OOD Detection on Medical Images and Explainable OOD

BERK IMRE, University of Twente, The Netherlands

This research paper investigates various out-of-distribution (OOD) detectors in machine learning (ML) and deep learning (DL), specifically in computer vision in the field of medical images, for which detection of OOD is very crucial. Given that DL models are more popularly used than ML models with state-of-the-art methods, DL-based OOD detection will be the main focus of the research. Systematic experiments will be conducted to analyze the performance and reliability of the OOD detectors on medical images with the increasing range of severity in terms of distribution drift. This paper also introduces an explainability approach to out-of-distribution (OOD) images in DL models to understand the behavior of models, particularly regarding how OOD data affects downstream tasks such as classification and sources of failure of DL models in the presence of OOD data. The findings of this research highlighted the strong performance of PyTorch-OOD in OOD detection in healthcare DL applications. The findings showed the substantial effect of OOD on the confidence and accuracy of DL models in classification tasks. Additionally, noise and misleading visual similarities were identified as the main sources of failure for DL models in classification tasks, in the presence of OOD input.

Additional Key Words and Phrases: Out-of-Distribution (OOD) In-Distribution, Explainability, Machine Learning (ML), Deep Learning (DL), Neural Network

1 Introduction

In the last decade, the use of deep learning models has become more and more popular in various domains [5]. It is used to solve complex real-world problems with smart data-driven solutions. However, with the integration of deep learning models in critical applications such as autonomous vehicles, medical diagnostics, cybersecurity, and many more, guaranteeing the robustness and reliability of the models became extremely crucial. These standards are not always ensured, often depending on how the models are trained. Training is often done with the assumption that the test data will be similar to the training data, although, this does not always hold especially when the model is deployed in real-world applications. In such circumstances, performance can drop severely. This performance drop can lead to quite catastrophic failures in the domains such as medicine or autonomous vehicles.

The general term for when a deep learning model encounters data that is unlike anything it has seen during training is called out-of-distribution (OOD). A simple example to OOD would be a model that is trained with images of dogs and cats, receives an image of a fish as an input. This may lead to unreliable prediction by the model since it has never seen a fish before in the training process. To avoid unreliable predictions caused by OOD, extensive amounts of data are used to train big neural networks. However, it is a big challenge to deal with evolving data distribution in many ways such as cost and resources. For this reason, the detection of OOD data becomes a crucial field since models cannot possibly avoid all OOD data. This research paper will also work in this field and conduct experiments to get an understanding of which OOD detection techniques perform better and why [9]. The study will be specifically done on medical images in which the detection of OOD is extremely critical.

Additionally, achieving a deeper understanding of why a model fails when encountering OOD data is very crucial. The black-box nature of DL models makes it difficult to address the failures of the models [3]. Explainability methods have been developed to address the black-box nature of deep learning (DL) models by highlighting which regions of images the models used for classification. By using these methods to reveal the reasons and conditions that DL models fail against OOD data, this research will contribute to developing more robust and reliable models in the future. Thus, this research paper will also focus on using explainability tools in order to reveal why DL models fail against OOD data.

In summary, this research paper will examine the performance of different OOD detection techniques in medical data sets. In addition, it will use explainability tools to reveal why and how a model fails with an OOD input. The remainder of this document is organized as follows: Section 2 explains the state of the art. Section 3 introduces the problem statement and presents the research questions. Section 4 explains the tools and datasets used in the research. Section 5 provides the methodology of this investigation. Section 6 explains the implementations done to carry out the experiment. Section 7 provides the experiment and the results. Finally, Section 8 will conclude the paper with a discussion and conclusion.

2 State of the Art

The current situation in the field of OOD contains various approaches and techniques to tackle the problem of detecting OOD instances. Current research in this area suggested various ways to enhance OOD detection. For instance, a small-scale study showed that adapting camera parameters according to OOD detector leads to an increase in performance [10]. Techniques such as temperature scaling and input perturbation used for OOD detection for neural networks enhance the detection of OOD by adjusting the softmax scores [16] [10]. The current state of the art provides generalized frameworks that make OOD detection more reliable [16]. All the ongoing research in this field is very essantial to make the field of OOD detection more dependable to have more reliable real-world DL applications, especially in medical imaging.

In addition, there are several libraries published for the detection of OOD in real-world applications. The Pytorch-OOD library is one of the promising libraries that provides high-accuracy OOD detection techniques [9]. These techniques have shown successful results in standard datasets such as CIFAR and ImageNet [9]. However, it remains uncertain which techniques are most effective in the domain of medical imaging in which the nature of the data can be quite different than the very standard datasets. For this reason,

TScIT 41, July 5, 2024, Enschede, The Netherlands
© 2024 University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

the research will explore the effectiveness of various OOD detection techniques in the domain of medical imaging.

