Open Journal of Medical Imaging, 2023, 13, 63-83
https://www.scirp.org/journal/ojmi
ISSN Online: 2164-2796
ISSN Print: 2164-2788

# A Breast Density Classification System for Mammography Considering Reliability Issues in Deep Learning

**Eri Matsuyama[1*], Megumi Takehara[2], Noriyuki Takahashi[3], Haruyuki Watanabe[4]**

[1]Faculty of Informatics, University of Fukuchiyama, Kyoto, Japan
[2]Breast Center Dokkyo Medical University Hospital, Tochigi, Japan
[3]School of Hearth Sciences, Fukushima Medical University, Fukushima, Japan
[4]School of Radiological Technology, Gunma Prefectural College of Health Sciences, Gunma, Japan
Email: *matsuyama-eri@fukuchiyama.ac.jp

**How to cite this paper:** Matsuyama, E., Takehara, M., Takahashi, N. and Watanabe, H. (2023) A Breast Density Classification System for Mammography Considering Reliability Issues in Deep Learning. *Open Journal of Medical Imaging*, **13**, 63-83.
https://doi.org/10.4236/ojmi.2023.133007

**Received:** June 5, 2023
**Accepted:** August 4, 2023
**Published:** August 7, 2023

## Abstract

In a convolutional neural network (CNN) classification model for diagnosing medical images, transparency and interpretability of the model's behavior are crucial in addition to high classification accuracy, and it is highly important to explicitly demonstrate them. In this study, we constructed an interpretable CNN-based model for breast density classification using spectral information from mammograms. We evaluated whether the model's prediction scores provided reliable probability values using a reliability diagram and visualized the basis for the final prediction. In constructing the classification model, we modified ResNet50 and introduced algorithms for extracting and inputting image spectra, visualizing network behavior, and quantifying prediction ambiguity. From the experimental results, our proposed model demonstrated not only high classification accuracy but also higher reliability and interpretability compared to the conventional CNN models that use pixel information from images. Furthermore, our proposed model can detect misclassified data and indicate explicit basis for prediction. The results demonstrated the effectiveness and usefulness of our proposed model from the perspective of credibility and transparency.

## Keywords

Explainable AI, t-SNE, Entropy, Wavelet Transform, Mammogram

## 1. Introduction

Breast cancer is a major public health problem for women worldwide and is currently the most common cancer globally. It is estimated that in 2022, approximately 287,850 new cases of invasive breast cancer were diagnosed among US women, and about 43,250 women died from the disease [1]. Early detection, prompt diagnosis, and effective treatment of breast cancer are expected to improve survival rates, reduce morbidity, and lower cost of care [2]. Mammography is the most widely used screening modality for the early detection of breast cancer. It has been shown to reduce breast cancer mortality by 38% - 48% among participants who were actually screened [3]. Breast density is an important factor in determining a woman's risk of breast cancer [4] [5] [6] [7]. It describes the relative amount of different types of breast tissues, *i.e.*, glandular tissue, fibrous connective tissue, and fatty breast tissue, as seen on a mammogram. Breast density is cited as one of the important indicators for predicting breast cancer risk. Dense breast tissue has relatively high amounts of glandular tissue and fibrous connective tissue and relatively low amounts of fatty breast tissue [8]. Dense breast is a prevalent and strong correlation factor for breast cancer. Therefore, classification of mammographic breast density is a very important task. The most commonly used tool for classifying breast density clinically is the Breast Imaging Reporting and Data Systems 5th edition (BI-RADS). BI-RADS classifies breast density into four categories, namely, (BD1; breast density 1) extremely fatty, (BD2: breast density 2) scattered density, (BD3: breast density 3) heterogeneously dense, and (BD4: breast density 4) extremely dense [9] [10]. Of these 4 categories, the assessment of BD1 and BD4 is highly consistent. However, there is greater variability distinguishing scattered density from heterogeneously dense parenchyma [11] [12] [13]. The reading process by radiologists is monotonous, tiring, lengthy, and costly. Moreover, there is large inter- and intra-radiologist variability in subjective density categorization [14]. To address these issues, the development of computer-aided methods for breast density classification is currently being actively pursued.

