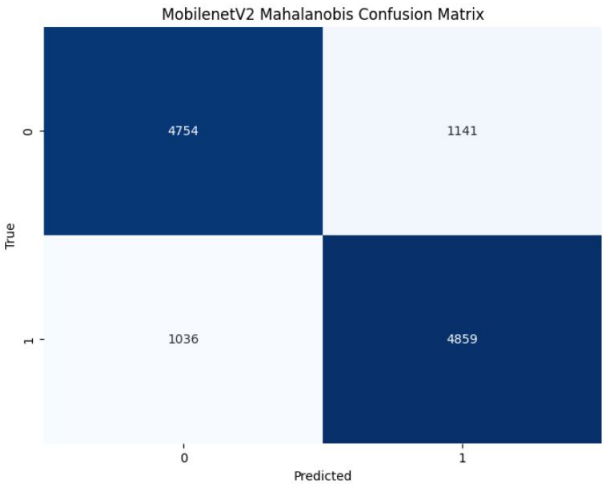

Fig. 4. Confusion matrix of Mobilenet - Mahalanobis.

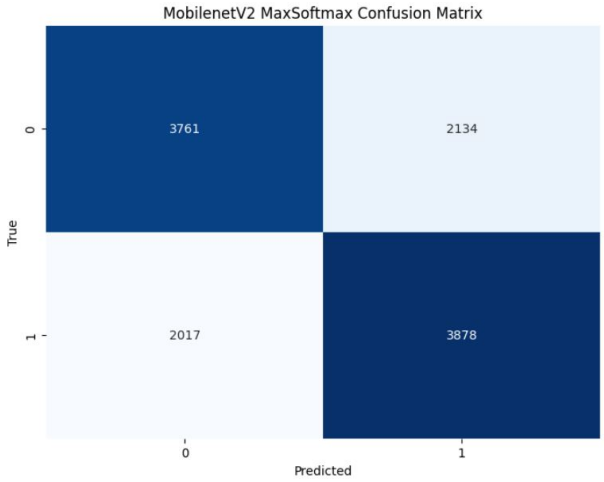

Fig. 5. Confusion matrix of Mobilenet - MaxSoftMax.

4.4.2 MobileNetV2. model is suitable in situations with a limitation in calculating resources. The general performance has decreased compared to the ResNet50 model. This model architecture is efficient in terms of both computation and memory. Kaggle displays GPU usage, and the MobileNetV2 model shows the lowest GPU RAM (vRAM) usage compared to other models. It is important to note that MobileNetV2 is available in most environments. However, its lightweight nature impacts its ability to handle OOD detection compared to a complex model. Like ResNet50, Figure 4 shows Mahalanobis performing most accurately and highly reliably. However, the performance appeared slightly lower than ResNet50 due to limitations in the model's simpler architecture.

Figure 5 shows the poor performance of the MaxSoftMax method with the MobileNetV2 model. A more straightforward architecture results in higher confidence scores for OOD data. Higher confidence scores for OOD data lead to more false positives. After getting a

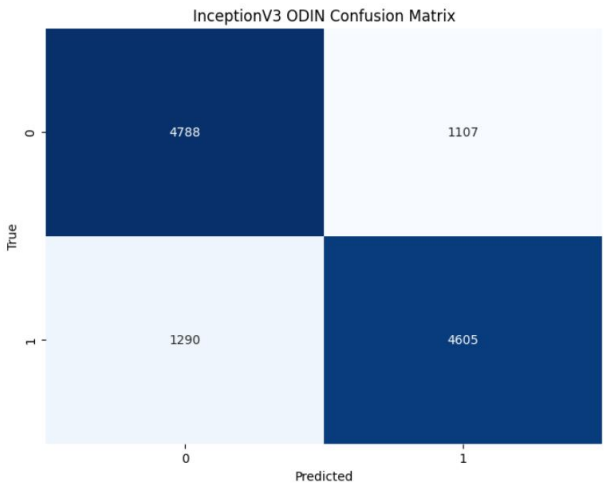

Fig. 6. Confusion matrix of Inception - ODIN.

poor result from the MaxSoftMax method, it was necessary to boost the performance and compare it with other valid methods. For this purpose, temperature scaling is applied. It adjusts the temperature of the softmax function to produce more calibrated probabilities. The implementation uses scaling during inference to smooth the confidence scores. The performance boost is achieved by dividing the logits by a temperature parameter T > 1 before applying the softmax function. The temperature scaling has the benefit of reducing the overconfidence of the model's predictions and making it easier to distinguish OOD samples.