Moreover, the question of how and why neural networks fail against OOD is still not quite answered in the field of OOD. There has been lack of investigation especially with using most recent explainability tools that reveals the behavior of complicated neural networks. These tools address the black-box nature of DL models by revealing which parts of the inputs were most influential in the classification. Methods like GradCAM and SmoothGrad proved to be effective in providing accurate explainability [13] [6]. However, there is still a need to integrate these methods into the field of OOD to provide some new insights. For this reason, the research will examine the effect of OOD in downstream tasks such as classification and explainability will be applied to provide insights into how DL models behave with OOD data.

3 PROBLEM STATEMENT AND RESEARCH QUESTIONS

The data distribution in real-world domains changes quickly, and training deep learning models with all possible distributions is expensive. As a result, it is critical to identify OOD data and ensure that this process is as accurate as possible. However, so far it has been a big challenge to understand which OOD detectors or techniques work better than others since different experiments are giving different results. The results so far mainly deviate when there is a domain shift or change in the severity of OOD [5]. Especially, when recommended OOD detectors are used on medical datasets, there is a significant deviation in how accurately they work with medical data. Therefore, it is still quite significant to understand which OOD detection methods are more effective specifically on medical images, yet research in this field is insufficient [1]. This paper aims to provide some insight into this gap in the literature.‘

Even though the field of OOD detection is rapidly growing, there still has been a huge gap in the explainability of OOD. There are numerous techniques to deal with OOD yet not enough studies have been done to analyze how OOD affects the performance of DL models in downstream tasks such as classification. Additionally, there is a lack of research on the sources of DL model failures in the presence of OOD data in classification. A detailed understanding of this will be very significant insight for building robust and reliable DL models in the future. Thus this paper aims to examine the performance of downstream tasks such as classification with and without the presence of OOD data and apply XAI to detect the sources of correct and erroneous detections.

3.1 Research Questions

(1) How accurately do the OOD detectors in the PyTorch-OOD library work on cases of OOD data of increasing difficulty in medical images?
(2) How does the presence of OOD data affect downstream tasks such as classification, for cases of OOD data of increasing severity?
(3) What can we learn about the sources of the failure of DL models in downstream tasks such as classification, in the presence of OOD inputs, for OOD data of increasing complexity?

4 Tools and Datasets

4.1 Tools

4.1.1 Pytorch-OOD

Pytorch-OOD [9] is a specialized toolkit designed for OOD detection. It provides comprehensive methods and techniques for OOD detection in deep learning models [9]. The range of algorithms and techniques it offers makes it a good fit for this study. Additionally, the library is designed to be flexible so that it can be customized for different domains and research. Furthermore, it provides additional tools for the performance evaluation of the OOD detectors, and these tools will be used to analyze different OOD detection methods on medical data.

4.1.2 Xplique

The Xplique [3] library offers powerful tools and techniques for explainability of how DL models behave, making it very useful in this research. It provides comprehensive methods and techniques such as GradCAM and SmoothGrad to reveal what happens in complicated deep neural networks. In this research, some explainability tools from this library will be used to reveal how a neural network fails with an OOD input.

4.2 Datasets

In order to have more generalized results, multiple medical image datasets will be used to explore the research questions. The datasets will differ in their size, dimensions of their data, and their content.

4.2.1 MedMNIST

The first tests will be conducted on Medical-MNIST (MedMNIST, a small and straightforward dataset that is ideal for use at the beginning of the study to produce preliminary results. MedMNIST is standardized to perform classification tasks on lightweight 28 x 28 images, which require no background knowledge, making it simple to experiment on [15].

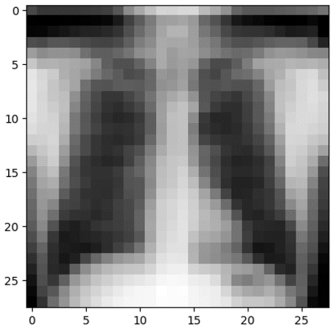

Fig. 1. An example image from MedMNIST dataset

4.2.2 Chest X-Ray Images (Pneumonia)

The dataset includes 5,863 chest X-ray images (JPEG), categorized into Pneumonia and Normal [8]. Images are organized into train, test, and validation sets. They were sourced from pediatric patients ages 1 to 5 at Guangzhou Women and Children's Medical Center.

Quality control was performed to exclude low-quality scans. Two expert physicians confirmed diagnoses, with a third expert verifying the evaluation set .

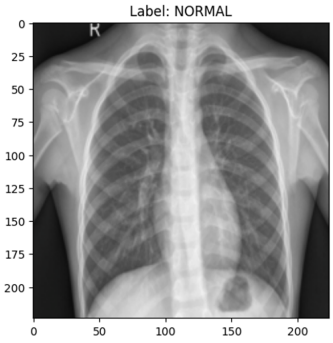

Fig. 2. An example image from Chest X-Ray dataset

4.2.3 *ISIC*

In this research, the ISIC 2018 Challenge dataset was used to evaluate out-of-distribution (OOD) detection and explainability methods. The dataset includes 1,000 dermoscopic images with labels for seven types of skin lesions, such as melanoma and benign keratosis [2]. Its diverse and well-annotated images make it suitable for testing how OOD detection models perform and understanding their limitations in medical image classification.

