scientific reports

Check for updates

OPEN

Effective data selection via deep learning processes and corresponding learning strategies in ultrasound image classification

Hyunju Lee[1], Jin Young Kwak[1] & Eunjung Lee[2]✉

In this study, we propose a novel approach to enhancing transfer learning by optimizing data selection through deep learning techniques and corresponding innovative learning strategies. This method is particularly beneficial when the available dataset has reached its limit and cannot be further expanded. Our approach focuses on maximizing the use of existing data to improve learning outcomes which offers an effective solution for data-limited applications in medical imaging classification. The proposed method consists of two stages. In the first stage, an original network performs the initial classification. When the original network exhibits low confidence in its predictions, ambiguous classifications are passed to a secondary decision-making step involving a newly trained network, referred to as the True network. The True network shares the same architecture as the original network but is trained on a subset of the original dataset that is selected based on consensus among multiple independent networks. It is then used to verify the classification results of the original network, identifying and correcting any misclassified images. To evaluate the effectiveness of our approach, we conducted experiments using thyroid nodule ultrasound images with the ResNet101 and Vision Transformer architectures along with eleven other pre-trained neural networks. The proposed method led to performance improvements across all five key metrics, accuracy, sensitivity, specificity, F1-score, and AUC, compared to using only the original or True networks in ResNet101. Additionally, the True network showed strong performance when applied to the Vision Transformer and similar enhancements were observed across multiple convolutional neural network architectures. Furthermore, to assess the robustness and adaptability of our method across different medical imaging modalities, we applied it to dermoscopic images and observed similar performance enhancements. These results provide evidence of the effectiveness of our approach in improving transfer learning-based medical image classification without requiring additional training data.

Keywords Data selection, Two-step decision making process, True network

Convolutional neural networks (CNNs) have emerged as the primary technique for deep learning in computer vision owing to their remarkable performance in the 2012 ImageNet competition[1]. Although CNN-based image analysis has been successful in medical image analysis, the full training of CNNs requires a large labeled dataset and extensive computational and memory resources, which can be challenging for medical data analysis[2]. Transfer learning has thus been widely used as an alternative, where the parameters of well-trained CNN models on the vast ImageNet dataset are transferred to a CNN model on medical images[3]. Specifically, the convolutional layers of a well-trained CNN model are frozen or fine-tuned, while the fully connected layers are trained for medical images. This approach has shown reasonable performance in various medical image modalities (e.g., X-ray, MRI, CT, Ultrasound, etc.) using well-trained CNN models such as AlexNet[4], VGGNet[5], ResNet[6], GoogLeNet[7], and Inception[8].

The experimental data in this paper is the thyroid nodule ultrasound images. Thyroid nodules are abnormal growths of cells in the thyroid gland that have become a significant medical concern, with a notable increase in incidence over the last few decades[9,10]. Several techniques have been developed to classify thyroid nodules

[1]Department of Radiology, Severance Hospital, Research Institute of Radiological Science, College of Medicine, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. [2]School of Mathematics and Computing, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. ✉email: eunjunglee@yonsei.ac.kr

as benign or malignant, with CNN being widely employed for nodule diagnosis based on medical imaging, especially ultrasound images[11]. Because ultrasound imaging is safe and cost-effective for rapid diagnosis, many types of research using CNN use ultrasound images for diagnosing thyroid nodules[12]. Many studies classified thyroid nodules by properly utilizing fine-tuning and feature extraction with GoogLeNet[13], VGGNet[14,15], ResNet[16], etc. In a previous study[17], it was explored that an ensemble method using the classification probabilities from multiple pre-trained networks achieved better performance than using a single pre-trained network.

Meanwhile, there has been continuous research in the computer vision field to integrate attention mechanisms into architectures inspired by CNN[18]. These attention-based transformer models have shown the ability to learn highly effective feature representations[19]. Recent studies have shown that such transformer modules, operating on a sequence of image patches, can completely replace the standard convolution in deep neural networks by generating Vision Transformers (ViTs)[20,21]. An advantage of ViT is its larger receptive field, which allows it to better understand contextual information than CNNs. This is particularly useful in medical imaging applications, as it considers not only the regions of interest but also the surrounding conditions when diagnosing health statuses. However, ViTs may require more data, which can be a disadvantage in resource-constrained medical imaging fields[22]. To address these drawbacks, some studies have utilized generative adversarial networks (GANs)[23,24]or devised other learning methods[23,24]to train ViT. A hierarchical ViT utilizing Shifted Windows, called Swin Transformer[25], has been employed for thyroid image classification[26], and a CNN-Transformer hybrid model, named MedViT, has also been devised for generalized medical image classification[27].

In medical imaging, labeled data is often scarce, making it essential to optimize the use of available datasets. Data augmentation techniques are widely applied to enhance model robustness and generalization, mitigating the challenges posed by limited labeled data[28,29]. Integrating data augmentation with transfer learning has proven effective for improving performance in medical image analysis[30,31]. However, clinical datasets often suffer from class imbalances, making it difficult to collect well-balanced training samples[28,32]. To address this, various resampling techniques have been employed, such as random under sampling, which reduces the majority class, and random oversampling, which increases the number of minority class instances[33]. Several studies have demonstrated that training on a carefully selected core dataset can maintain, or even improve, model performance compared to training on full-scale datasets[34–37]. By reducing redundancy and focusing on essential data points, data selection enhances generalization and lowers computational costs, making it particularly valuable for medical imaging applications where labelled data is limited. This approach ensures data-efficient deep learning by constructing a core dataset composed of the most essential data points, optimizing training without sacrificing performance.

This paper proposes a new strategy to improve deep learning performance in ultrasound image classification by systematically filtering out possibly misclassified cases. The proposed method does not require an additional dataset and is therefore efficient in clinical applications. To validate its effectiveness, we evaluate the method using ResNet101 and Vision Transformer architectures on ultrasound thyroid nodule images. Furthermore, we assess its generalizability by applying it to another medical dataset, dermoscopic images, and conducting experiments across eleven different pre-trained neural networks.