Recent advances in machine learning have provided opportunities to use deep learning (DL) techniques to address the challenge of breast cancer detection. Deep convolutional neural network (DCNN) is one of the most popular algorithms for DL and has been successfully applied in various fields, achieving high performance in image recognition and classification [15] [16]. Many studies have attempted to apply DCNNs to assess and classify mammographic breast density [17] [18] [19] [20]. However, there are important flaws in current DCNN-based models. These DCNN-based models are black box models that generalize the data transmitted to it and learn from the data. Thus, the relational link between input and output is unobservable [21]. For these reasons, artificial intelligence (AI)-based methods have yet to be widely used in medical practice. A medical diagnosis system needs to be transparent, understandable, and explainable to earn the trust of medical professionals and patients [22]. Therefore, the explainability and interpretability of black box models need to be seriously addressed. In recent years, explainable artificial intelligence (XAI) has been a hot research topics. Medical imaging researchers are increasingly using XAI to explain the results of algorithms, and XAI techniques have been developed using a variety of approaches [23]-[29].

In a previous study, we proposed a method for automatically classifying breast density by employing a wavelet transform-based and fine-tuned DCNN method [30]. The experimental results showed that the proposed method achieves promising classification performance in distinguishing between scattered density and heterogeneous density. However, a flaw of this paper is that the proposed DCNN-based model operated essentially as a black box without interpretability. Also, analytical reliability evaluations were not performed. A reliable classification model should provide a measure of uncertainty associated with the prediction so that physicians can make a well-informed decision. To overcome the shortcoming of the previous study, in this study, we newly propose an interpretable DCNN-based model for the classification of breast density in mammographic images. DCNN-based models typically learn raw image pixels represented in the spatial domain and perform image classification tasks by directly providing pixels as classification inputs. However, in this case, the spectral information content of images is not utilized for classification. In the present work, as inputs to the DCNN models, we adopt using the wavelet coefficients of original images instead of using pixel value information from original images. Our study focuses on distinguishing between the two most difficult categories of the BI-RADS density classification, namely, scattered density (BD2) vs. heterogeneous density (BD3) [31]. The classification of scattered density and heterogeneous high-density is extremely important. Differentiating between these two types of breast tissue helps radiologists identify potential abnormalities. Radiologists can make informed decisions and take appropriate actions based on the characteristics of the observed phenomena. Our goals are to assess the breast density classification performance of a newly proposed wavelet-based DCNN model, and to perform an analytical reliability evaluation of the model as well as visualize and characterize the model's behavior.

The main contributions of this study are as follows: 1) Evaluate the reliability of the proposed model using reliability diagrams, a measure of whether the output probability of the model is consistent with the true probability or not [32]; 2) Use t-distributed stochastic neighbor embedding (t-SNE) [33], a nonlinear, iterative dimensionality reduction method, to gain a better understanding of the behavior and response of the proposed model; 3) Analyze the ambiguity (uncertainty) of the proposed model's judgments from the perspective of Shannon's information entropy; 4) Use gradient-weighted class activation mapping (Grad-CAM) [34], a class discriminable localization method, to visually understand where the proposed model is looking and where mammograms are being evaluated.

## 2. Methods

In this study, we evaluate the reliability of the proposed model, investigate its behavior and judgment ambiguity, and visualize the areas that influence its judgment. The proposed model is based on ResNet-50 [35], which has been demonstrated to be effective in many medical imaging tasks. ResNet-50 is a pre-trained model that has been trained on over one million images from the ImageNet database [36] and is designed for large-scale natural images. As natural images have inherent differences from mammograms, we conduct retraining (fine-tuning) throughout all layers of the architecture.

Image data set used is comprised of 1300 mammograms (MMG). A spectral information learning model for the images (hereinafter referred to as "wavelet-model") is constructed. For comparison, we also construct a conventional original image (pixel information)-based learning model (hereafter referred to as "original-model"). Details of the spectral information of the images will be described in Section 2.2. Both models perform fine-tuning and 10-fold cross- validation to distinguish between scattered density (BD2) and heterogeneous high-density (BD3). In cross-validation, the original images of each 10-fold dataset (10 subsets) are unified for both models for performing training and validation. In the following, we describe the dataset used, the method of extracting the image spectral information used, the architecture of the proposed method, the analytical reliability evaluation method, the behavior visualization method, the ambiguity measurement, and visualization of judgement regions.