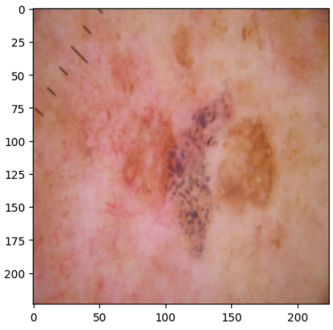

Fig. 3. An example image from ISIC dataset

5 Methods of Research

This section will explain the methodology used in this research in order to explain each research question. Separate experiments are conducted to address each research question, requiring separate methodologies.

5.1 Part 1: Making a Comparison of Various OOD Detection Methods On Medical Images

This part of the research will aim to answer research question 1. By adding some Gaussian noise to random images in our datasets mentioned in thesection4.2, OOD data will be created with increasing levels of severity in the context of distribution. There will be three levels of noise: low severity where the image is minimally distorted, medium severity with a substantial amount of noise, and high severity where the image is heavily distorted, making the features difficult to identify. Next, a pre-trained model will be imported and with using transfer learning it will be fine tuned for the dataset being used. Then various OOD detectors from the PyTorch-OOD library will be compared for accuracy and reliability on these medical images and the model. The tools from the same library will be used to assess and make comparison among the OOD detectors. After obtaining concrete results from preliminary experiments on MedMNIST,a smaller dataset, further tests will be conducted on the more extensive and complex datasets which are Chest X-Ray and ISIC.

5.2 Part 2: Analyzing the Effects of OOD Data In Classification Tasks

This part of the research will aim to answer research question 2. To evaluate the robustness of the DL model against varying severities of OOD data, we introduced different levels of noise into the input images. This approach simulates various OOD scenarios by altering the input data in a controlled manner. Gaussian noise was added to the images at incremental levels, ranging from low to high severity. These noisy images were then fed into the trained classification model to assess its performance for accuracy and confidence. By systematically increasing the noise, we aimed to understand how the model's accuracy and confidence degrade with increasing OOD severity and identify the thresholds at which the model's predictions become unreliable. This analysis provides insights into the robustness of the model and its ability to handle unexpected variations in the input data.

5.3 Part 3: Analyzing the Source of Failure of DL Models in Classification Tasks With OOD Inputs

This part of the research will aim to answer research question 3. To evaluate the sources of failure in downstream classification tasks with DL models, different types and severities of OOD data will be introduced as input to a pre-trained model. Various OOD scenarios will be simulated by adding Gaussian noise and also introducing new, unseen classes at varying levels of intensity. The model's performance will be assessed to identify sources of failure. To gain deeper insights into these failures, we will employ explainability tools such as GradCAM and SmoothGrad from the Xplique library. These tools will help visualize the regions of the input data that the model focuses on, highlighting the sources of correct and erroneous predictions. By analyzing these saliency maps, we aim to understand the underlying reasons for the failure of the model with OOD data.

6 Implementation

To answer the research questions mentioned insection3.1, several Python notebooks are created using Torch, Cuda, PyTorch, Xplique, and many other libraries. The local Jupyter server of the University of Twente and Google Colab environments are used for these notebooks and GPU-accelerated computing is utilized to deal with computationally intensive tasks.

6.1 Implementation of Part 1

For each dataset, the same process that is explained in the following subsubsections was applied.

6.1.1 Preperaing the Data

Once the dataset is loaded, all the images first go to a transformation. Initially, the images are resized to have identical dimensions, followed by their conversion into tensors, and finally, normalization is performed. Later, the images are split into train, test, and validation datasets to be used in the model training.

6.1.2 Gaussian Noise

Gaussian noise is a type of statistical noise characterized by a normal distribution, often used in image processing to simulate random variations in pixel intensity. When applied to images, it produces variations with a Gaussian (bell-curve) distribution [11]. Gaussian Noise is implemented in this study to add random noise to the images to create OOD instances with different severities. Low, medium, and high noise is used to use three different severities in this study. An example of an OOD instance created by noise can be seen in Figure 4.

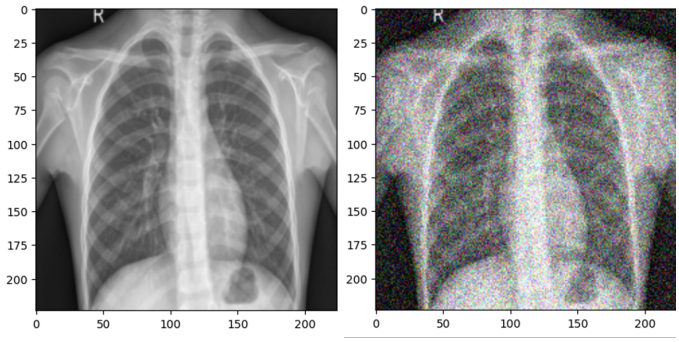

Fig. 4. Normal image (left) and a OOD instance created by noise (right)

6.1.3 Fine Tuning the Model

Wide ResNet-50-2 model, pre-trained on imagenet-1k, is used in the experiment [17]. The fact that Wide ResNet-50-2 was trained on big datasets and has a deep and wide architecture makes it suitable for transfer learning which will be used in our experiment [17]. The model is later moved to the specified device which was the GPU to leverage GPU acceleration. However, this model was not very familiar with our datasets. Therefore, we have done transfer learning and fine-tuned the model for more accurate results. For the fine-tuning process, we employed a training loop with a batch size of 32 and ran for 10 epochs, using the Adam optimizer with a learning rate of 0.001. The model was trained using cross-entropy loss, and the learning rate was dynamically adjusted with a scheduler based on validation performance. After each epoch, the model's performance was evaluated on the test set to monitor accuracy and ensure proper learning progression.

