# Energy-Balance-Based Out-of-Distribution Detection of Skin Lesions

Jiahui Sun[1], Guan Yang[2*], Yishuo Chen[3], Hongyan Wu[4], Xiaoming Liu[5]
School of Computer Science, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China[1, 2]
School of Artificial Intelligence, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China[1, 2, 5]
Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou, Henan 450007, China[1, 2, 5]
School of Textiles, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China[3, 4]

***Abstract*—Skin lesion detection plays a crucial role in the diagnosis and treatment of skin diseases. Due to the wide variety of skin lesion types, especially when dealing with unknown or rare lesions, models tend to exhibit overconfidence. Out-of-distribution (OOD) detection techniques are capable of identifying lesion types that were not present in the training data, thereby enhancing the model's robustness and diagnostic reliability. However, the issue of class imbalance makes it difficult for models to effectively learn the features of minority class lesions. To address this challenge, a Balanced Energy Regularization Loss is proposed in this paper, aimed at mitigating the class imbalance problem in OOD detection. This method applies stronger regularization to majority class samples, promoting the model's learning of minority class samples, which significantly improves model performance. Experimental results demonstrate that the Balanced Energy Regularization Loss effectively enhances the model's robustness and accuracy in OOD detection tasks, providing a viable solution to the class imbalance issue in skin lesion detection.**

*Keywords—Balanced energy regularization loss; skin lesions; out-of-distribution detection; convolutional neural networks*

## I. Introduction

Early detection and regular monitoring of skin lesions, one of the most common diseases in daily life, is of great importance; this not only helps to improve cure rates and develop accurate treatment plans but also effectively reduces mortality [1]. Especially for melanoma, the most lethal form of skin cancer, this importance is particularly pronounced in study [2]. When using convolutional neural networks (CNNs) for skin lesion detection, only data with known specific distributions are exposed during training; however, due to the characteristics of skin lesions, including the diversity and complexity of their presentation [3], the actual distribution of data in the clinical setting is often uncertain. This poses a challenge with limited training data, which is often insufficient to fully cover the variety of skin lesions encountered. In addition, the distributions of the training data and the actual clinical data may differ significantly, further complicating the task. The predictive performance of the model is greatly reduced when faced with the significant challenges posed by different data distributions [4]. By employing an effective out-of-distribution (OOD) detection method, the model can identify new data that are different from the distribution of the training data. This detection mechanism enables the model to perform the special treatment or directly reject these OOD data for prediction, avoiding making wrong judgments on uncertain data, thus significantly improving the robustness and safety of the model [4] [5]. Effective OOD detection not only enhances the generalization ability of the model but also improves the reliability and robustness of the system in practical applications [6] [7]. Therefore, an effective tool in the diagnosis and management of skin lesions will be methods that can accurately detect OOD images of skin lesions.

In recent years, CNNs have made significant advances in the use of medical image data for disease diagnosis and analysis [8], and these models demonstrate performance comparable to that of professional physicians in the identification and classification of a wide range of common skin lesions [9], especially in binary and multiclassification tasks [10]. Specifically, CNNs can accurately differentiate between malignant melanoma, basal cell carcinoma, or other types of skin lesions [11]. However, OOD detection remains a challenging problem when confronted with skin lesions of unknown characteristics. Furthermore, for assessing the performance of models in different datasets, cross-dataset validation has not yet been widely applied to the OOD detection of skin lesions, which is crucial to ensure the generalizability and validity of the models given the potential differences in the data information in different datasets.

CNNs achieve great success in natural language processing and image recognition tasks, mainly due to their excellent feature learning, large-scale data processing, and generalization capabilities [12] [13]. Although the use of techniques such as self-attention mechanisms [14], regularization [15], and transfer learning [16] can significantly improve the performance of a model, their performance in OOD detection may still be unsatisfactory. This is because the features of the OOD samples are significantly different from the features of the distribution of the training data, which causes the model to be prone to prediction errors when dealing with these samples, thus reducing the robustness and reliability of the model. Therefore, a method called Balanced Energy Regularization Loss (BERL) [17] is applied to CNNs for OOD detection of skin lesions; The model is named energy-balance based OOD detection (EBOD).

In skin lesion image datasets, the class distribution is often highly imbalanced, with some lesion categories having a large number of samples, while others have relatively few. Traditional OOD detection methods struggle to effectively handle the disparity between majority and minority categories under such imbalanced distributions. The BERL method addresses this issue by introducing regularization based on the prior class

*Corresponding Author

probabilities, which enhances the regularization of majority class samples, thereby allowing the model to focus more on minority class samples and improving the detection accuracy for these categories. This method effectively mitigates the detection bias caused by class imbalance and optimizes the overall OOD detection performance. During the training process, the prior probability of each class is computed, allowing the model to adjust for the class distribution imbalance, thereby improving detection accuracy. Particularly in high-dimensional image datasets, the BERL enhances the model's robustness to unseen skin lesions, providing strong support for OOD detection of skin lesions. The successful application of this method not only generates significant impacts in skin lesion detection but also provides a new perspective for OOD detection in other medical imaging tasks. Through this technology, medical imaging systems are enabled to more accurately identify emerging disease types or lesions, thereby enhancing the intelligence and precision of medical diagnosis.

The remainder of this paper is structured as follows: Section II reviews the research advancements in the relevant field; Section III provides a detailed description of the proposed model's architecture, data acquisition and preparation process, as well as the evaluation metrics; Section IV discusses the model evaluation, computational cost, and presents the experimental results, accompanied by a comparative analysis with existing methods; Section V summarizes the main contributions of this paper, clarifies the motivation and potential advantages of the proposed method, and discusses the limitations of the research; Section VI outlines the directions for future research.

## II. Related Work

During the past few years, a variety of methods based on CNNs have emerged in the field of OOD detection. These methods are not only innovative in theory but also show excellent performance in practical applications. These methods can be broadly classified into the following groups: output score-based methods, generative model-based methods, adversarial training-based methods, and feature space-based methods, according to their basic principles and application characteristics.