## 2.1. Image Data Set

The image dataset used was mammogram X-ray DICOM images acquired from The Cancer Imaging Archive (TCIA) [37]. TCIA is a large archive of publicly available medical images (cancer images) on the web. Therefore, there are no ethical issues in this study, and the requirement to obtain informed consent was waived.

This dataset includes images with and without calcification/masses, with a definitive diagnosis of benign or malignant. In this study, 650 BD2 and BD3 images each (1300 images in total, maximum 2 images from the same patient) were collected. The collected images were manually segmented by a certified breast specialist to remove background areas not needed for diagnosis and labeled as BD2 and BD3. The sizes of these segmented images varied. **Figure 1** shows an example of the segmented images.

## 2.2. Extraction of Image Spectral Information Extraction Using Redundant Discrete Wavelet Transform

In this study, two-dimensional redundant discrete wavelet transform (2D-RDWT) was used to extract spectral information from original images. In the medical field, two-dimensional discrete wavelet transform (2D-DWT) has been applied for data compression, image enhancement, and noise reduction [38]. In 2D-DWT, an image is initialized at decomposition level 0 and decomposed into four components: one low-frequency component and three high-frequency components at

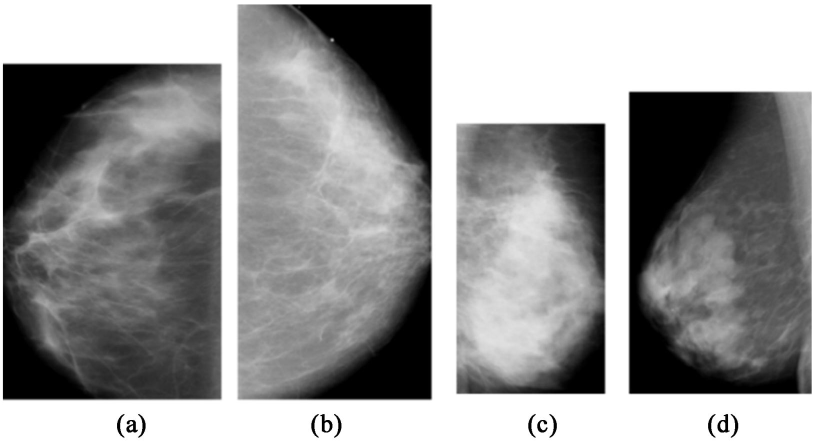


Figure 1. Example of the segmented images. (a) BD2 (benign); image size 3468 × 2370 pixels; (b) BD2 (malignant); image size 4698 × 2352 pixels; (c) BD3 (malignant); image size 2238 × 1260 pixels; (d) BD3 (benign); image size 3245 × 1968 pixels.

decomposition level 1. The low-frequency component (LL) produces a smoothed image, while the three high-frequency components, namely the low-high (LH), high-low (HL), and high-high (HH) components, produce three detailed images. When further decomposition is repeated, it is performed on the LL component. As the decomposition level increases, the resolution of the image decreases. More details of the 2D-DWT can be found in the literature [39].

The conventional 2D-DWT involves downsampling, decomposing the image into four components, and reducing each component to 1/4 size. The result of such 2D-DWT lacks shift-invariance and may cause problems such as loss of image contours. To avoid this problem, we used 2D-RDWT, which does not involve downsampling. The basic algorithm of 2D-RDWT applies the transformation at each point in the image, stores the detailed coefficients, and uses the approximation coefficients for the next level. Therefore, the size of the coefficient array does not decrease for each level [38] [39]. As a result, shift invariance is maintained and the size of each of the four decomposed components remains the same as the original image. There are various wavelet basis functions in discrete wavelet transform. For example, there are Haar, Daubechies, biorthogonal spline, Coiflet, Meyer, etc. In this study, Daubechies order 2 (db2) was used.