6.1.4 Evaluation

Various Pytorch-OOD detectors are implemented to be assessed for their performance of OOD detection. Additionally, OOD metrics for the Pytorch-OOD library is implemented to calculate various metrics.

6.2 Implementation of Part 2

To answer the second research question 2, the MobileNet-v2 model, pre-trained on Imagenet is implemented [12]. This model was a suitable choice for this research because of its efficiency, high accuracy, and versatility. Subsequently, the input images were prepared to be suitable for feeding into the model. Next, Gaussian Noise, as discussed in section 6.1.2, was applied to generate multiple versions of the same image, each with progressively higher noise levels, indicating the increasing severity of OOD.

6.3 Implementation of Part 3

To answer the third research question 3, the same steps as section6.2 are applied to obtain a model and preprocessing data. Furthermore, Xplique explainability tools mentioned in section 4.1.2 were implemented, such as GradCAM and SmoothGrad to obtain an answer to the third research question.

7 Experiment and Results

7.1 Part 1

In this section, the experiment carried out to answer the research question 1 and its results are explained.

7.1.1 Experiment of Part 1

Once all the steps in section 6.1 are completed, including data preprocessing, generating some OOD instances by introducing noise, and fine-tuning the models, the experiment is ready to be carried out. For each dataset, the same five OOD detectors were used in order to detect OOD occurrences. These detectors were chosen to represent a diverse range of approaches to OOD detection. For assessing their performance, a function that returns the metrics used in the state of the art was also implemented. The metric function computes AUROC, AUPR IN, AUPR OUT, which should ideally be high, and FPR@95TPR, which should ideally be low for optimal performance. Refer to the documentation of the PyTorch-OOD library for detailed explanations of these metrics [9]. OOD detectors that were assessed are:

- Monte Carlo Dropout (MCD)
- Maximum Softmax (MaxSoftmax)
- Maximum Logit (MaxLogit)
- Energy Based (EnergyBased)
- Entropy
- Nearest Neighbor (KNN)

These OOD detectors listed above implement different approaches and methods to detect OOD instances. MCD forward-propagates the input through the model N times with activated dropout and averages the results [4]. MaxSoftmax detects OOD instances by implementing the Maximum Softmax Probability Thresholding baseline [9]. MaxLogit implements the method mentioned in the paper Scaling Out-of-Distribution Detection for Real-World Settings which uses the highest raw logit value to measure confidence [7]. EnergyBased calculates the negative vector of logits to be later used as an outlier score [9]. Entropy method calculates the entropy based

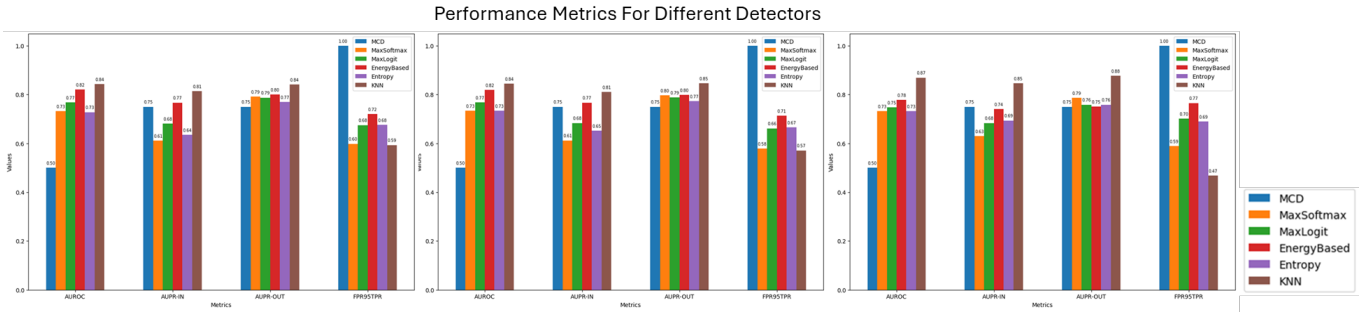

Fig. 5. MedMNIST results for sequentially low, medium, and high severity of OOD. X axis provides different metrics (respectively AUROC, AUPR IN, AUPR OUT, and FPR@95TP), and the detection methods are listed in each metric. The Y axis provides the values of each detector. The KNN method outperformed the other detectors by having higher accuracy in detecting in and out of distribution data and low values for FPR@95TPR. The Energy Based method also demonstrated high accuracy in detecting OOD instances.