4.4.3 InceptionV3. is a convolutional neural network architecture known for its efficiency and high performance in image classification tasks. Factorized convolutions and aggressive regularization reduce computational costs while maintaining high accuracy. Given the architectural strength of InceptionV3, a high accuracy rate was expected in all three OOD methods. However, the result did not show up as expected.

Figure 6 shows the values of false positive and false negative are relatively high with the InceptionV3 model. Unlike the other two methods, ODIN shows unexpected performance degradation. Multiple attempts have occurred, but the result remains the same.

Figure 7 shows that the t-SNE graph visually represents high-dimensional data in two-dimensional space. It was possible to observe the clustering and separation of different datasets. The graph shows that the AI struggles to distinguish chestCT and abdomenCT datasets. These two datasets have features similar to those of the others. The image is 299 X 299 in size, and both CT scans have circular shapes centered with a black background. The characteristic of image shape confuses AI in distinguishing the correct label. Compared to ResNet50's architecture, the residual connection helps maintain the integrity of feature representations across layers. This stable feature representation can benefit the ODIN method since the method heavily depends on subtle changes in softmax scores. In contrast, InceptionV3's varied and parallel convolutional paths produce more variable and less predictable feature maps, making it

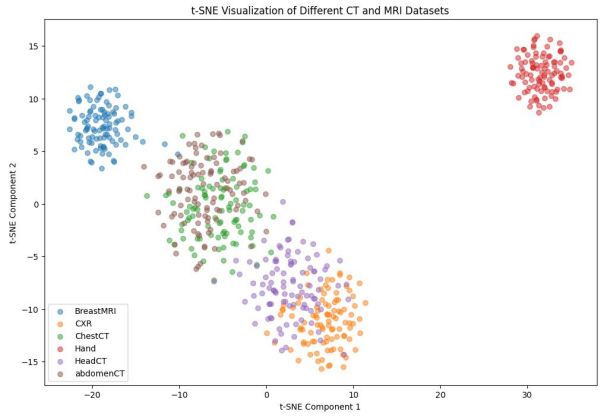

Fig. 7. t-SNE graph of Inception - ODIN.

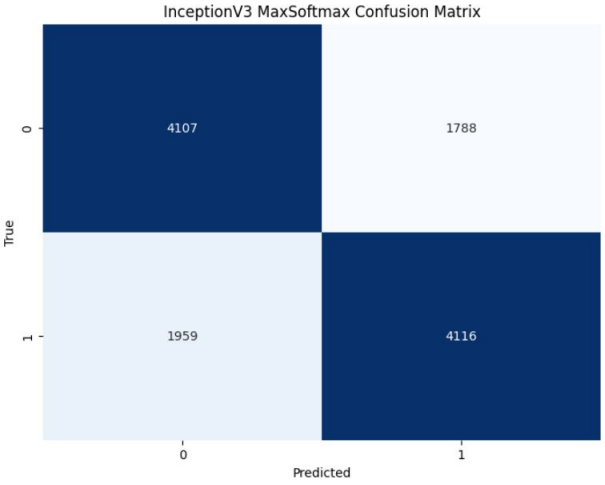

Fig. 8. Confusion matrix of Inception - MaxSoftMax.

harder for ODIN to distinguish between in-distribution and OOD samples. The complex model design does not always bring high performance in image classification.

The other two methods show similar results to those of the ResNet50 model. Minor differences were found between fewer false negatives and more false positives depending on which part of the picture the AI focused on.

4.5 Optimizing OOD method

4.5.1 PCA optimization. The evaluation of each model and method provides the most suitable combination for the medical MNIST dataset. The combination of Resnet50 and Mahalanobis detection showed the highest accuracy and reliability. The medical MNIST dataset is high-dimensional data that may contain redundant information. The redundant information can affect the accuracy of the OOD topology. It increases computational inefficiency and the risk of overfitting. Trivedi et al. [19] explain overfitting as an undesirable

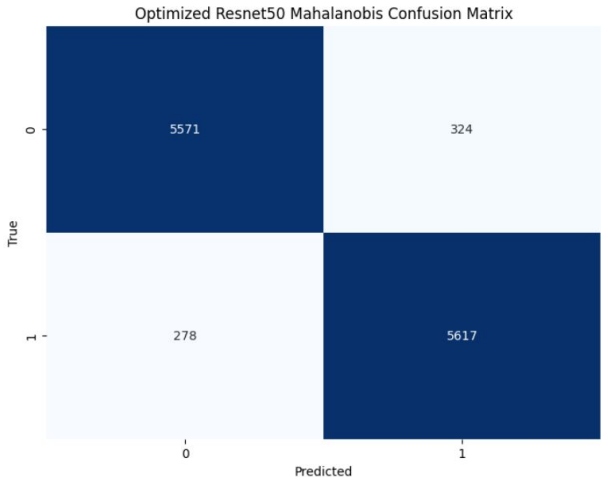

Fig. 9. Confusion matrix of optimized Mahalanobis.