Materials and methods

Image dataset

The proposed model is evaluated using a dataset of ultrasound images for thyroid nodules. This study exclusively focuses on thyroid nodule classification using ultrasound imaging, ensuring consistency in dataset selection. A total of 18,269 ultrasound thyroid nodule images were collected from Severance Hospital, Yonsei University Health System, between 2016 and 2020. This dataset includes 13,560 images (6400 benign, B, and 7160 malignant, M) from a previous study[17], which were reanalyzed and expanded for this research. The institutional review board (IRB) of Severance Hospital, Seoul, South Korea granted approval for this retrospective study with a waiver of the requirement for informed consent. Each thyroid nodule was diagnosed as either benign or malignant based on aspirate cytology or surgical histology. The thyroid nodule dataset was divided into 17,607 images for training and 662 images for testing. The test set, consisting of 123 benign images and 539 malignant images, was randomly chosen to maintain an unbiased evaluation. The training dataset contained 10,447 benign images and 7,160 malignant images. Model performance was assessed using a fixed train-test split to ensure consistency in the training and validation process. No k-fold cross-validation was applied. To address the class imbalance, we applied a combination of oversampling and data augmentation techniques. Oversampling was performed by random duplication of underrepresented class samples, while data augmentation included random left-right reflections and random rotations within −30 to 30 degrees. These techniques help prevent overfitting and improve model generalization.

Data selection strategy for decision-making in deep learning framework

The proposed method outlines a two-step framework for decision-making. Initially, a decision is made using the original training process. If the classification confidence level, defined by the probability difference between the two output classes, falls below a predefined threshold, a second decision-making step is triggered that employs a network trained with selectively chosen 'good data'. This subsection elaborates on the methodology for identifying 'good data', which forms the cornerstone of our approach.

To identify 'good data', we first train five pre-trained convolutional neural networks, AlexNet[4], VGG16[5], ResNet101[6], GoogLeNet[7], and Inception-v3[8], using the same initial dataset. Each network undergoes training for a maximum of 10 epochs, regardless of whether it reaches 100% accuracy. This limitation prevents overfitting and ensures that False Positive (FP) and False Negative (FN) cases remain present in the dataset. The composition of the dataset, including the number of benign and malignant images used for training, is provided in Table 1. After completing this initial training phase, we reassess each network using the same dataset to evaluate its

Dataset	Network	Training dataset		Architecture
Thyroid nodule	Original network	Benign	10,447	The same architecture (Pre-trained network in MATLAB)
		Malignant	10,447 (7160 before augmentation)	
	True network 3	Benign (True dataset 3)	9862	
		Malignant (True dataset 3)	9862 (6616 before augmentation)	
	True network 5	Benign (True dataset 5)	8790	
		Malignant (True dataset 5)	8790 (5782 before augmentation)	

Table 1. List of experiments, their corresponding dataset and architecture for training networks.

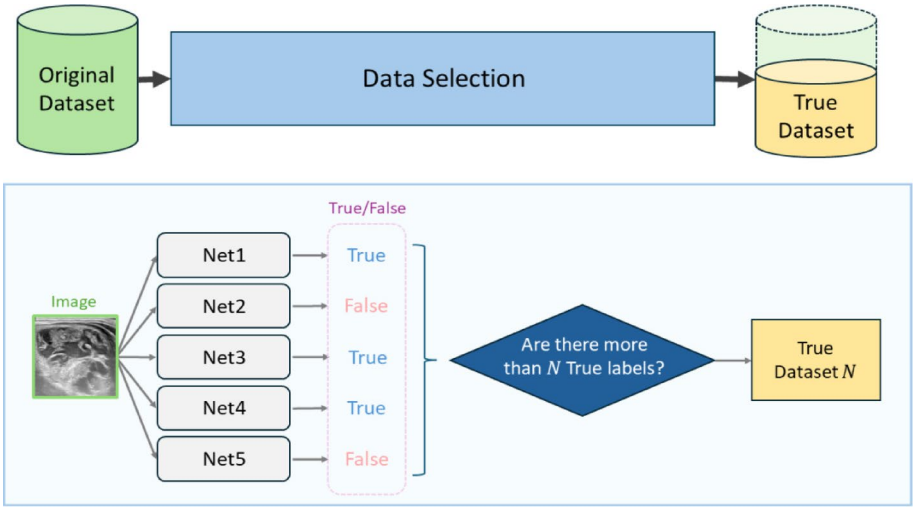

Fig. 1. Data selection process for constructing a True dataset. Five networks classify each image as True or False for both positive and negative cases. Images labeled as 'True' by at least three networks (N = 3) are included in True dataset 3, while images labeled as 'True' by all five networks ($N = 5$) form True dataset 5. This figure illustrates the hierarchical filtering process used to create consensus datasets.

classification performance. Specifically, each image is reclassified by the trained network and assigned to one of four categories: True Positive (TP), True Negative (TN), False Positive (FP), or False Negative (FN). This categorization forms the basis for selecting consensus data for the second decision-making step. To mitigate the risk of overfitting, we limit the maximum number of training epochs to 10, ensuring that sufficient FP and FN samples remain in the dataset.

The dataset comprising only TP or TN labeled images is termed the 'good data', or the 'True dataset'. The data flow diagram for creating a True dataset is illustrated in Fig. 1. We create two variations of the True dataset: True dataset 3 and True dataset 5. True dataset 3 consists of images identified as TP or TN by at least three networks, while True dataset 5 includes images classified as TP or TN by all five networks. Images that satisfy the conditions of both datasets are naturally included in both datasets. Therefore, True dataset 5 is naturally included in True dataset 3. Networks trained on these True datasets are referred to as 'True networks'. These True networks are then used to enhance the original network's classification performance, which was trained with the initial dataset. We conduct a comparative analysis using two True network 3 and True network 5 to evaluate the impact of the number of networks used in data selection.

Proposed method

The proposed method is designed with the original network as the primary classifier, with the True Network acting as a refinement step rather than a replacement. The goal is to enhance the original network's decision-making by introducing a structured verification process that selectively refines only uncertain classifications, rather than reclassifying all data. Instead of relying solely on a single model, this multi-step approach ensures that the original network maintains broad generalization, while the True Network improves precision by focusing on cases flagged as ambiguous. This targeted correction mechanism prevents the True Network, which is trained on a filtered dataset, from becoming over-specialized. Additionally, this selective application reduces computational overhead and make the approach efficient.

The ambiguity in the original network's decision is determined by a specific criterion: if the difference in classification probability between the two classes for a given image is less than a predetermined threshold T^*, the image is flagged for reclassification using the True network. The overall process is depicted in the following steps.