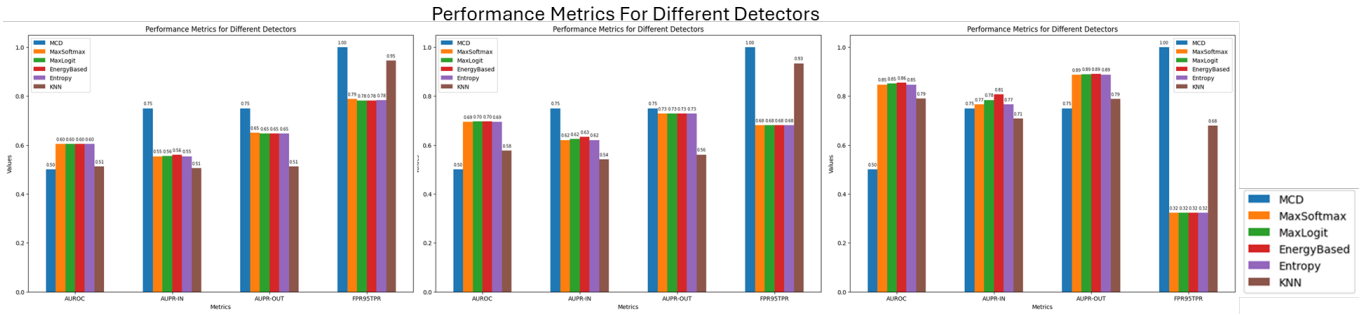

Fig. 6. Chest X-ray results for sequentially low, medium, and high severity of OOD. X axis provides different metrics (respectively AUROC, AUPR IN, AUPR OUT, and FPR@95TP), and the detection methods are listed in each metric. The Y axis provides the values of each detector. Overall, the detectors performed very well. Particularly, Maximum Softmax, Maximum Logit, Energy Based, and Entropy achieved high accuracy in detecting in and out distribution instances, with low values for FPR@95TPR, as desired.

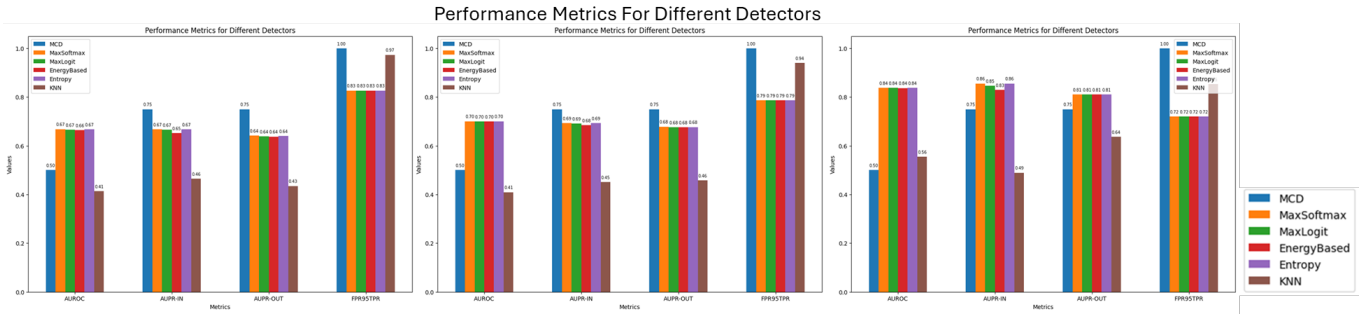

Fig. 7. ISIC results for sequentially low, medium, and high severity of OOD. X axis provides different metrics (respectively AUROC, AUPR IN, AUPR OUT, and FPR@95TP), and the detection methods are listed in each metric. The Y axis provides the values of each detector. The Nearest Neighbor method performed exceptionally poorly, worse than random guessing with mostly having lower accuracy than 0.5 on detecting in and out distribution instances and having high value for FPR@95TPR. In contrast, all other methods demonstrated consistent and effective performance in detecting out-of-distribution instances.

on logits of a classifier to asses data for being OOD [9]. Nearest Neighbor is a distance-based method that calculates the kth nearest neighbor distance between the embedding of each test image and the training set and use a threshold to decide if an input is OOD or not [14].

7.1.2 Results of Part 1

This section presents the performance evaluation of various OOD detection methods across three distinct datasets: MedMNIST, Chest X-ray, and ISIC. The results of the datasets can be observed in Figure 5, Figure 6, and Figure 7 consecutively.

Overall, all methods showed strong performance across all datasets, with a few minor exceptions. They consistently achieved high performance, with in- and out-of-distribution detection accuracy values between 0.7 and 0.8, particularly in datasets with high severity of out-of-distribution (OOD) samples. Moreover, their performance was still quite acceptable with in- and out-of-distribution detection accuracy values mostly above 0.6 in datasets with low severity of OOD, where the distinction between in- and out-of-distribution data was minimal, making them very challenging to distinguish from each other.