### A. *Methodology Based on Output Scores*

Output score-based methods rely heavily on the output probability distribution of the classifier for the detection of OOD samples. This type of approach works by analyzing the confidence level of the classifier as it processes the samples, and samples with a low confidence level are considered to be possible OOD samples. Hendrycks and Gimpel [5] propose this method, which is widely used due to its simplicity and low computational cost. However, in some cases, such as skin lesion OOD detection, certain OOD samples may have higher softmax values, resulting in detection errors.

### B. *Generative Modeling-Based Methods*

Generative models detect OOD samples by learning the latent distribution of the data, and the variation autoencoder (VAE) method proposed by An and Cho [18] is a typical example. The VAE detects samples that do not match the distribution of the training data by reconstructing the data. This method is particularly suitable for OOD detection of medical image data and can identify rare or unseen lesion types.

### C. *Adversarial Training-Based Methods*

The adversarial training-based approach utilizes Generative Adversarial Networks (GANs) for OOD detection. The method proposed by Schlegl, Seeböck, and Waldstein [19] generates samples that are similar to normal data employing GANs and identifies OOD samples utilizing reconstruction errors. This approach significantly improves the sensitivity of the model to OOD samples and is well-suited for application in complex clinical settings.

### D. *Feature Space-Based Methods*

Feature space-based methods include the Mahalanobis distance-based method introduced by Lee et al. [20] and the One-Class Support Vector Machine (One-Class SVM) method developed by Schölkopf et al. [21]. The former identifies OOD samples by calculating the Mahalanobis distance of the input data in the feature space, which is suitable for feature extraction of high-dimensional data, but requires careful tuning of the distance metric and high computational cost. The latter separates most of the training data by constructing hyperplanes or decision boundaries to distinguish between normal and abnormal data, and is suitable for initial screening for OOD detection of skin lesions, but may not perform well on large and complex datasets.

In addition, several studies explore ways to enhance OOD detection by integrating multiple models. Xu et al. [22] present a deep integrated learning approach to enhance the accuracy and robustness of detection by combining the predictions of multiple deep neural network models. Dai et al. [23] designed a multimodal detection method utilizing multiple data sources (e.g., images, text, and clinical data) to improve the detection performance, and the combination of patient history, symptom descriptions, and image data in OOD detection of skin lesions can significantly improve the accuracy of OOD detection.

## III. Methodology

This section describes the three modules of the experiment in this study: data acquisition, model architecture, and model evaluation. The first section outlines the methodological steps adopted by the researchers in the data collection and analysis process, which is the core part of the study. The next subsections explain the specific steps of the study in terms of model architecture design and model evaluation, respectively. The experimental workflow shown in Fig. 1 provides a clear overview of the experimental process.

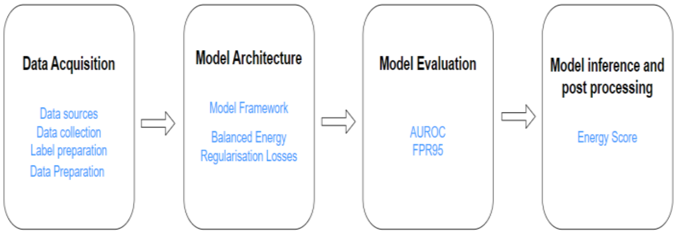

Fig. 1. Experiment module diagram.

## A. Data Acquisition

*1) Data sources:* The medical datasets used in this study are obtained from open-source databases, including but not limited to the International Skin Imaging Collaboration (ISIC), the Harvard University database (https://data.harvard.edu.dataverse), the Portuguese Pedro Hispano Hospital Dermoscopy Image Database (https://www.fc.up.pt/addi/project.html), and Stanford AIMI Shared Database (https://stanfordaimi.azurewebsites.net). These databases provide rich and diverse images of skin lesions for this study, ensuring broad applicability and reliability for model training and evaluation. To validate the external generalization ability of the model, several datasets of non-skin lesions are also obtained from the Kaggle platform (www.kaggle.com) for OOD detection. In terms of data use, this study strictly follows ethical principles to ensure that patient privacy is adequately protected, data security is effectively guaranteed, and patients' rights are respected. At the same time, this study actively promotes data sharing and open scientific research and assumes corresponding responsibilities and obligations.

*2) Data collection:* In this study, several open-source skin lesion datasets are used. These include images of different types of skin lesions as well as images of normal skin from different locations with manually created or corrected annotation information. From the ISIC2018, ISIC2019, ISIC2020, HAM10000 [24], PH2 [25], DDI [26], Dermnet [27], UMCG [28], and PAD-UFES-20 [29] datasets, several datasets containing multiple lesion types are screened. The prevalence of various skin lesions varies due to differences in the number of images of various lesion types in different datasets. In addition, abdominal MRI, brain tumor, kidney stone, and Places365 datasets are obtained from the Kaggle platform. These are used as auxiliary datasets [30] along with the skin lesion datasets for model training. Table I provides a summary of the fundamental characteristics of the datasets employed in this study. To ensure the breadth of the training datasets, 13 datasets containing four diseases and one non-disease are used during model training. For OOD data, three other disease datasets and two object detection datasets are used in this study: the colon adenocarcinoma dataset, the gastrointestinal disease dataset, the cataract dataset, the Street View House Numbers (SVHN) dataset [31], and the Cifar10 dataset [32].

TABLE I. BASIC CHARACTERISTICS OF EACH SKIN LESION DATASET

| Dataset | Count of Types | Number of Photos | Source |
|---|---|---|---|
| ISIC2018 | 7 | 11720 | ISIC |
| ISIC2019 | 8 | 25331 | ISIC |
| ISIC2020 | 5 | 33126 | ISIC |
| HAM10000 | 7 | 10015 | Harvard Dataset |
| PH2 | 2 | 200 | PH2 Dataset |
| Dermnet | 5 | 579 | Kaggle |
| PAD-UFES-20 | 5 | 2298 | GitHub |
| UMCG | 2 | 170 | MED-NODE Dataset |
| DDI | 6 | 656 | Stanfordaimi AIMI |

*3) Label preparation:* In the many open-source datasets on skin lesions, many different types of skin lesions are usually covered. However, the types of skin lesions that are recorded in the different datasets are not the same. It is not possible to train directly with these raw datasets as the model needs to be trained by lesion type for classification when performing training. For example, the ISIC2019 database contains eight different types of skin lesions. The HAM10000 database [24] contains seven types, and the lesion types differ between the two. To solve this problem, we use a strategy that requires a two-stage labeling process. Firstly, skin lesion types are classified according to the label files in each data set. When the same lesion is encountered but in different locations, it is combined into the same lesion type, and the original label file is reordered to generate a new label file based on the lesion type. Subsequently, the images in each of the datasets are then retrieved and organized according to these new label files. By using this method, each data set is divided into sub-datasets that contain multiple types of lesions. Finally, all datasets are regrouped and fused by skin lesion type to better support model training. Images of each type of skin lesion are shown in Fig. 2.