Step 1:	**Initial Classification.** **The image is processed by the original network.** **The probabilities for class B (benign) and M (malignant), denoted as p_O^B and p_O^M, are obtained.**
Step 2:	Uncertainty Check. Computing the difference $p_O^B - p_O^M$. If $\left\|p_O^B - p_O^M\right\| < T^*$, the image is considered ambiguous and is forwarded to the True network. Otherwise, the original classification result is retained.
Step 3:	Re-evaluation with the True network. The True network processes the image, yielding new probabilities p_T^B and p_T^M.
Step 4:	Final Probability Calculation. Calculate the new probability p^B and p^M. Method 1) $p^B = p_T^B$ and $p^M = p_T^M$ Method 2) $p^B = mean\left(p_O^B, p_T^B\right)$ and $p^M = mean\left(p_O^M, p_T^M\right)$
Step 5:	Final Decision. If $p^B > p^M$, the image is classified as benign. Otherwise, it is classified as malignant.

To further elaborate the flowchart in Fig. 2, suppose that the original network classifies an image to the class B (benign), where the probability p_O^B of class B is greater than the probability p_O^M of class M (malignant) from the original network. If the probability p_O^B is not greater than p_O^M by a threshold value T^*, we consider that the original network might have misclassified the image. For instance, $T^* = 0.5$ means $p_O^B - p_O^M = p_O^B - \left(1 - p_O^B\right) < 0.5$, that is $p_O^B < 0.75$. This indicates that when p_O^M is below 0.75, the classification may be unreliable. The image is then passed through the True network for a second evaluation, resulting in the determination of the probabilities p_T^B and p_T^M for the two respective classes (the subindex T refers to the output from the True network). We then set the classification probabilities p^B and p^M using one of two methods. The first method involves assigning the probability from the True network directly, i.e., $p^B = p_T^B$, $p^M = p_T^M$. The second method entails computing the average probability value between the original and True networks, i.e., $p^B = mean\left(p_O^B, p_T^B\right)$, $p^M = mean(p_O^M, p_T^M)$. The final classification decision is then made based on the updated probabilities. If $p^B > p^M$, the image remains classified as benign; otherwise, it is reclassified as malignant. We present two models that follow the flow chart depicted in Fig. 2, and the sole difference between these models is the method used to derive the updated probability.

Evaluation

For performance evaluation of the proposed method, we used three standard metrics: specificity (SPE), sensitivity (SEN), and accuracy (ACC). Additionally, we used F1 score and AUC (Area under the curve). The F1 score is defined as the harmonic mean of precision $(TP/(TP + FP))$ and recall (sensitivity). The F1 score measures the model's accuracy and have the range from 0 to 1; higher F1 scores are generally better. AUC is the area under a ROC (Receiver operating characteristic) curve which represents an (FP, TP)-plot that illustrates the performance of a binary classifier model at varying threshold values. The closer the AUC is to 1, the better the classifier performance. The definitions of the four metrics we used are as follows:

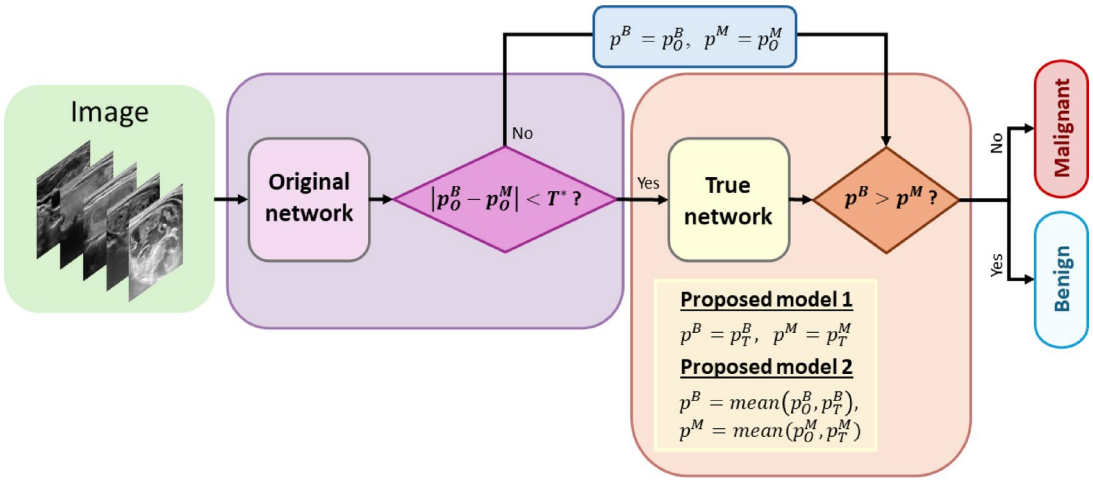

Fig. 2. Flowchart of the proposed model for binary classification into Benign(B) and Malignant(M). The probabilities p_O^B and p_O^M represent the average classification probabilities over the last three epochs for the original network. Similarly, p_T^B and p_T^M are obtained from the True network. The threshold value T^* is chosen within the range $[0,1]$ to determine ambiguity in classification, triggering further evaluation through the True network.

$$\text{Specificity} = \frac{TN}{FP+TN} \times 100\,\%\,,\ \text{Sensitivity} = \frac{TP}{TP+FN} \times 100\,\%,$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\,\%\,,\ \text{F1 score} = \frac{2TP}{2TP+FP+FN} \times 100\,\%.$$

Experimental results

True dataset

Five networks were trained on the original dataset to classify the training data into four categories: TP, TN, FP, and FN. To assess classification consistency across the five models, we analyzed the number of images that received identical labels from all models. We observed that 8,790 benign images and 5,782 malignant images were consistently labeled by all five models forming True dataset 5. Similarly, True dataset 3 was constructed by selecting images that were labeled identically by at least three out of the five models, resulting in 9,862 benign and 6,616 malignant images. The construction of True dataset 3 and True dataset 5 ensures that the selected images have a higher level of classification agreement, thereby improving the reliability of the training process.

Experimental settings

The proposed algorithm was evaluated using ResNet101 and Vision Transformer as the original network, and the outcomes are subsequently discussed. The findings from the use of other pre-trained CNNs can be found in Supplementary Information. The True network has the same architecture as the original network, but it is a model trained with a True dataset, which is a subset of the original dataset. Table 1 shows the training data used for each network to apply the proposed method. Pre-trained networks provided by MATLAB were used, specifically 'resnet101' and 'visionTransformer' (base-16-imagenet-384). The training process was standardized across both models with common settings: the maximum of 10 epochs, the initial learning rate of 0.0001, and batch normalization statistics set to 'moving'. When training ResNet101 in MATLAB, the input image size is 224×224 pixels, with the minibatch size is 100, and the detailed conditions of the Adam optimizer are the gradient decay factor of 0.9, the squared gradient decay factor of 0.999, and the epsilon of 10^{-8}. The training options for the Vision Transformer in MATLAB include the input image size of 384×384 pixels, the mini-batch size of 8 and the SGDM optimizer with the following details: 'LearnRateSchedule' set to piecewise, 'LearnRateDropFactor' set to 0.2, and 'LearnRateDropPeriod' set to 5.