Across all datasets, the Probability-based approach (including Maximum Softmax and Entropy) and the Logit-based approach (including Maximum Logit and Energy Based methods) consistently demonstrated high accuracy in detecting out-of-distribution instances, making them the most preferable approaches for OOD detection in medical datasets.

Overall, the methods used from the PyTorch-OOD library performed accurately on the three chosen medical datasets. Most methods, particularly probability-based and logit-based approaches, accurately distinguished between in-distribution and out-of-distribution medical images, even at low levels of severity of OOD.

7.2 Part 2

In this section, the experiment carried out to answer the research question 2 and its results are explained.

7.2.1 Experiment of Part 2

One of the classes on which the MobileNet-v2 model was trained on, the Loggerhead Turtle, was selected to be used in the experiment. This choice was primarily made because of its distinct shape, features, and clear background. Subsequently, increasing levels of Gaussian Noise were introduced into the input image of the Loggerhead Turtle to create varying degrees of out-of-distribution severity. Eventually, 17 different input images were obtained. The initial image did not have noise, representing the original image. Subsequently, noise was incrementally added, resulting in increasing severity of out-of-distribution (OOD) conditions. The final image had significant noise, representing a completely out-of-distribution state. Then for each image, predictions were obtained by the model outputting the top 3 classes with their confidence level.

7.2.2 Results of Part 2

In Figure 8, it is observed that initially, without any noise, the image was classified as a Loggerhead Turtle with a confidence level of 95%. As the noise level increased and the image became more out-of-distribution, the confidence in accurately identifying the correct class diminished. At the noise level 10, the prediction showed equal confidence between the true class and another class, Bubble. After level 10, the input image became completely out-of-distribution and was predicted as various classes, mainly as a Bubble. This is because the high noise levels made the image of the Loggerhead Turtle resemble a round, blurry object.

In conclusion, it is clear that out-of-distribution (OOD) data significantly affect the model's classification performance. As the severity of OOD increases, both the confidence level and the accuracy of the classification drop substantially.

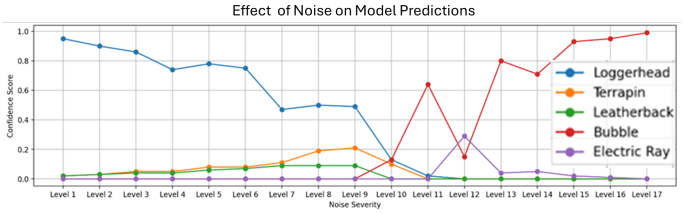

Fig. 8. Effect of increasing severity of OOD on model predictions. This graph illustrates how increasing Gaussian noise severity affects confidence scores for various classes (Loggerhead, Terrapin, Leatherback Turtle, Bubble, Electric Ray) in model predictions. Each line represents the confidence score of a single class.

7.3 Part 3

7.3.1 Experiment of Part 3

In this experiment, the sources of failure of DL models in classification, in the presence of OOD inputs were tried to be revealed by the explainibilty tools from Xplique 4.1.2. GradCAM and SmoothGrad were the two explainability tools that were used.

GradCAM generates visual explanations for convolutional neural networks by highlighting important regions in an image that influence the network's prediction [13].

SmoothGrad improves gradient-based explanations that build on work like Saliency Maps and GradCAM, by reducing noise. It averages saliency maps from multiple noisy versions of the input image, resulting in clearer and more reliable visualizations [6].

These two methods will be used on multiple input images in this experiment. These input images consist of one in-distribution image (the image of a cat) and various out-of-distribution images. These images will exhibit increasing severity of out-of-distribution characteristics by initially adding some noise, as shown in Figure 9, and including some completely unknown images and out-of-distribution for the model, as illustrated in Figure 10.

7.3.2 Results of Part 3

This section presents the result of the experiment that aims to answer the research question 3.

Figure 9 and 10 present the results for the part 3 experiment. Red or mainly the warmer areas are the areas that were most influential for the prediction of the model. It is observed that for the in-distribution image (the first image of the Figure 9), the warmer regions of the saliency maps correspond to the relevant and meaningful features of the image. This is particularly noticeable in the more detailed saliency maps generated by SmoothGrad.

However, in the case of all other out-of-distribution images created by noise, the warmer areas are dispersed throughout the image rather than being concentrated on meaningful features. In Figure 9, from left to right, the focus on significant features such as the eyes and chin becomes unclear and dispersed, thus the model cannot identify them as it can be observed from the saliency maps produced by SmoothGrad.