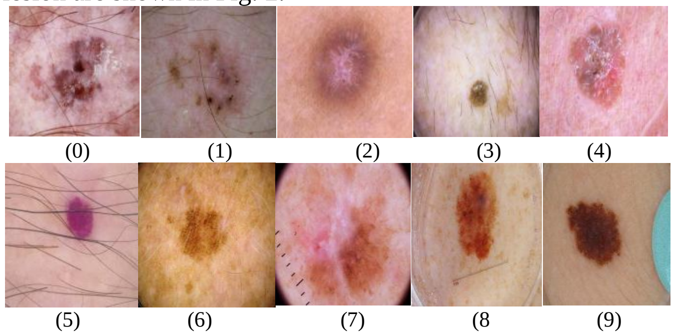

Fig. 2. Images of each type of skin lesion. 0: actinic keratosis; 1: basal cell carcinoma; 2: dermatofibroma; 3: seborrheic keratosis; 4: benign keratosis; 5: vascular lesions; 6: freckles; 7: squamous cell carcinoma; 8: melanoma; 9: melanocytic nevus.

*4) Data preparation:* All datasets used in this study show significant heterogeneity [33], involving differences in lesion characteristics, lesion sites, and recording devices. To reduce the impact of the differences between these datasets on the model, appropriate corrective measures are taken to standardize the datasets to ensure their compatibility with the model. When determining the size of the input image, the standard 224 × 224 size is selected. If the dimensions of the input image exceed 224 × 224, the image undergoes a process of cropping and scaling, referred to as center crop scaling [34], in order to align with the specifications of the model. The rationale behind the selection of this size is that image dimensions fluctuate across the datasets, rendering a uniform input size conducive to the model's capacity to discern pivotal characteristics. This uniformity simplifies data manipulation, reduces the complexity of computational operations, accelerates the convergence process, enhances the model's capacity to generalize, and facilitates the improvement of the model's performance. In both steps of training and detecting the model, the images are cropped and the newly generated data obtained from cropping is used for training and detection.

### B. Model Architecture

*1) Model framework:* The current model, similar to other models for OOD detection, uses neural networks to design appropriate OOD detection scores [4] [5] [35]. However, unlike most previous OOD detection methods that focus on designing OOD scores or introducing multiple outlier samples to retrain the model [36] [37], this study delves into the obstacle factors in OOD detection from the perspective of class imbalance in the auxiliary datasets [38]. To address the imbalance problem, a BERL is used, with different regularizations for each category of auxiliary data, to achieve reliable uncertainty estimates.

The training model employed in this study is primarily ResNet18 [39]. In comparison with other neural network models, ResNet18 demonstrates notable advantages in terms of feature extraction and generalizability. ResNet18 is a relatively deep network architecture that is capable of learning more abstract and complex feature representations through multi-level convolutional operations and feature extraction. This allows ResNet18 to extract more information from skin lesion data. In addition, ResNet18 is connected in a way that helps mitigate the problem of vanishing gradients and allows the network to be deeper. In the field of skin lesion OOD detection, residual connectivity and CNNs are both useful in enhancing the efficiency of feature capturing in images, which in turn leads to improved performance and greater model generalization. Although ResNet18 itself is not specifically designed for the detection of OOD, it can be appropriately adapted and enhanced to make it applicable to the detection of OOD [40]. ResNet18, with its deep convolutional architecture, is capable of effectively capturing latent patterns within skin lesion images, including both the intricate details and the overall morphology of the lesions. These patterns aid the model in distinguishing between in-distribution (known) and OOD (unknown) samples. During training, the residual connections in ResNet18 effectively alleviate the issue of overfitting, particularly when dealing with complex textures or lesion structures. This enables the extraction of robust and discriminative features, thereby enhancing the performance of OOD detection. Specifically, ResNet18 learns multi-scale features through convolution operations at different layers, from fine details to global representations. At lower layers, the network is capable of identifying minute textures and details on the skin surface, while at higher layers, it can capture the broader shapes of larger skin lesion areas. This hierarchical feature learning enables ResNet18 to accurately identify unknown lesion types in OOD detection tasks and effectively avoid misclassifying them as known types. The restrained ResNet18 model extracts feature from the image and feeds those features into a separate classifier to map the features of the image into a specific space and use the classification boundaries in that space to distinguish between known and unknown data. Finally, an energy function is introduced into the model for the calculation of OOD scores. Since the energy function does not require labeling information between known and unknown data, OOD detection can be performed without unknown data labels. In addition, energy functions usually have a good generalization ability to deal with different types of unknown data and to establish reasonable boundaries between the known data and the unknown data. The model structure is schematically shown in Fig. 3.

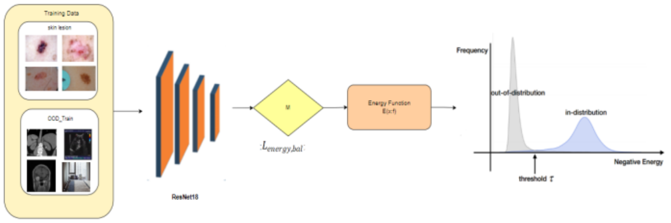

Fig. 3. Schematic diagram of the balanced energy regularization loss OOD detection model.