All experiments were conducted using MATLAB R2022a on an NVIDIA GeForce RTX 3090 GPU processor. To ensure the consistency of the model, all networks use the average probability of the networks obtained in the last three epochs out of the maximum epoch. For average probability calculation, we use the 8 th, 9 th, and 10 th probabilities out of 10 epochs for the thyroid nodule dataset. To assess classification certainty, the original network assigns probability values to each class through a SoftMax activation function. This produces two probabilities, p_O^B and p_O^M, corresponding to benign and malignant classifications, respectively. We define the classification confidence as the absolute difference between these two probabilities:

$$\text{confidence} = \left| p_O^B - p_O^M \right|.$$

A higher confidence value indicates a more certain classification, while a lower value suggests ambiguity. If this probability difference falls below a predefined threshold T^*, the image is considered ambiguous and is reclassified using the True network.

Figure 3 shows the histograms and cumulative distributions of class probability differences $|p_O^B - p_O^M|$ in the training data. The plot titled 'Total' is the result of all probability differences across the five networks. The remaining five plots are the results for each representative network. The distributions that the majority of probability differences cluster around 0.9 and above. Based on this analysis, we set the threshold value T^* to 0.9, which led approximately 38% of the training data being reclassified. If the class probability difference of the original network in proposed methods 1 and 2 is less than 0.9, the image is considered ambiguous and reclassified using the True network. The next section will show the results of applying this reclassification strategy to representative classification models such as ResNet101 and Vision Transformer.

Classification results

Table 2 presents the classification results of applying the original network, the True network 3, the average method which classifies using the mean of the probabilities of these two networks, and our proposed models 1&2 in Fig. 2 to ResNet101 and Vision transformer. The classification performance of the original networks significantly differs between ResNet101 and Vision Transformer. ResNet101 exhibits a notably high accuracy of 86%, whereas Vision Transformer exhibits a comparatively lower accuracy of 73%. This discrepancy aligns with existing findings that Vision Transformer often underperform when trained on medium-sized datasets (comprising fewer than 10 million images) without robust regularization techniques in place. However, both models show performance improvements when utilizing True network 3 increases.

In particular, in the case of Vision Transformer, we succeeded in increasing the accuracy by approximately 7% compared to the original network by using only True network 3. ResNet101 showed performance improvement in the proposed models compared to the model proposed using only the True network 3. However, in the case of Vision Transformer, the proposed models did not outperform the True network 3. Nevertheless, both proposed models contribute to performance enhancement over the original network. Furthermore, examining Table S-1 (Supplementary Information), which presents the results from 11 additional CNNs, it can be observed that models utilizing both the original and true networks, such as Average and Proposed models, tend to outperform those using only the original or true networks individually.

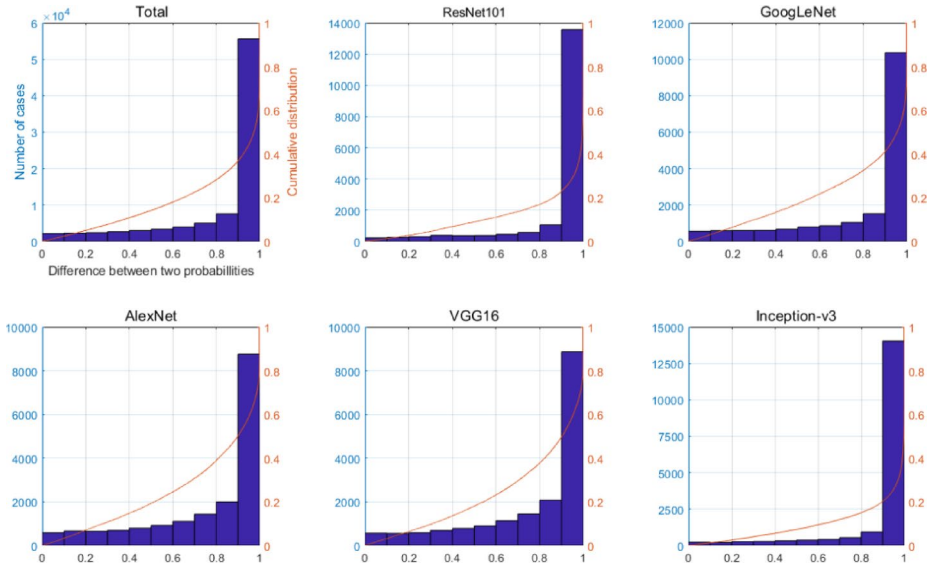

Fig. 3. Plots with overlapping histogram and cumulative distribution of two probability differences $|p_O^B - p_O^M|$ for each five representative networks used to create the True dataset. The plot titled 'Total' is the result of all probability differences across the five networks.

Network	Method	ACC	SPE	SEN	F1	AUC
ResNet101	**Original**	86.10	82.11	87.01	91.07	0.9038
	True network 3	88.52	82.93	89.80	92.72	0.9125
	Average	88.82	86.99	89.24	92.86	0.9205
	Proposed model 1	89.12	86.18	89.80	93.08	0.9170
	Proposed model 2	88.82	86.99	89.24	92.86	0.9221
Vision Transformer	**Original**	73.87	86.99	70.87	81.54	0.8722
	True network 3	85.05	87.80	84.42	90.19	0.9128
	Average	82.48	88.62	81.08	88.28	0.9145
	Proposed model 1	84.89	87.80	84.23	90.08	0.9118
	Proposed model 2	82.48	88.62	81.08	88.28	0.9137

Table 2. Comparison results using the original network, true network 3, average of original and true network and proposed models 1&2 for the test dataset. The architectures are ResNet101 and vision transformer. The threshold value for the proposed models is fixed to 0.9.

As a result, while the Vision Transformer showed significant performance improvement with True network 3, its accuracy remains 4% lower than the best result achieved by ResNet101. Therefore, for small size of dataset of thyroid nodule ultrasound images, ResNet101 appears to be a more effective choice over the Vision Transformer. However, the results presented in this section are based solely on True Network 3. In the next section, we analyse the performance of True Network 5 and compare its effectiveness relative to True Network 3 to better understand the impact of training data size and selection strategy on classification accuracy.