In the case of all out-of-distribution images that were completely unknown to the model, saliency maps have clustered in areas of the image that are not directly relevant or significant. This situation

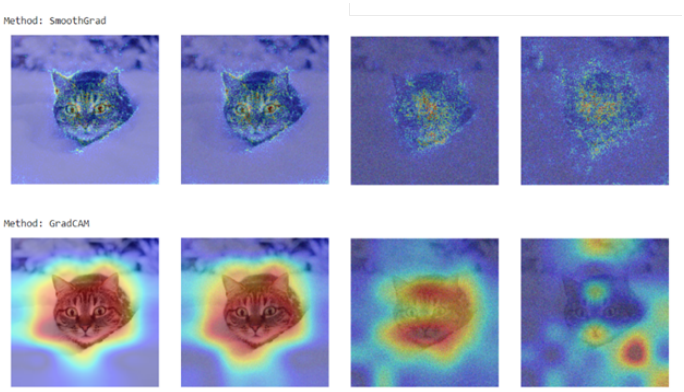

Fig. 9. Saliency maps created by methods SmoothGrad and GradCAM on different OOD instances. Warmer colors indicate the areas of focus.

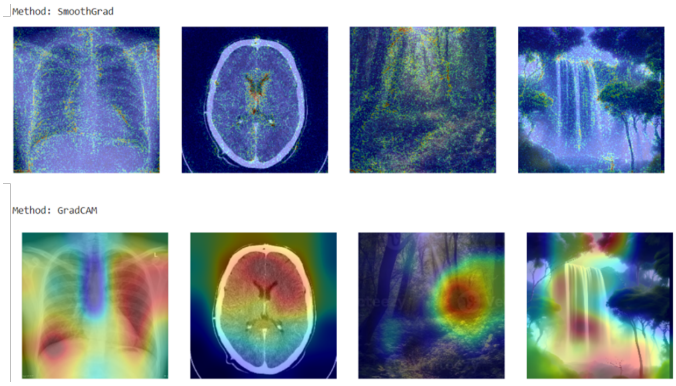

Fig. 10. Saliency maps created by methods SmoothGrad and GradCAM on different OOD instances created by noise and ordered with increasing severity of OOD. Warmer colors indicate the areas of focus.

occurs mainly because the model confuses certain parts of the image with features of other known classes. For example, this was seen in the brain scan in which the model incorrectly identified the image as a sundial due to its similar shape.

8 DISCUSSION AND CONCLUSION

8.1 Conclusion to Research Question 1

To answer the research question 1, the experiment explained in Section 7.1 was carried out. Multiple out-of-distribution detectors from the PyTorch-OOD library were evaluated on three distinct medical datasets that differ in size, shape, and content. The findings from the experiment indicated that OOD detectors implemented in the PyTorch-OOD library are highly effective across varying levels of OOD in medical images. In particular, the probability-based and logit-based approaches consistently performed with high accuracy. Even in a low severity of OOD where the distinction between in- and out-of-distribution is very challenging, the performance of the methods from these approaches remained acceptable.

KNN, a fetaure-based method, was the only one that was not consistent across all the datasets performing well in small images such as MedMNIST and poorly in bigger images from ISIC and Chest X-Ray. This is reasonable considering that small images such as MedMNIST have small future spaces that are less complex and it is easier to find an optimal threshold for distance metrics to make accurate decisions. However, these conditions are exactly the opposite in bigger images resulting in poor performance and creating inconsistency across datasets.

Overall, OOD detectors in the PyTorch-OOD library proved their reliability in medical imaging.

8.2 Conclusion to Research Question 2

To answer the research question 2, the experiment explained in Section 7.2 was carried out. The findings based on the results suggest that OOD data profoundly impacts model performance. As the severity of OOD increases, the model's confidence and accuracy level in the classification drops significantly. This highlights the need for robust OOD detection systems in real-world applications in which the OOD data are common.

8.3 Conclusion to Research Question 3

To answer the research question 3, the experiment explained in section 7.3 was carried out. Different cases and severity of out-of-distribution data were given to the model. For the in-distribution image, the explainability tools effectively highlighted the meaningful and relevant features, indicating that the model predictions were based on sensible and appropriate parts of the image, as expected.

In the case of out-of-distribution images created by noise, the focus of the model was dispersed. The model was unable to maintain focus on relevant parts and identify them. This led to a decrease in the confidence level in prediction for low severity of OOD and the failure to accurately classify for high severity of OOD.

In the case of out-of-distribution images that are completely unknown to the model, the saliency maps showed focus on random parts of the image that resemble features of known classes due to the model's tendency to incorrectly associate unfamiliar patterns with familiar ones. Thus, misleading visual similarities were concluded to be the source of failure in classification.

8.4 General Conclusion

This study comprehensively examined the effectiveness of OOD detection methods from the PyTorch-OOD library, particularly in medical image classification tasks. Through a series of experiments on OOD detection across various medical datasets, PyTorch-OOD demonstrated high accuracy and reliability in identifying OOD instances in medical imaging.

Further analyses were conducted to investigate the impact of OOD on classification tasks. The experimental results demonstrated a substantial effect of OOD on model performance, particularly in prediction confidence and accuracy.

Finally, explainability tools were utilized to investigate the sources of failure in classification tasks in which OOD inputs are involved. As mentioned in section 8.3, models encounter difficulties and sometimes fail in classification tasks when faced with OOD data containing noise or misleading visual similarities.