AI behavior when it produces accurate prediction results for the training dataset but relatively poor results for new data.

PCA optimizes the dataset's feature space and mitigates this issue. The reduced dimensionality of the feature space helps better capture the distribution of in-distribution and OOD data.

Figure 8 shows the better performance of the optimized method. However, the result was hardly unstable and heavily dependent on the quality of the model's training and exhibited instability. Insight from [20] provides the potential problem of PCA. Loss of important variance information during the dimensionality reduction can lead to this ambiguous result. PCA prioritizes the directions of maximum variance that might not align with those most helpful in distinguishing between in-distribution and OOD data. The misguided value can lead to reduced effectiveness of the OOD detecting task. The important features may be underestimated in the reduced feature space. PCA can be more effective in a controlled environment similar to a realistic medical dataset. For example, PCA performs sufficiently when the data has clear linear separations or when computational efficiency is prioritized over absolute detection performance.

4.5.2 ODIN optimization. Figures 9 and 10 illustrate the performance of the ODIN method with the InceptionV3 model before and after optimization. Before the optimization, the ODIN method demonstrated unexpected and downgraded performance. The in-distribution and OOD data scores were closely clustered and significantly overlapping. Since the reason for this overlap was not inevitable, multiple optimization methods have been applied. Firstly, adversarial training is used to enhance the model's feature space. Even though data augmentation was already applied to the training dataset, extra augmentation has been applied for better results. The training process helped the model learn generalized features. Secondly, the threshold is essential to balance between the true positive rate and the false positive rate. A grid search over various threshold values has been performed using a calibration set comprising both in-distribution and OOD data. AUROC value was used to evaluate these thresholds, and the closest to 1 was selected as the best value.

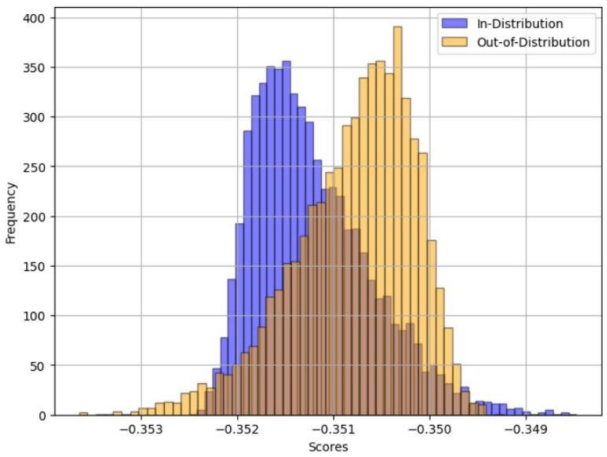

Fig. 10. OOD score histogram - before optimization.

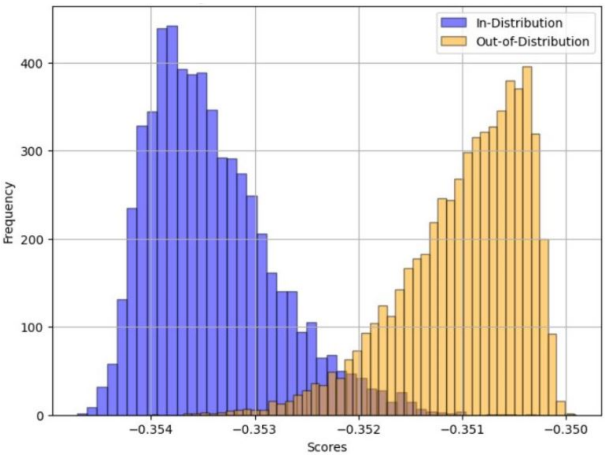

Fig. 11. OOD score histogram - after optimization.

As a result, the histogram shows distinguishable results between in-distribution and OOD data. There are still some overlapping cases between the two datasets, but the accuracy increases after optimization. One issue was that the overlapped area percentage was unstable and moved between 15% to 35%, depending on the test dataset.

5 CONCLUSION

The results of the experiments demonstrate that the ResNet50 model offers the most accurate and powerful performance among the three models. ResNet50 consistently outperforms in various matrices compared to other models and establishes itself as a robust choice for medical image classification tasks. The Mahalanobis method is the most accurate and reliable OOD detection approach. It shows distinct class boundary classification and feature distribution through its effectiveness. However, it is important to note that the superiority of the Mahalanobis method may vary with different classification datasets. Other methods can perform better under different conditions of input datasets.