*2) Balanced energy regularization loss:* When regularizing auxiliary data, due to the imbalance of its class distribution, this may result in the model not being able to effectively learn information about the data of a few classes, thus affecting the model's ability to generalize to OOD data. In order to solve this problem, a variable M is introduced to measure whether the sample of the auxiliary data belongs to the majority class or the minority class [17]. In addition, the prior probabilities of the distributions of the auxiliary data were used to determine which class was to be categorized as a minority class. When this model is used to make inferences on the OOD auxiliary data, a statistical value $N_i$ can be obtained that indicates the number of samples that are classified in class $i$ . Then, the prior probability of the OOD distribution can be calculated using the following formula:

$$P(y=i\mid o)=\frac{N_i}{N_1+N_2+\cdots+N_K}\,. \tag{1}$$

For a neural network classifier $f$ , the a posteriori probability that the input image x belongs to a class $i$ is acquired by performing a softmax operation on the production of $f$ , which is,

$$P(y=i\mid \mathbf{x},o)=\frac{e^{f_i(\mathbf{x})}}{\Sigma_{j=1}^{K}e^{f_j(\mathbf{x})}}\,. \tag{2}$$

As the posterior probability that $\mathbf{x}$ belongs to the class $i$ increases, the probability that $\mathbf{x}$ belongs to the class $i$ increases accordingly. Similarly, if the prior probability of category $i$ is higher, then the probability of category $i$ becoming the majority category will increase. Thus, as the probability that $\mathbf{x}$ belongs to the majority class $i$ increases, the product of $\mathrm{P}\left(y=i\mid o\right)$ and $\mathrm{P}\left(y=i\mid \mathbf{x},o\right)$ must increase. Based on this result, the metric $\mathrm{M}$ for measuring the probability that $\mathbf{x}$ belongs to the majority class is defined as follows:

$$M=\Sigma_{j=1}^{K}P(y=j\mid \mathbf{x},o)P(y=j\mid o)\,. \tag{3}$$

Moreover, a hyperparameter $\lambda$ is introduced to model an additional generalized prior probability, which is used to regulate the degree of prior difference between categories. Ultimately, the generalized form $M_\lambda$ is as follows:

$$M_\lambda = \Sigma_{j=1}^{K} P(y = j \mid \mathbf{x}, o) P_\lambda (y = j \mid o) . \tag{4}$$

Where $P_\lambda(y = i \mid o) = L^1 norm\{ P^\lambda(y = i \mid o) \}$ . In order to ensure numerical reliability, the $\lambda$ -th power operation was performed on the a priori probability $\mathrm{P}\left(y = i \mid o\right)$, and the L1-normalization was applied [41]. When $\lambda$ = 0, the uniform prior probability model is established and $M_\lambda$ becomes a constant value $\frac{1}{K}$ , resulting in equal regularization strength across samples of different classes during the regularization process. When $\lambda$ > 0, $M_\lambda$ amplifies the regularization strength for the majority class while reducing it for the minority class, thus directing the model's focus more towards enhancing the OOD detection capability of majority class samples. In contrast, when $\lambda$ < 0, the inverse distribution of the prior probability will be modeled, and the regularization strength for the minority class samples is increased, allowing the model to better adapt to OOD samples from the minority class. The optimal value $\lambda$ varies across different OOD datasets. Generally, a larger value $\lambda$ indicates a greater prior difference between classes, which is more suitable for datasets dominated by the majority class. Conversely, smaller or negative values $\lambda$ are better suited for scenarios where the class distribution is more balanced or where the minority classes are of greater importance. With $\lambda$ growth, the prior probability gap among classes grows. According to the $M_\lambda$ component, the BERL is given by:

$$\begin{aligned} L_{\text{energy,bal}} &= L_{in,hinge} + L_{out,bal} \\ &= E_{(\mathbf{x}_{in},y)\sim D_{in}^{train}}\left[\left(\max\left(0, E\left(\mathbf{x}\right) - m_{in}\right)\right)^2\right] \\ &+ E_{\mathbf{x}\sim D_{in}^{train}}\left[\left(\max\left(0, E\left(\mathbf{x}\right) - m_{out} - \alpha M_\lambda\right)\right)^2 M_\lambda\right] \end{aligned} \tag{5}$$

where $E(x; f) = -T \cdot \log(\Sigma_{j=1}^{K} e^{f_j(\mathbf{x})/T})$ . In this formula, $f_j\left(x\right)$ denotes the logit output of the model for the input sample $x$ in class $j$ , while $T$ representing the temperature parameter, which is employed to smooth the logits distribution, By performing an exponentially weighted summation of the logits across all classes, the energy value for each sample is computed. Lower energy values are generally associated with OOD samples, whereas higher energy values are typically indicative of in-distribution samples. The energy loss $L_{\text{energy,bal}}$ of the model is the sum of both $L_{in,hinge}$ and $L_{out,bal}$ .

The key characteristic of the energy function lies in its ability to compute an energy value for each sample that is associated with its corresponding class. For in-distribution samples, the energy values are typically low because the predictions for these samples are usually accurate and consistent with the distribution of the training data. In contrast, the energy values for OOD samples are generally higher, as they do not conform to the distribution of the training data, thereby reflecting the greater uncertainty the model has regarding these samples. For instance, consider a pre-trained model designed to classify cats, dogs, and birds. When an image of a car is input, the model's logit output tends to be more dispersed, resulting in a higher energy value, which indicates that the model has lower confidence in classifying this sample. Conversely, when an image of a cat is input, the model's logit output is more concentrated, leading to a lower energy value, thus demonstrating the model's higher confidence in classifying this sample.

### *C. Model Evaluation*

The key metrics that researchers typically focus on when performing model evaluations include Recall, the area under the receiver operating characteristic curve (AUROC), and the false positive rate at 95% true positive rate (FPR95). Recall is used to measure the proportion of OOD samples that are correctly detected by the model out of all actual OOD samples.

*1) AUROC:* AUROC, on the other hand, represents the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR), calculated as the area under the ROC curve, which can synthesize the performance of the model under different classification thresholds. The value of AUROC ranges from 0.5 to 1. The closer the value is to 1, the better the ability of the model to discriminate.

*2) FPR95:* On the other hand, FPR95 focuses on the false alarm situation under high recall conditions, specifically calculating the false positive rate while maintaining a 95% true positive rate. FPR95, as a performance evaluation index, reflects the ability of the model to control the false alarm rate under the premise of guaranteeing a high detection rate in practical applications, and the smaller its value is, the lower the false alarm rate of the model is, thus proving the better performance of OOD detection.