Results using the other true network

Table 3 presents the results of applying True network 5 to our method. Comparing these results with those utilizing True network 3 in Table 2, it models employing True network 3 demonstrate better performance for both ResNet101 and Vision transformer compared to those utilizing True network 5. However, the observed differences depend on both the architecture and dataset characteristics. For ResNet101, the performance difference between the two True Networks is minimal, suggesting that ResNet101's classification ability is less affected by the training subset size. In contrast, Vision transformer, which is more sensitive to the amount of training data, achieves approximately 5% higher accuracy with True network 3 trained on approximately 1,900 more images than True network 5. This result suggests that dataset size plays a more significant role in Vision Transformer's performance than in ResNet101. Therefore, for ultrasound image classification, training True networks using True dataset 3, which includes a larger amount of data than True dataset 5, appears to be more effective in improving performance. The Grad-CAM visualizations for the original network and True network 5 using ResNet101 are presented in Fig. 4. Grad-CAM generates heatmaps highlighting the important regions used by the CNN for classification. Figure 4 depicts instances where the True network effectively classifies images

Network	Method	ACC	SPE	SEN	F1	AUC
ResNet101	Original	86.10	82.11	87.01	91.07	0.9038
	True network 5	86.56	83.74	87.20	91.35	0.9069
	Average	88.97	88.62	89.05	92.93	0.9217
	Proposed model 1	87.61	86.99	87.76	92.02	0.9175
	Proposed model 2	89.12	88.62	89.24	93.04	0.9242
Vision Transformer	Original	73.87	86.99	70.87	81.54	0.8722
	True network 5	79.31	92.68	76.25	85.71	0.8786
	Average	78.70	92.68	75.51	85.24	0.9069
	Proposed model 1	79.15	92.68	76.07	85.59	0.8777
	Proposed model 2	78.55	92.68	75.32	85.12	0.9057

Table 3. Comparison results using the original network, true network 5, average of original and true network and proposed models 1&2 for the test dataset. The architectures are ResNet101 and vision transformer. The threshold value for the proposed models is fixed to 0.9.

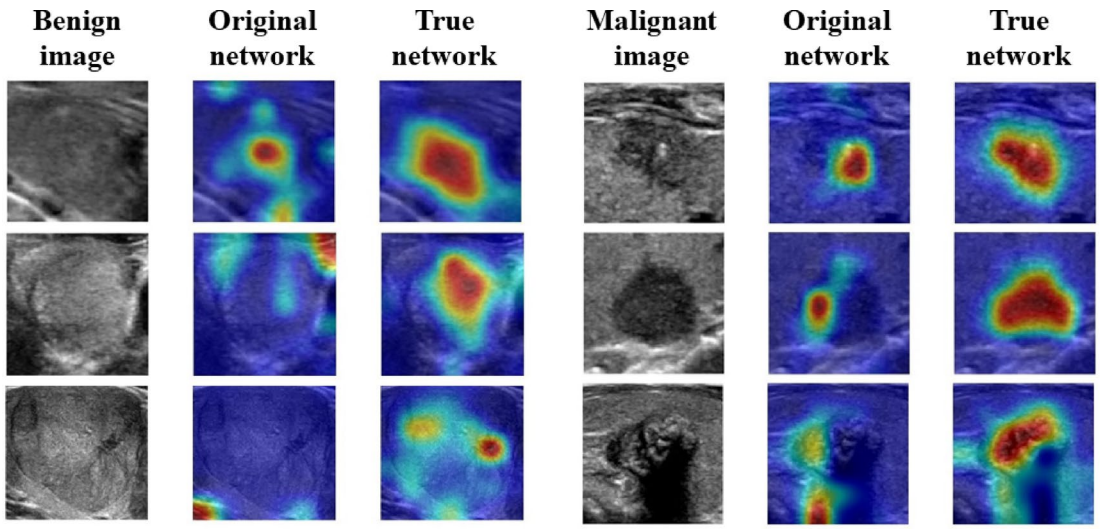

Fig. 4. Grad-CAM visualization for ResNet101 applied to six ultrasound images (three benign and three malignant). The figure is structured in a 3 × 6 format, where the first column contains benign images, the second column shows the corresponding Grad-CAM results from the original network, and the third column presents the Grad-CAM results from the True network. Similarly, the fourth column contains malignant images, followed by Grad-CAM results from the original network in the fifth column and those from the True network in the sixth column. The color scale from red to blue represents the importance of different image regions, where red indicates the most critical areas used for classification. The selected images include cases where the original network misclassified or had a small probability difference, whereas the True network provided correct classification with a higher probability difference.

with significant differences in probability compared to those misclassified or with small probability differences by the original network. While the original network trained on the entire dataset tends to focus on small parts of the nodule or areas without nodules, the True network demonstrates an overall focus on the entire nodule, potentially leading to more reliable classification.

Modifications

This section extends the previous experiments by introducing False networks, which are trained on False datasets comprising misclassified images. With the goal of improving the performance of the original network, we employed the True network, which was trained using accurately classified images. Similarly, we aimed to apply this technique to the proposed model by generating, so-called, a false dataset through the collection of poorly classified images.

Initially, we created a false dataset by gathering images that the five pre-trained networks were unable to classify effectively. Five false datasets were constructed based on the number of pre-trained networks that classified images with the same label (FP or FN). False dataset 1 consisted of images classified as FP or FN by any of the five networks. False datasets 2–5 were similarly created. The number of ultrasound images in False datasets 1–5 were (1657,1378), (953,823), (585,544), (283,331), and (74,99) for FP or FN, respectively. In the experimental

setting, False dataset 3 is chosen for use because there is a scarcity of images that receive unanimous classification as false from all five networks. Next, we trained a network with the same architecture and training options as the original network for the False dataset 3. A network trained on the false dataset 3 is called a False network 3.

The proposed algorithm depicted in Fig. 2 utilizes the True network classification result if the difference in probabilities between classes of the original network is less than 0.9 ($\left|p_O^B - p_O^M\right| < 0.9$). To further enhance classification performance, we introduced an additional condition: if the final probability difference between p^B and p^M of the proposed method was less than 0.1, the image was classified once more using False network 3. The threshold of 0.1 was selected based on experimental evaluations ranging from 0.1 to 0.9, with 0.1 yielding the best results.

Experiments were implemented using ResNe101 architecture and on the False dataset 3, and the results were illustrated in Table 4. After analyzing the results, we can deduce that incorporating false data into the learning process in order to prevent networks from making incorrect decisions, by teaching them something that may not be accurate due to the inclusion of unreliable data, does not effectively enhance the performance of the networks. This suggests that False networks may not provide the same level of beneficial correction as True networks and that training on correctly classified images remains the more effective strategy for improving classification accuracy.