Figure 2 provides an overview of level 1 2D-RDWT and dataset creation. Figure 2(a) displays the original image and the four component images of the level 1 2D-RDWT. Figure 2(b) is the 3-channel data for the wavelet-model, consisting of a combination of LL, LH, and HL images. Figure 2(c) is the 3-channel data for the original-model.

## 2.3. Proposed Architecture Based on Fine-Tuned ResNet50 Using Wavelet Coefficients

We modified ResNet50 by introducing a wavelet coefficient 3-channel algorithm in the input layer. In addition, we equipped an algorithm for dimensionality reduction and 2D mapping of high-dimensional data immediately after the activation

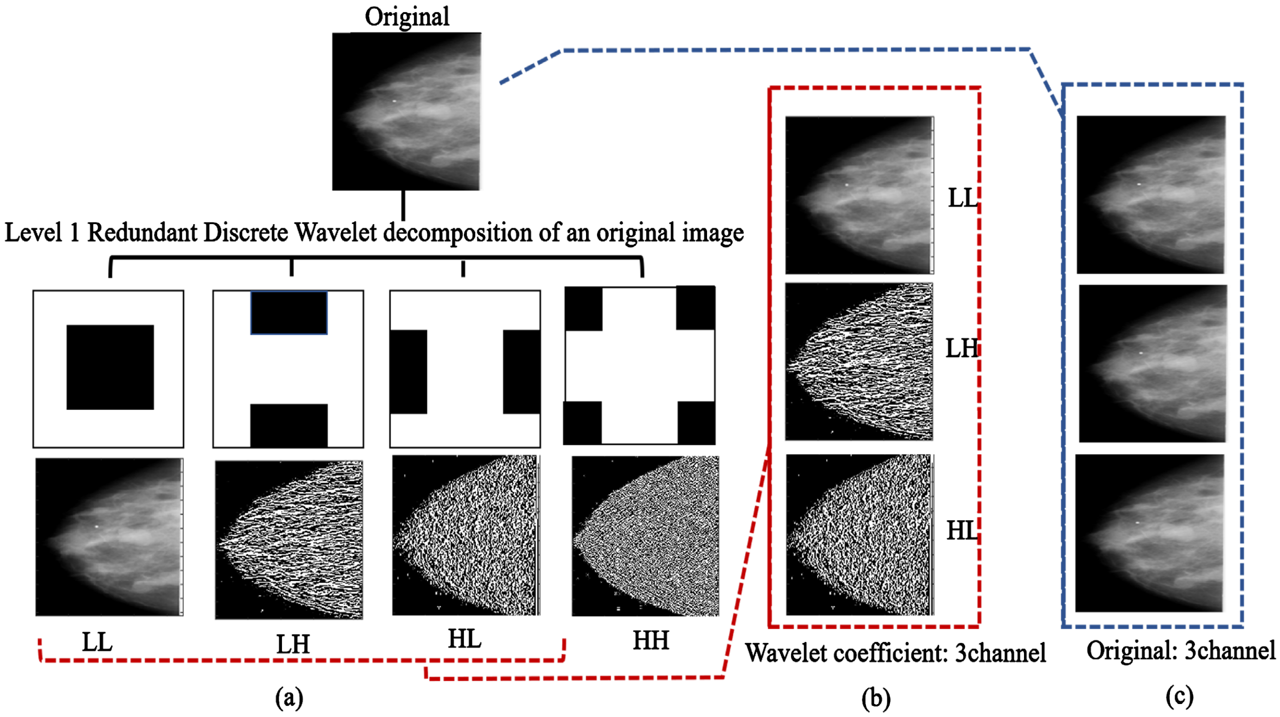

Original
Level 1 Redundant Discrete Wavelet decomposition of an original image
LL
LH
HL
HH
(a)
LL
LH
HL
Wavelet coefficient: 3channel
(b)
Original: 3channel
(c)

Figure 2. Redundant discrete wavelet decomposition and combination of input data. (a) Original image and level 1 wavelet decomposition; (b) 3-channel input data for the wavelet-model, consisting of a combination of LL, LH, and HL images; (c) 3-channel input data for the original-model, consisting of three identical original images.