Therefore, in this study, AUROC and FPR95 are used as the core metrics to assess the reliability of the model in OOD image detection of skin lesions. The FPR95 evaluates the ability to achieve high recall while controlling for false positives, while the AUROC provides an overall performance evaluation showing the average performance of the model across all thresholds. Recall as a base metric plays an important role in the calculation of AUROC and FPR95 as it has a direct impact on the results and performance analysis of these two metrics. The formulas for these indicators are shown below:

$$\mathrm{Re}\,call = TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$FPR95 = FPR\left(TPR = 0.95\right) \tag{8}$$

$$AUROC = \int_0^1 TPR(x)\, dx \ \left(x = FPR\right) \tag{9}$$

Where TP denotes true positive, TN denotes true negative, FP denotes false positive and FN denotes false negative.

### *D. Model Inference and Post Processing*

Although the model is trained using only the skin lesion datasets, it is equally capable of handling datasets from other diseases. The output generated by the model is an energy score that reflects the difference between the input sample data distribution and the known sample data distribution. Since the model is designed for OOD detection of skin lesions, its output can be interpreted as a measure of the model's classification accuracy in distinguishing between known and unknown samples, i.e., the model's ability to determine whether or not a sample is an OOD. The level of the energy score then indicates how confident the model is in classifying the input samples. Therefore, a post-processing step was used so that when the energy score is low, the model has a higher confidence that the input samples belong to known data, and conversely, when the energy score is high, the model has a higher confidence that the input samples belong to unknown data [42]. The course of the training and testing steps is summarized in Fig. 4.

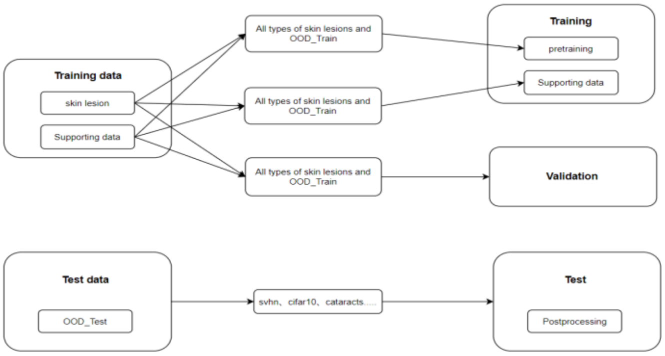
Training data
skin lesion
Supporting data
All types of skin lesions and OOD_Train
All types of skin lesions and OOD_Train
All types of skin lesions and OOD_Train
Training
pretraining
Supporting data
Validation
Test data
OOD_Test
svhn、cifar10、cataracts.....
Test
Postprocessing

Fig. 4. Overview of training and testing procedures.

In the context of OOD detection for skin lesions, the model may encounter challenging scenarios such as noise, occlusion, and small-sized OOD samples. When confronted with noise and small-sized OOD samples, the application of energy regularization loss results in greater regularization being imposed on these samples, while less regularization is applied to other samples, thereby maintaining the model’s ability to effectively detect OOD samples. In the case of occlusion, the integration of an attention mechanism enables the model to adaptively focus on the key regions of interest during training. Furthermore, by adjusting the class distribution of auxiliary data and weighting each sample according to the prior probability of its respective class, the model can effectively control the regularization strength across different classes, ensuring fairness in the data distribution.

## IV. Results

### *A. Performance Evaluation*

In order to evaluate the performance of the models, this experiment presents and analyzes the detection results of five models on OOD datasets from multiple domains, both medical and non-medical; the models used in this analysis are MSP [5], OE [43], OECC [44], Energy OOD [45], and EBOD. The FPR95 and AUROC metrics for the five models on each of the OOD data sets are shown in Table II. The experimental results show a slight difference in the performance of EBOD on OOD data in non-medical domains. However, the difference is not significant compared to its performance on OOD data in medical domains. This means that although there may be some differences in the performance of the EBOD model on OOD data from different domains, over the performance is relatively stable and has strong generalization capabilities. Compared to the other four models, the EBOD model shows significant performance gains on most OOD datasets. A sample of the OOD data used in the study is shown in Fig. 5, where colon_aca represents colon adenocarcinoma images, stomach represents gastrointestinal disorders, and cataracts represent cataract images.

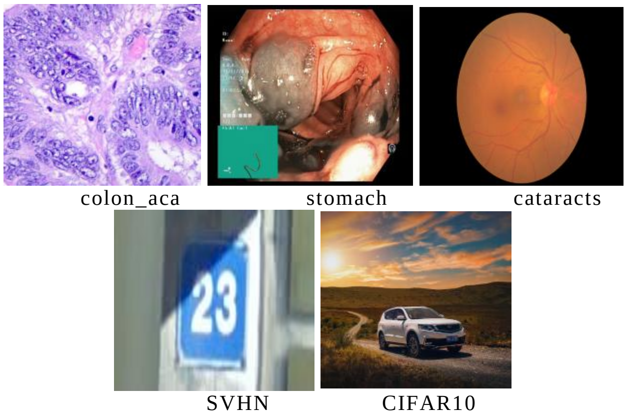
colon_aca stomach cataracts
SVHN CIFAR10

Fig. 5. Sample graph of OOD data.

In exploring the factors that enhance the performance of the model, the significant difference between the EBOD model compared to the WideResNet-based Energy OOD model [45] is the use of BERL for the auxiliary datasets. This innovative approach plays a key role in the optimization process and has a profound impact on the OOD detection performance of the model. A comparison of the FPR95 and AUROC metrics for the Energy OOD and EBOD models reveals that the incorporation of BERL is a pivotal factor in enhancing the OOD detection capabilities of the models. The EBOD model introduces BERL, which serves to balance the energy distribution of the auxiliary datasets during the training process. This significantly improves the model's generalization ability on different datasets. Table II details the results of the comparison of the models on different datasets. The FPR95 metrics for multiple OOD datasets are significantly lower when only the energy model is used and no BERL is introduced.