Result for skin cancer image dataset

To assess the applicability of the proposed technique, in this section, we examine the outcomes obtained by applying our method to classify images of melanoma, a form of skin cancer, rather than ultrasound images. Melanoma is an aggressive form of skin cancer that originates in the melanocytes and can rapidly metastasize, making early detection and accurate diagnosis critical[38]. Since the development of deep learning, various classification methods using CNN have been developed for melanoma lesion images. The International Skin Imaging Collaboration (ISIC) Archive provides skin lesion dermoscopic images for research[39]. ISIC provides high-resolution dermoscopic images, which minimize artifacts such as shadows and light reflections, allowing for more reliable image-based classification[40,41]. For our experiment, we used a subset of the ISIC Challenge dataset[42,43], specifically selecting images labeled as benign nevi and malignant melanoma. There are 12,875 benign and 4,522 malignant images, so 10% of them are separated as test data (B:1,287, M:452). The training dataset is created by randomly selecting 4,070 images of each class from each class to ensure balanced distribution. When learning the original network for melanoma data, the data augmentation method was not used because the number of images for each class in the training data was the same.

Since the training data is very small at about 8,000 pieces, this data set was chosen as True dataset 3, which has more data than True dataset 5. True dataset 3 has 3,914 benign and 3,684 malignant images. True dataset 5 has 3,264 benign and 2,801 malignant images. Augmentation was performed to make fewer malicious images equal to benign images. The training options are the same as the training options for thyroid nodule images except that the max epoch is set to 20. Because dermoscopic images are color images, unlike grayscale ultrasound images, extending the number of epochs was necessary to allow the network to better capture relevant features. Consequently, the probability values of each network used in the experiment was the average of the probabilities of the last three networks among the entire epoch, so the 18, 19, and 20 networks were used.

Table 5 shows the results of applying the proposed method to ResNet101 and Vision Transformer for melanoma classification. The proposed models obtained by fine-tuning the ResNet101 architecture leveraging True network 3 showed improved accuracy compared to the original network. Conversely, Vision Transformer performed better with True network 5, though its sensitivity for detecting malignant cases remained low, limiting its practical applicability. To further validate the effectiveness of our proposed methods, we extend our evaluation to 11 additional pre-trained architectures beyond ResNet101 and Vision Transformer. The results are presented in Supplementary Table S-2. When using True network 3, proposed model 1 achieves higher accuracy than the original network in 10 out of 11 cases (90.9%) and outperforms True network 3 alone in 10 cases. Similarly, proposed model 2 surpasses the original network in 10 out of 11 cases, though it shows improvements over the average of the original and True network in only 3 cases. Furthermore, in 7 out of 11 architectures, proposed model 2 performs better than proposed model 1 which suggests that the averaging-based approach can be beneficial in some scenarios. When using True network 5, a similar trend was observed. The proposed model 1 shows superior accuracy over the original network in 9 out of 11 cases (81.8%) and surpasses True network 5 alone in all 11 cases. Likewise, proposed model 2 outperforms the original network in 11 cases and exceeds the average of the original and True network in 11 cases, showing a more consistent improvement. Additionally,

ResNet101					
Method	ACC	SPE	SEN	F1	AUC
False network 3	29.00	21.95	30.61	41.25	0.1758
Proposed model 1 + True 3 + False 3	88.82	85.37	89.61	92.88	0.9130
Proposed model 2 + True 3 + False 3	88.67	84.55	89.61	92.80	0.9089
Proposed model 1 + True 5 + False 3	87.61	86.99	87.76	92.02	0.9090
Proposed model 2 + True 5 + False 3	89.12	88.62	89.24	93.04	0.9194

Table 4. Classification results for test data when the false network 3 was used in proposed methods 1&2 for each true network.

ResNet101					
Networks	ACC	SPE	SEN	F1	AUC
Original	84.76	83.06	89.60	75.35	0.9170
True network 3	86.03	87.41	82.08	75.33	0.8924
Average (Original, True3)	85.91	86.25	84.96	75.81	0.9266
Proposed model 1	86.26	86.64	85.18	76.31	0.8929
Proposed model 2	86.08	86.25	85.62	76.18	0.9184
Vision transformer					
Networks	ACC	SPE	SEN	F1	AUC
Original	84.93	95.23	56.40	66.50	0.8684
True network 3	83.09	98.44	40.56	55.99	0.7856
Average (Original, True3)	83.73	97.65	45.12	59.51	0.8579
Proposed model 1	83.38	98.44	41.65	57.06	0.7888
Proposed model 2	83.84	97.65	45.55	59.91	0.8583
True network 5	84.93	91.78	65.94	69.89	0.8170
Average (Original, True5)	85.91	93.35	65.29	71.07	0.8650
Proposed model 1	85.22	92.25	65.73	70.22	0.8171
Proposed model 2	85.80	93.35	64.86	70.77	0.8643

Table 5. Comparison results for melanoma dermoscopic image dataset. The experiments are obtained using the original network, true network, average of original and true network and proposed models 1&2 for the test dataset. The threshold value is fixed to 0.9.

proposed model 2 performs better than proposed model 1 in 9 out of 11 architectures and reinforces the effectiveness of probability averaging when True network 5 is used.

These findings suggest that while the True network alone does not always surpass the original network, integrating it into the proposed decision-making framework leads to consistent performance gains. Furthermore, the comparative results between proposed models 1 and 2 indicate that different network architectures may benefit more from either direct correction (proposed model 1) or probability averaging (proposed model 2), depending on the dataset and model characteristics.

Discussion

In this work, we propose a methodology to improve the classification results of a network for ultrasound images. We generate a dataset excluding images that confuse learning and use it to train a new network, True network. The True network is used to correct the original network's classification result. In the proposed approach, the original network serves as the primary classifier, while the True Network functions as a secondary refinement mechanism, rather than a full replacement. Instead of applying the True Network to the entire dataset, it is selectively used for instances where the original network shows uncertainty. This selective refinement process helps maintain a balance between generalization and precision which allows the original network to draw upon its broad knowledge while the True Network focuses on improving classification in ambiguous cases. By integrating this verification step, the method enhances overall accuracy while keeping computational demands manageable, as demonstrated by our experimental results. Table 2 illustrates that even when using only the True network, it outperforms the original network. In other words, it means that networks trained on selected data exhibit better performance than the original network. Furthermore, upon examining Table 3, it appears that selecting as much good data as possible within limited datasets can contribute to performance improvement.