output of the initial pooling layer, final fully convolutional layer, and softmax layer. Furthermore, we introduced an entropy calculation algorithm on the output side of the softmax layer. These allow for visualization of the model's behavior and prediction ambiguity. The proposed network's structure is shown in Figure 3. The ResNet50, on which it is based, is composed of 16 processing blocks and implements two types of shortcut connections. One is a module called the "conv (convolution) block" that places a convolutional layer in the shortcut (the input dimension becomes smaller than the output dimension), and the other is the "identity block" that does not place a convolutional layer in the shortcut (the input and output have the same dimension). Both modules consist of a bottleneck building block structure with three layers (1 × 1, 3 × 3, and 1 × 1 convolutional layers) and enable reduction of the number of parameters without decreasing performance. In this study, we performed retraining of the entire network. In other words, we conducted fine-tuning without placing frozen layers (layers without weight updates) and performed a two-category classification. Therefore, we replaced the final fully connected layer (Full conv) and the final classification layer (Classification) with new ones that match the number of categories.

ResNet50 requires the input data size to be 224 × 224 due to its structure, so the entire image sizes were unified using bi-cubic interpolation. Model performance evaluation was performed using 10-fold cross-validation (10% of the total data as validation data, the remaining 90% as training data), and the average

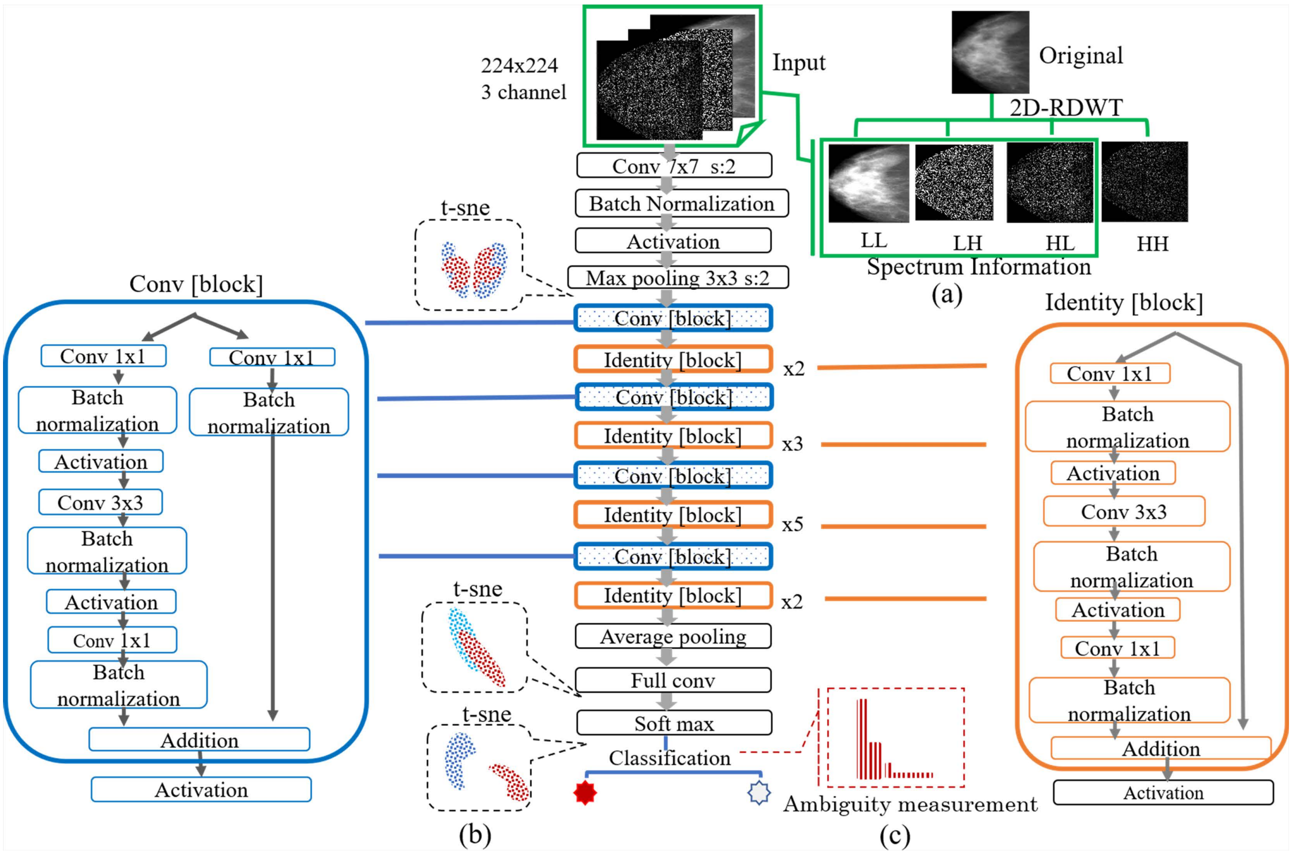


**Figure 3.** Overall structure of the proposed network. The ×2, ×3, and ×5 on the right side of the diagram indicate the number of blocks, the blue rectangle on the left side shows the structure of the Conv (Convolution) block where the input dimension changes, and the orange rectangle on the right side shows the structure of the Identity block where the input dimension remains the same. (a) Input (wavelet transform) layer; (b) 2D mapping display of activation output in the 3 different layers; (c) Visualization of prediction ambiguity.

accuracy of the 10 subsets is used as the model's accuracy. The mini-batch size is 10, and the optimizer is Adam (momentum SGD+RMSprop). In the retraining with mammograms, the parameters were adjusted so that the learning speed increases in the newly replaced fully connected layer and decreases in the transfer layer, and so that the learning speed decreases every 5 epochs. Additionally, an L2 regularization term was added to the cost function (loss function) to prevent overfitting. The number of epochs was determined by performing accuracy validation after every iteration, and re-training stops if the accuracy is lower than the highest achieved accuracy for 5 consecutive iterations.