Overall, the findings underscored the critical role of OOD detection in enhancing the robustness and reliability of deep learning models in medical imaging for accurate and reliable applications in healthcare. Furthermore, the findings highlighted the strong performance of PyTorch-OOD in OOD detection in the context of healthcare DL applications.

8.5 Value of This Study

This study contributes to the field of medical imaging and deep learning by addressing a critical issue of OOD detection. This study offers insights into the effectiveness of different OOD detection methods from the PyTorch-OOD library in medical imaging. In this field, this library was not frequently evaluated. Additionally, this study provides insight into how deep learning models perform in classification tasks with OOD data and identifies sources of failure. These insights contribute to advancing OOD detection methods, aiming to develop more reliable and accurate systems against OOD data.

8.6 Shortcomings and Future Work

While this study offered valuable insights into OOD detection in medical imaging, it encountered several limitations that will be addressed in this section to support future research.

8.6.1 Part 1

Even though several medical datasets with varying contents were used, the diversity of these datasets does not fully represent the diversity in real-world clinical settings. The experiments conducted in this study do not cover a wide range of medical imaging contents, and thus the results may not be fully reliable for broader generalizations. Future studies with more resources should include a wide range of medical image datasets making the research more comprehensive.

This study explored certain methods that have been shown to work well in regular datasets assuming that they will also work well in medical datasets. However, that might not be the case since the performance of methods sometimes deviates drastically with domain shifts. In other words, certain methods that were not the best in regular datasets could be effective on medical datasets. Therefore, for future studies, it is recommended to explore as many OOD detection techniques as possible to the extent that resources allow it to obtain more comprehensive and universally applicable results.

Another limitation of this study was its use of one type of OOD in this part which was created by noise. Even though this provides useful insights into the OOD detection techniques, many more different types of OOD instances occur in the real medical world. For example, one of the common OOD instances is having scans from different hospitals that vary in imaging protocols, equipment, or patient demographics. These type of cases should also be tested to ensure comprehensive evaluation and robustness. For future studies, it is advised to communicate with clinicians and gather information on which type of OOD is faced the most in the real world and have a more comprehensive study on these various types.

8.6.2 Part 2

This section of the research was also limited by the lack of variety with datasets and different types of OOD instances, just as mentioned in section 8.6.1. Furthermore, the experiment is only done with one single pre-trained model. Since different models behave differently, it is better to investigate the research question 2 in a wide range of models. In further studies, the research should be carried out with a bigger variety of datasets, types of OOD, and models.

8.6.3 Part 3

This section of the research was also limited by the use of a single model and limited variation in types of OOD instances. The paper has provided useful insight into the sources of failure in deep learning models in the presence of OOD data. It has considered several variations in types of OOD. However, the diversity of models used and variations in types of OOD instances is highly recommended to be extended for future studies to identify all possible sources of failure.

References

[1] Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. 2020. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250* (2020).

[2] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. 2018. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368* (2018). https://arxiv.org/abs/1902.03368

[3] Thomas Fel, Lucas Hervier, David Vigouroux, Antonin Poche, Justin Plakoo, Remi Cadene, Mathieu Chalvidal, Julien Colin, Thibaut Boissin, Louis Bethune, et al. 2022. Xplique: A deep learning explainability toolbox. *arXiv preprint arXiv:2206.04394* (2022).

[4] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059. https://proceedings.mlr.press/v48/gal16.html

[5] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. 2023. A framework for benchmarking class-out-of-distribution detection and its application to ImageNet. *arXiv preprint arXiv:2302.11893* (2023).

[6] Erick Galinkin. 2022. Robustness and usefulness in AI explanation methods. *arXiv preprint arXiv:2203.03729* (2022).

[7] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2019. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132* (2019).

[8] Daniel Kermany, Kang Zhang, and Michael Goldbaum. 2018. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. https://doi.org/10.17632/rscbjbr9sj.2

[9] Konstantin Kirchheim, Marco Filax, and Frank Ortmeier. 2022. Pytorch-ood: A library for out-of-distribution detection based on pytorch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4351–4360.

[10] Simon Kristoffersson Lind, Rudolph Triebel, Luigi Nardi, and Volker Krueger. 2023. Out-of-Distribution Detection for Adaptive Computer Vision. In *Scandinavian Conference on Image Analysis*. Springer, 311–325.

[11] Ameen Mohammed Abd-Alsalam Selami and Ahmed Freidoon Fadhil. 2016. A study of the effects of gaussian noise on image features. *Kirkuk Journal of Science* 11, 3 (2016), 152–169.

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[14] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*. PMLR, 20827–20840.
[15] Jiancheng Yang, Rui Shi, and Bingbing Ni. 2021. MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 191–195.
[16] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision* (2024), 1–28.
[17] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

A Appendix: Use of AI Tools

During the preparation of this work, Berk Imre used ChatGPT, TexGPT and QuillBot to generate example code and paraphrase certain pieces of text. Almost all the sections of this paper utilized AI assistance to increase coherency and readability while keeping the originality and credibility of its ideas. After using these tools, Berk Imre reviewed and edited the content as needed and take full responsibility for the content of the work.