In contrast, the InceptionV3 model's complex architecture was not necessarily connected to better performance. This observation points out that high complexity in model architecture does not always equate to higher accuracy.

MobileNetV2 shows the lowest accuracy and performance among pre-trained models. However, showing the lowest calculation resource consumption level. It is particularly suitable for medical environments where computational resources may be limited.

Finally, optimizing each OOD method showed different methods for different models. XAI techniques and statistical analysis are essential to find them. Methods that rely on thresholds are sensitive to statistical measures like AUROC and need to be changed. The improved performance after finding the correct threshold shows the importance of optimizing the optimization. Overall, the experiment found that selecting a suitable model and method based on the specific characteristics of the dataset is important. For example, datasets with clear class boundaries and distinct features may benefit from the Mahalanobis method and models like ResNet50. Conversely, environments with limited computational resources might favor models like MobileNetV2 despite their lower accuracy compared to other models. Therefore, tailoring the choice of model and OOD detection method to the dataset and application context is crucial for achieving optimal performance in medical image classification tasks.

6 APPENDIX

6.1 Usage of AI technique

During the preparation of this work, the author Gyum Cho used Grammarly[6] to perform a general grammar check of the thesis and to ensure an effective writing tone.

During the preparation of this work, the author Gyum Cho used DeepL[5] as an auto-translation tool to translate parts of the thesis written in the author's mother tongue into English and to translate related research articles.

During the preparation of this work, the author Gyum Cho used ChatGPT[13] in order to generate an Overleaf format for tables, figures, mathematical equations, and references (citations). Part of the background research, mathematics evaluation, and XAI technique paragraph rephrased to correctly explain what contents were introduced by each research. For programming, AI assists with the environment setup for both Google Colab and Kaggle virtual machines, including configuring pre-trained models and printing output graphs.

During the preparation of this work, the author Gyum Cho used QuillBot[14] in order to rewrite ambiguous sentences for better clarity and intuition.

After using this tool/service, the author Gyum Cho reviewed and edited the content as needed and take full responsibility for the content of the work.

REFERENCES

[1] François Chollet et al. 2023. InceptionV3 - Keras Documentation. https://keras.io/api/applications/inceptionv3/ Accessed: 2024-06-30.

[2] François Chollet et al. 2023. MobileNet - Keras Documentation. https://keras.io/api/applications/mobilenet/ Accessed: 2024-06-30.
[3] François Chollet et al. 2023. ResNet - Keras Documentation. https://keras.io/api/applications/resnet/ Accessed: 2024-06-30.
[4] Francesco D'Angelo and Christian Henning. 2021. Evaluation of deep learning models for real-time object detection on embedded systems. *Sensors* 21, 14 (2021), 4772.
[5] DeepL GmbH. 2024. DeepL Translator. https://www.deepl.com/translator
[6] Inc. Grammarly. 2024. Grammarly. https://www.grammarly.com/
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.
[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).
[9] Kaggle. 2024. Kaggle: Your Home for Data Science. https://www.kaggle.com/. Accessed: 2024-06-30.
[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.
[11] Shiyu Liang, Yixuan Li, and R Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.
[12] Andrew Mvd. 2020. Medical MNIST. https://www.kaggle.com/datasets/andrewmvd/medical-mnist. Accessed: 2024-06-30.
[13] OpenAI. 2024. ChatGPT. https://chat.openai.com/
[14] QuillBot. 2024. QuillBot. https://quillbot.com/
[15] Glenn Shafer and Vladimir Vovk. 2008. *A tutorial on conformal prediction*. Journal of Machine Learning Research.
[16] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.
[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 2818–2826.
[19] Udai Bhan Trivedi, Milind Bhatt, and Prerna Srivastava. 2021. Prevent Overfitting Problem in Machine Learning: A Case Focus on Linear Regression and Logistics Regression. In *Innovations in Information and Communication Technologies (IICT-2020)*, Pradeep Kumar Singh, Zdzislaw Polkowski, Sudeep Tanwar, Sunil Kumar Pandey, Gheorghe Matei, and Daniela Pirvu (Eds.). Springer, Cham. https://doi.org/10.1007/978-3-030-66218-9_40
[20] McKell Woodland, Nihil Patel, Mais Al Taie, Joshua P. Yung, Tucker J. Netherton, Ankit B. Patel, and Kristy K. Brock. 2023. Dimensionality Reduction for Improving Out-of-Distribution Detection in Medical Image Segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, Carole H. Sudre, Christian F. Baumgartner, Adrian Dalca, Raghav Mehta, Chen Qin, and William M. Wells (Eds.). Springer Nature Switzerland, Cham, 147–156.