TABLE II. Evaluation Results of the Model on the OOD Datasets

| | MSP | OE | OECC | Energy OOD | EBOD (Ours) |
|---|---|---|---|---|---|
| colon_aca | 6.28 | 5.43 | 4.26 | 3.61 | 1.07 |
| stomach | 4.76 | 2.57 | 0.79 | 1.21 | 0.41 |
| cataracts | 3.55 | 1.93 | 2.87 | 2.08 | 0.49 |
| SVHN | 43.11 | 35.41 | 29.68 | 36.39 | 14.03 |
| CIFAR10 | 45.26 | 38.71 | 32.52 | 29.01 | 5.85 |

(a) FPR95

| | MSP | OE | OECC | Energy OOD | EBOD (Ours) |
|---|---|---|---|---|---|
| colon_aca | 88.53 | 92.81 | 95.83 | 95.02 | 99.36 |
| stomach | 89.37 | 93.46 | 94.29 | 92.38 | 99.85 |
| cataracts | 88.61 | 95.36 | 99.57 | 97.54 | 99.76 |
| SVHN | 86.83 | 90.19 | 90.68 | 91.49 | 95.83 |
| CIFAR10 | 88.15 | 91.72 | 92.64 | 92.78 | 98.53 |

(b) AUROC

### B. Calculate Cost

In the practical deployment of OOD detection for skin lesions, the computational cost is a crucial factor that determines model selection and deployment efficiency. To ensure the effectiveness and scalability of OOD detection methods in real-world applications, it is essential to optimize training and inference times, as well as reduce memory consumption, thereby minimizing computational overhead. Through the implementation of effective optimization strategies, the usability of the model in resource-constrained environments can be enhanced while maintaining its accuracy, thus meeting the demands of practical applications. Regarding training time, the BERL function proposed in this paper, compared to traditional methods, mitigates overfitting on minority class samples and excessive training on all class samples by precisely adjusting the regularization strength for each sample. This accelerates model convergence, thus reducing training time. In terms of inference time, the introduction of the M-value to quantify the likelihood of each sample belonging to a specific class, coupled with the adjustment of the loss function based on this value, reduces the computational complexity required for each sample during inference, thereby optimizing inference speed. Table III lists the calculated costs of different skin lesion OOD detection techniques.

TABLE III. CALCULATED COST OF DIFFERENT SKIN LESION OOD DETECTION TECHNIQUES

| | Training time (Sheets/ms) | Reasoning time (Sheets/ms) | MEM |
|---|---|---|---|
| La-OOD [46] | 4.57 | 1.46 | 6.87% |
| Bayesian [47] | 5.13 | 1.51 | 7.63% |
| Ours | 3.28 | 1.32 | 6.50% |

### C. Ablation Study

To validate the effectiveness of the proposed method, two ablation experiments were conducted. First, the BERL was introduced into DenseNet [48] for OOD detection of skin lesions. Subsequently, the BERL was removed from the EBOD model (i.e., EBOD-), and the same experiment was repeated. The experimental results demonstrate that the EBOD model exhibits superior performance in the OOD detection of various types of skin lesions, with the specific results summarized in Table IV.

TABLE IV. RESULTS OF ABLATION EXPERIMENTS

| | DenseNet | EBOD- | EBOD (Ours) |
|---|---|---|---|
| colon_aca | 2.61 | 4.39 | 1.07 |
| stomach | 1.82 | 2.07 | 0.41 |
| cataracts | 2.69 | 1.85 | 0.49 |
| SVHN | 34.50 | 42.37 | 14.03 |
| CIFAR10 | 27.14 | 16.29 | 5.85 |

(a) FPR95

| | DenseNet | EBOD- | EBOD (Ours) |
|---|---|---|---|
| colon_aca | 92.21 | 95.77 | 99.36 |
| stomach | 93.09 | 96.18 | 99.85 |
| cataracts | 95.61 | 94.80 | 99.76 |
| SVHN | 90.04 | 89.45 | 95.83 |
| CIFAR10 | 91.62 | 93.37 | 98.53 |

(b) AUROC

In contrast to previous studies, few studies synthesize multiple datasets on skin lesions by lesion type and further validate them using OOD datasets from a variety of different domains. Therefore, the present study is based on this innovation and improvement. Firstly, this study employs a methodology analogous to that employed in previous studies, whereby the training and test sets are rationalized in order to ensure a balanced data distribution and representative samples. Furthermore, this study introduces auxiliary datasets, which are employed to facilitate the model's ability to characterize a more expansive range of data, thereby enhancing its performance in the context of novel and previously unseen data. This enhanced generalization capacity enables the model to more effectively adapt to diverse application contexts and data distributions, and remains stable in the presence of noise, outliers, and other disturbances, thereby enhancing the model's resilience.

## V. DISCUSSION

This study introduces BERL into CNNs, aiming to accurately detect OOD data in skin lesions. EBOD achieves state-of-the-art performance in cross-database evaluations and demonstrates a high degree of accuracy, even under a wide range of special conditions. In comparing the model performance under different training methods, the introduction of energy balance regularization [17] plays an important role in improving the excellent performance of the model. In addition, EBOD has potential applications in other clinical OOD detection situations.

CNNs produce good results in several areas, such as natural language processing and image recognition [12][13]. In addition, its application is extended to the medical field, bringing great convenience to medical research and clinical practice [49][50]. CNNs have significant potential to improve the accuracy of medical image data analysis, which may have far-reaching implications in the field of medical image diagnosis [51]. The performance of EBOD may be overestimated due to the exclusion of certain types of skin lesions (e.g., common nevi) from the dataset, but the model demonstrates excellent performance in high noise and strong motion environments. Interpreting skin lesions in practical clinical applications is a challenging and complex task. Therefore utilizing this model can potentially reduce the misdiagnosis rate of skin lesions in clinical practice and enhance patient care.

It is well known that the application of energy regularization techniques plays a crucial role in CNNs. Despite the demonstrated efficacy of energy regularization techniques in many other fields, there remains a paucity of research investigating their application to the domain of medical image processing. Inspired by the obstacles in OOD detection caused by a class imbalance in auxiliary datasets and internal mechanisms of the model, this study employs BERL to enhance the performance of CNNs in OOD detection. The introduction of BERL lays the foundation for applying CNNs to OOD detection of medical images. If OOD detection is performed on multiple categories of medical images in the future, it is possible to achieve more efficient performance. Skin lesions are complex in shape and type, and especially under actual clinical conditions, OOD detection of skin lesions is more important than identification to minimize panic in the minds of patients and the rate of misdiagnosis.