Various experimental results in Tables 2, 5, and S-1&2 illustrate that improved results can be obtained using the original and True networks together for CNNs. CNNs and Vision Transformer exhibited different trends in the experiments. For ultrasound images, it can be observed that Vision Transformer, contrary to most CNN models, shows superior performance with True network trained only on selected data, compared to both the original network and proposed models. However, for melanoma images, we found that neither Vision Transformer nor CNN achieved significant performance improvement. Therefore, the proposed method shows variability depending on which pre-trained architecture and true network are selected for a specific medical image.

While our method provides consistent performance improvements for ultrasound image classification, we also observe certain limitations. Our method primarily focuses on refining decision-making using strategically selected data based on TN and TP images, the FN and FP images can also contain valuable information. To explore this, we attempted to apply the same True network-based strategy to FN and FP datasets, expecting that training on these challenging cases might help correct them. However, our experiments revealed that the proposed approach was less effective in capturing FN and FP characteristics. This limitation arose due to the small number of images that all five networks misclassified in common, making it difficult to extract meaningful features for proper reclassification.

As an extension of our method, we introduced False networks trained on misclassified images (False datasets) to assess whether training on difficult cases could enhance decision-making. Specifically, if an image remained

ambiguous even after classification by the True network (i.e., the final probability difference was below 0.1), it was re-evaluated using False network 3. However, as shown in Table 4, adding the False network resulted in only a marginal improvement over the True network alone. This suggests that while misclassified images contain useful information, they do not contribute as effectively to classification refinement as correctly classified samples. These findings reinforce the idea that consensus-based selection provides a more stable foundation for training robust networks. Future research could explore hybrid methods that integrate selected false samples in a controlled manner to balance robustness and diversity.

Our method employs multiple pre-trained networks in the data selection process, but it differs from traditional ensemble learning. In ensemble learning methods, such as bagging and boosting, multiple models are combined at inference to improve classification performance. The proposed approach does not rely on multiple models at inference but instead uses them only for refining the training dataset. The selected networks evaluate the dataset to identify reliably classified samples, which are then used to train a single final True Network. This consensus-based data selection improves training data quality without requiring multiple networks to contribute directly to inference. By filtering out potentially misclassified samples before training, our method enhances classification accuracy efficiently while avoiding the computational overhead of ensemble learning.

One potential concern in data selection is whether the consensus-based approach may inherently select easier-to-classify samples while excluding harder cases, leading to a distribution shift between the original dataset and the selected subset. While our method does not explicitly enforce class balance, the reliance on multiple independent networks mitigates individual classifier bias, reducing the likelihood of systematically skewing the dataset. Our results suggest that while some difficult samples were excluded, this did not negatively impact overall performance.

Furthermore, in Step 2 of our proposed method, the selection of an appropriate confidence threshold (T^*) is currently empirical. While our experiments show that this threshold effectively refines classification decisions, its lack of systematic optimization is a limitation. Developing an adaptive, data-driven approach, such as Bayesian optimization or reinforcement learning, for setting T^* could enhance the generalizability and robustness of our method.

Using the average of the probabilities of the original and true networks usually improves performance compared to using them separately. Since the averaging method utilizes the results of the two networks without conditions on each network, a more significant effect can be obtained if the two networks' classification results are similar. However, simple averaging can yield suboptimal results if the two networks exhibit opposite classification tendencies. Our proposed method can compensate for these shortcomings. The proposed methods apply the True network condition only to images classified with low confidence by the original network, it can play a stable filtering role even for two networks with opposite tendencies. Another advantage of our approach is that it achieves accuracy comparable to the averaging method while being computationally more efficient. Unlike ensemble-based methods that classify all images using both networks, our approach applies the True network only to a subset of the testing dataset that can significantly reduce computational cost. The more images classified without requiring re-evaluation, the greater the efficiency gains.

In summary, while our method offers a practical, efficient, and effective strategy for improving ultrasound and dermoscopic image classification, future work should explore ways to optimize threshold selection and leverage FN and FP datasets more effectively to further enhance performance.

Conclusion

This study presents a novel two-step for enhancing the classification performance of deep learning models that include CNNs and Vision Transformer in ultrasound image analysis The proposed approach begins with an initial classification using the original network. If the classification confidence of this decision falls below a predefined threshold, a second decision-making step is triggered which utilizes a network trained on a strategically selected dataset, referred to as the True network. This True network maintains the same architecture as the original network but optimally leverages existing data without requiring additional data collection. Our findings suggest that this method provides an effective means for improving classification performance, particularly in cases where data availability is limited and computational resources are constrained. Although this study focuses on ultrasound and dermoscopic image classification, we expect that our proposed method could also be applicable to other medical imaging domains and modalities.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to reasons of sensitivity but are available from the corresponding author on reasonable request.

Received: 6 August 2024; Accepted: 28 April 2025
Published online: 08 May 2025

References

1. Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. **9**, 611–629. https://doi.org/10.1007/s13244-018-0639-9 (2018).
2. Yu, X. et al. Transfer learning for medical images analyses: A survey. *Neurocomputing* **489**, 230–254. https://doi.org/10.1016/j.neucom.2021.08.159 (2022).
3. Morid, M. A., Borjali, A. & del Fiol, G. A scoping review of transfer learning research on medical image analysis using imagenet. *Comput. Biol. Med.* **128**, 104115. https://doi.org/10.1016/j.compbiomed.2020.104115 (2021).
4. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems* Vol. 25 (eds Pereira, F. et al.) 1–9 (Curran Associates, Inc., 2012).

5. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition 2014. https://doi.org/10.48550/arxiv.1409.1556
6. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016).
7. Szegedy, C. et al. Going Deeper With Convolutions. Proceedings of the IEEE Conference on Computer and Pattern Recognition (CVPR), (2015).
8. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016).
9. Alexander, E. K. & Cibas, E. S. Diagnosis of thyroid nodules. *Lancet Diabetes Endocrinol.* **10**, 533–539. https://doi.org/10.1016/S2213-8587(22)00101-2 (2022).
10. Megwalu, U. C. & Moon, P. K. Thyroid Cancer incidence and mortality trends in the united States: 2000–2018. *Thyroid* **32**, 560–570. https://doi.org/10.1089/thy.2021.0662 (2022).
11. Anari, S., Tataei Sarshar, N., Mahjoori, N., Dorosti, S. & Rezaie, A. Review of deep learning approaches for thyroid Cancer diagnosis. *Math. Probl. Eng.* **2022**, 5052435. https://doi.org/10.1155/2022/5052435 (2022).
12. Sharifi, Y. et al. Deep learning on ultrasound images of thyroid nodules. *Biocybern Biomed. Eng.* **41**, 636–655. https://doi.org/10.1016/j.bbe.2021.02.008 (2021).
13. Chi, J. et al. Thyroid nodule classification in ultrasound images by Fine-Tuning deep convolutional neural network. *J. Digit. Imaging.* **30**, 477–486. https://doi.org/10.1007/s10278-017-9997-y (2017).
14. Qin, P., Wu, K., Hu, Y., Zeng, J. & Chai, X. Diagnosis of benign and malignant thyroid nodules using combined conventional ultrasound and ultrasound elasticity imaging. *IEEE J. Biomed. Health Inf.* **24**, 1028–1036. https://doi.org/10.1109/JBHI.2019.2950994 (2020).
15. Zhang, X., Lee, V. C. S., Rong, J., Lee, J. C. & Liu, F. Deep convolutional neural networks in thyroid disease detection: A multi-classification comparison by ultrasonography and computed tomography. *Comput. Methods Programs Biomed.* **220**, 106823. https://doi.org/10.1016/j.cmpb.2022.106823 (2022).
16. Zhou, H., Wang, K. & Tian, J. Online transfer learning for differential diagnosis of benign and malignant thyroid nodules with ultrasound images. *IEEE Trans. Biomed. Eng.* **67**, 2773–2780. https://doi.org/10.1109/TBME.2020.2971065 (2020).
17. Koh, J. et al. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. *Sci. Rep.* **10**, 15245. https://doi.org/10.1038/s41598-020-72270-6 (2020).
18. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1–11 (2017).
19. Chaudhari, S., Mithal, V., Polatkan, G. & Ramanath, R. An attentive survey of attention models. *ACM Trans. Intell. Syst. Technol. (TIST).* **12**, 1–32 (2021).
20. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Preprint ArXiv* :201011929 (2020).
21. Shamshad, F. et al. Transformers in medical imaging: A survey. *Med. Image Anal.* **8**, 102802 (2023).
22. Li, J. et al. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Med. Image Anal.* **85**, 102762 (2023).
23. Sun, J. et al. Classification for thyroid nodule using ViT with contrastive learning in ultrasound images. *Comput. Biol. Med.* **152**, 106444 (2023).
24. Jerbi, F., Aboudi, N. & Khlifa, N. Automatic classification of ultrasound thyroids images using vision Transformers and generative adversarial networks. *Sci. Afr.* **20**, e01679 (2023).
25. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–22. (2021).
26. Huang, L., Xu, Y., Wang, S., Sang, L. & Ma, H. SRT: Swin-residual transformer for benign and malignant nodules classification in thyroid ultrasound images. *Med. Eng. Phys.* **124**, 104101 (2024).
27. Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B. & Ayatollahi, A. MedViT: a robust vision transformer for generalized medical image classification. *Comput. Biol. Med.* **157**, 106791 (2023).
28. Raj, R., Mathew, J., Kannath, S. K. & Rajan, J. Crossover based technique for data augmentation. *Comput. Methods Programs Biomed.* **218**, 106716. https://doi.org/10.1016/j.cmpb.2022.106716 (2022).
29. Kora, P. et al. Transfer learning techniques for medical image analysis: A review. *Biocybern Biomed. Eng.* **42**, 79–107. https://doi.org/10.1016/j.bbe.2021.11.004 (2022).
30. Wang, L., Zhang, L., Zhu, M., Qi, X. & Yi, Z. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. *Med. Image Anal.* **61**, 101665. https://doi.org/10.1016/j.media.2020.101665 (2020).
31. Kassem, M. A., Hosny, K. M. & Fouad, M. M. Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. *IEEE Access.* **8**, 114822–114832. https://doi.org/10.1109/ACCESS.2020.3003890 (2020).
32. Gao, L., Zhang, L., Liu, C. & Wu, S. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artif. Intell. Med.* **108**, 101935. https://doi.org/10.1016/j.artmed.2020.101935 (2020).
33. Abdani, S. R., Zulkifley, M. A. & Zulkifley, N. H. Undersampling and Oversampling Strategies for Convolutional Neural Networks Classifier. In: Md. Zain Z, Sulaiman MohdH, Mohamed AI, Bakar MohdS, Ramli MohdS, editors. Proceedings of the 6th International Conference on Electrical, Control and Computer Engineering, Singapore: Springer Singapore; pp. 1129–37. (2022).
34. Xia, X. et al. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. The Eleventh International Conference on Learning Representations, (2022).
35. Huang, J., Huang, R., Liu, W., Freris, N. & Ding, H. A novel sequential coreset method for gradient descent algorithms. International Conference on Machine Learning, pp. 4412–22. (2021).
36. Lei, S., He, F., Yuan, Y. & Tao, D. Understanding deep learning via decision boundary. *IEEE Trans. Neural Netw. Learn. Syst.* **36**(1), 1533–1544 (2025).
37. Lyu, Y. & Tsang, I. W. Curriculum Loss: Robust Learning and Generalization against Label Corruption. International Conference on Learning Representations, (2019).
38. Apalla, Z., Nashan, D., Weller, R. B., Castellsagué, X. & Skin Cancer Epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. *Dermatol. Ther. (Heidelb).* **7**, 5–19. https://doi.org/10.1007/s13555-016-0165-y (2017).
39. ISIC Archiv site. Available: https://www.isic-archive.comn.d
40. Al-masni, M. A., Kim, D-H. & Kim, T-S. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Comput. Methods Programs Biomed.* **190**, 105351. https://doi.org/10.1016/j.cmpb.2020.105351 (2020).
41. Codella, N. et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC) (2019).
42. Codella, N. C. F. et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–72. (2018). https://doi.org/10.1109/ISBI.2018.8363547
43. Combalia, M. et al. BCN20000: Dermoscopic Lesions in the Wild (2019).

Acknowledgements
This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government NRF-2021R1 A2 C2007492 and RS-2023-00282764. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions
All authors contributed to the study conception and design. Material preparation and data analysis were performed by Hyunju Lee and Eunjung Lee. Data collection was performed by Jin Young Kwak. The first draft of the manuscript was written by Hyunju Lee and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

Ethics approval
The institutional review board (IRB) of Severance Hospital, Seoul, South Korea granted approvalfor this retrospective study with a waiver of the requirement for informed consent. All methods wereperformed in accordance with the relevant guidelines and regulations.

Competing interests
The authors declare no competing interests.

Disclosure of conflicts of interest
All of the authors declare that they have no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00416-5.

Correspondence and requests for materials should be addressed to E.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.