In the case of ancillary data, especially in real-world scenarios, there is often an imbalance in the distribution of classes in the ancillary OOD data, e.g. there is a significant difference in the amount of data in the two classes. It can be difficult to effectively capture the diversity of the auxiliary samples of the OOD data using the traditional methods of cross-entropy loss and regularized loss. The model tends to learn the features of samples from a numerically larger number of categories while ignoring the features of samples from a numerically smaller number of categories, which reduces the model's ability to generalize when dealing with samples from unknown distributions. To overcome this problem, this study uses BERL to apply higher regularization to the data of the more numerous categories in the auxiliary data to ensure that the features of the data of the less numerous categories are also adequately extracted. This approach improves the robustness of the model and makes its performance more stable in the face of a variety of unknown distribution samples.

The evaluation metrics are determined based on the information from each OOD data after it has been tested by the model, which is critical to understanding and measuring the overall performance of the model. Previous research has shown that many detection models tend to produce overly confident prediction results when confronted with OOD data, resulting in less reliable detection of this OOD data. If the skin lesion features are similar in each training dataset, the model may suffer from overfitting, which weakens its generalization ability and leads to poor performance in OOD detection. As open-source datasets continue to proliferate in the medical field, researchers are able to utilize datasets with more comprehensive and diverse lesion characteristics, providing a valuable resource for OOD detection research. The current study involves skin lesion information from multiple datasets, which helps to ensure diversity in the training dataset and significantly improves the stability and robustness of the model. In the OOD detection model, the diversity of training data samples has a particularly significant impact on the detection results. The diversity of the training data allows the model to better recognize data with unknown distributions, demonstrating the potential value and reliability of the model in real-world applications.

Despite the progress made in this study in the field of OOD detection of skin lesions, however, there are still some limitations that need to be further explored and addressed. First, the limited size of the dataset used may not adequately represent the diversity of skin lesions in actual clinical settings. Therefore, when confronted with certain rare or emerging skin lesion types, the generalization ability and detection accuracy of the model may be insufficient. Moreover, in the OOD detection of skin lesions, the similarity between images presents a complex and challenging issue. In particular, the high similarity between certain non-skin lesion images and skin lesion images often leads to misclassifications in OOD detection. For example, varicose veins may cause the appearance of red or purple net-like patches on the skin surface, accompanied by localized swelling, which is prone to be misjudged as hemangiomas or purpuric skin lesions. Similarly, in some cases, lymphadenitis may lead to the formation of pustule-like or erythematous areas on the skin surface, especially when skin changes caused by enlarged lymph nodes closely resemble those of skin lesions, leading to incorrect identification as ulcers or nodules. Given the visual similarity between these non-skin lesions and actual lesions, effectively distinguishing these similar types of lesions has become a significant challenge in the OOD detection task for skin lesions. Finally, factors such as noise and image quality differences that may be encountered in practical applications are not yet fully considered in this study. Therefore, future research should focus on expanding the size of the dataset, improving the generalization ability of the model, exploring more different models, and testing them under conditions closer to clinical application scenarios to further validate and improve the performance of the model.

## VI. Conclusion and Future Research

The current study is testing the accuracy of depth models for OOD detection of skin lesions, specifically for the detection of unknown distribution skin lesion images from known distribution skin lesion images. The results of the study show that the models tested exhibit almost similar performance. However, the best-performing model was EBOD, which significantly outperformed the other models in both the AUROC and FPR95 metrics. Future research should focus on the creation of larger databases and the expansion of the variety of skin lesions used to train the models, which could help the models learn a wider range of features and allow them to better understand the differences between known distribution data and unknown data, thus improving their performance in OOD detection. This means that the model not only performs well on training data but also maintains high accuracy on unknown distribution data. As more and more research is devoted to OOD detection modeling, researchers believe that more advanced algorithms will emerge to improve the accuracy and stability of the models and make them perform better in the face of different types of OOD data. This study provides a basis for further research on OOD detection of skin lesions using depth modeling and demonstrates its great potential for medical applications.

## References

[1] A. Brunssen, A. Waldmann, N. Eisemann, and A. Katalinic, "Impact of skin cancer screening and secondary prevention campaigns on skin cancer incidence and mortality: a systematic review," Journal of the American Academy of Dermatology, vol. 76, no. 1, pp. 129-139, 2017.

[2] C. Di Raimondo, F. Lozzi, P. P. Di Domenico, E. Campione, and L. Bianchi, "The diagnosis and management of cutaneous metastases from melanoma," International Journal of Molecular Sciences, vol. 24, no. 19, p. 14535, 2023.

[3] A. Courtney, D. J. Lopez, A. J. Lowe, Z. Holmes, and J. C. Su, "Burden of disease and unmet needs in the diagnosis and management of atopic dermatitis in diverse skin types in Australia," Journal of Clinical Medicine, vol. 12, no. 11, p. 3812, 2023.

[4] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," In Proc. 6th Int. Conf. Learning Representations (ICLR), 2018.

[5] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," International Conference on Learning Representations, 2022.

[6] M. Yi et al., "Improved OOD generalization via adversarial training and pretraining," International Conference on Machine Learning, PMLR, pp. 11987-11997, Jul. 2021.

[7] J. Liu et al., "Towards out-of-distribution generalization: A survey," arXiv preprint arXiv:2108.13624, vol. 1, no. 1, p. 1, 2021.

[8] C. Zhang et al., "Enhancing lung cancer diagnosis with data fusion and mobile edge computing using DenseNet and CNN," Journal of Cloud Computing, vol. 13, no. 1, p. 91, 2024.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017.

[10] L. Hoang, S. H. Lee, E. J. Lee, and K. R. Kwon, "Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare," Applied Sciences, vol. 12, no. 5, p. 2677, 2022.

[11] D. Moturi, R. K. Surapaneni, and V. S. G. Avanigadda, "Developing an efficient method for melanoma detection using CNN techniques," Journal of the Egyptian National Cancer Institute, vol. 36, no. 1, p. 6, 2024.

[12] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," Artificial Intelligence Review, vol. 53, pp. 5455-5516, 2020.

[13] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6999-7019, 2021.

[14] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3286-3295.

[15] R. Moradi, R. Berangi, and B. Minaei, "A survey of regularization strategies for deep models," Artificial Intelligence Review, vol. 53, no. 6, pp. 3947-3986, 2020.

[16] S. A. Munoz, J. Park, C. M. Stewart, A. M. Martin, and J. D. Hedengren, "Deep transfer learning for approximate model predictive control," Processes, vol. 11, no. 1, p. 197, 2023.

[17] H. Choi, H. Jeong, and J. Y. Choi, "Balanced energy regularization loss for out-of-distribution detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15691-15700.

[18] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," Special Lecture on IE, vol. 2, no. 1, pp. 1-18, 2015.

[19] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," International Conference on Information Processing in Medical Imaging, Cham: Springer International Publishing, pp. 146-157, May 2017.

[20] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," Advances in Neural Information Processing Systems, vol. 31, pp. 7167-7177, 2018.

[21] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Neural Computation, vol. 13, no. 7, pp. 1443-1471, 2001.

[22] C. Xu, F. Yu, Z. Xu, N. Inkawhich, and X. Chen, "Out-of-Distribution Detection via Deep Multi-Comprehension Ensemble," arXiv preprint arXiv:2403.16260, vol. 1, no. 1, p. 1, 2024.

[23] Y. Dai, H. Lang, K. Zeng, F. Huang, and Y. Li, "Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection," Findings of the Association for Computational Linguistics*: EMNLP 2023*, pp. 5292-5305, Dec. 2023.

[24] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, no. 1, pp. 1-9, 2018.

[25] E. M. Senan, M. E. Jadhav, and A. Kadam, "Classification of PH2 images for early detection of skin diseases," In 2021 6th international conference for convergence in technology (I2CT), pp. 1-7, Apr. 2021.

[26] R. Daneshjou et al., "Disparities in dermatology AI performance on a diverse, curated clinical image set," Science Advances, vol. 8, no. 31, p. eabq6147, 2022.

[27] A. Aboulmira, H. Hrimech, and M. Lachgar, "Comparative Study of Multiple CNN Models for Classification of 23 Skin Diseases," International Journal of Online & Biomedical Engineering, vol. 18, no. 11, 2022.

[28] A. A. D. Alsaeed, "On the development of a skin cancer computer aided diagnosis system using support vector machine," Biosci. Biotechnol. Res. Commun., vol. 12, pp. 297-308, 2019.

[29] A. G. Pacheco et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," Data in Brief, vol. 32, p. 106221, 2020.

[30] J. Mitros and B. Mac Namee, "On the importance of regularisation and auxiliary information in OOD detection," International Conference on Neural Information Processing, Cham: Springer International Publishing, pp. 361-368, Dec. 2021.

[31] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," Proceedings of the 21st international conference on pattern recognition (ICPR2012), pp. 3288-3291, Nov. 2012.

[32] V. Thakkar, S. Tewary, and C. Chakraborty, "Batch normalization in convolutional neural networks—A comparative study with CIFAR-10 data," 2018 fifth international conference on emerging applications of information technology (EAIT), pp. 1-5, Jan. 2018.

[33] L. Peng, G. Wang, and C. Zou, "Measuring, testing, and identifying heterogeneity of large parallel datasets," Statistica Sinica, vol. 33, no. 4, pp. 2787-2808, 2023.

[34] T. N. Minh, M. Sinn, H. T. Lam, and M. Wistuba, "Automated image data preprocessing with deep reinforcement learning," arXiv preprint arXiv:1806.05886, vol. 1, no. 1, p. 1, 2018.

[35] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 13911-13922, Dec. 2019.

[36] Y. Zhu et al., "Boosting out-of-distribution detection with typical features," Advances in Neural Information Processing Systems, vol. 35, pp. 20758-20769, 2022.

[37] T. Wei, B. L. Wang, and M. L. Zhang, "EAT: Towards Long-Tailed Out-of-Distribution Detection," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 14, pp. 15787-15795, Mar. 2024.

[38] S. M. Xie et al., "In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness," arXiv preprint arXiv:2012.04550, vol. 1, no. 1, p. 1, 2020.

[39] A. Ullah, H. Elahi, Z. Sun, A. Khatoon, and I. Ahmad, "Comparative analysis of AlexNet, ResNet18 and SqueezeNet with diverse modification and arduous implementation," Arabian Journal for Science and Engineering, vol. 47, no. 2, pp. 2397-2417, 2022.

[40] S. Regmi et al., "ReweightOOD: Loss Reweighting for Distance-based OOD Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 131-141, 2024.

[41] S. Wu et al., "L1-norm batch normalization for efficient training of deep neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 7, pp. 2043-2051, 2018.

[42] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," arXiv preprint arXiv:1802.04865, vol. 1, no. 1, p. 1, 2018.

[43] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," arXiv preprint arXiv:1812.04606, vol. 1, no. 1, p. 1, 2018.

[44] A. A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier exposure with confidence control for out-of-distribution detection," Neurocomputing, vol. 441, pp. 138-150, 2021.

[45] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," Advances in Neural Information Processing Systems, vol. 33, pp. 21464-21475, 2020.

[46] H. Wang, C. Zhao, X. Zhao, and F. Chen, "Layer adaptive deep neural networks for out-of-distribution detection," Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 526-538, May 2022.

[47] A. T. Nguyen, F. Lu, G. L. Munoz, E. Raff, C. Nicholas, and J. Holt, "Out of distribution data detection using dropout bayesian neural networks," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 7, pp. 7877-7885, June 2022.

[48] Y. Zhu and S. Newsam, "Densenet for dense flow," 2017 IEEE International Conference on Image Processing (ICIP), pp. 790-794, Sep. 2017.

[49] A. Nogales, A. J. Garcia-Tejedor, D. Monge, J. S. Vara, and C. Anton, "A survey of deep learning models in medical therapeutic areas," Artificial Intelligence in Medicine, vol. 112, p. 102020, 2021.

[50] C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," Multimedia Systems, vol. 27, no. 4, pp. 599-613, 2021.

[51] S. S. Kshatri and D. Singh, "Convolutional neural network in medical image analysis: a review," Archives of Computational Methods in Engineering, vol. 30, no. 4, pp. 2793-2810, 2023.