## 2.4. Methods for Visualizing Confidence of CNN Models

We evaluated confidence calibration using reliability diagrams based on the predicted scores for the wavelet-model and original-model inputs. We also used t-distributed stochastic neighbor embedding (t-SNE) to visualize the behavior of the models (changes in data distribution as the layers deepen). Entropy was computed from the output of the softmax layer to identify the ambiguity of the model's decisions and ambiguous images. Furthermore, we used gradient-weighted class activation mapping (Grad-Cam) [34] to highlight the regions of the model's final prediction.

### 2.4.1. Confidence Calibration

Confidence calibration refers to the process of assessing how well a deep learning model's predicted probabilities reflect the true probabilities of the events being predicted [32]. In other words, it measures whether the model is overconfident or underconfident in its predictions. A well-calibrated model assigns high probabilities to events that are likely to occur, and low probabilities to events that are unlikely to occur. There are several methods to evaluate confidence calibration. Reliability diagram is a visual tool used to evaluate the confidence calibration of a deep learning-based model. The diagram plots the predicted probability on the x-axis and the observed frequency of positive labels on the y-axis. In this experiment, a positive label is BD3 and a negative label is BD2. The diagram is divided into a set of equally spaced bins based on the predicted probabilities. In each bin, the mean predicted probability and observed frequency of positive labels are calculated and plotted as points on the diagram. Ideally, the points on the reliability diagram should be on the diagonal, indicating that the model's predicted probabilities and observed frequencies of positive labels are perfectly calibrated. Points above the diagonal line indicate an underestimate, meaning the model's predicted probability is lower than the true frequency. Points below the line indicate overconfidence, meaning the model's predicted probability is larger than the true frequency. Points on the diagonal line indicate perfect calibration. Note that in this experiment the positive label is BD3 and the negative label is BD2.

The general procedure for creating a reliability diagram is as follows:

1) Divide the prediction probabilities into equally spaced bins;

2) For each bin, calculate the average predicted probability and the observed frequency of BD3 labels of the prediction (actual accuracy);

3) Plot the average predicted probability on the x-axis and the observed frequency of BD3 labels on the y-axis. Plot one point in each bin;

4) Draw a diagonal line from the bottom left corner to the top right corner of the graph. This line represents a perfectly calibrated model.

### 2.4.2. T-Distributed Stochastic Neighbor Embedding (t-SNE)

T-distributed stochastic neighbor embedding (t-SNE) is an unsupervised nonlinear dimensionality reduction algorithm developed by Laurens van der Maaten and Geoffrey Hinton in 2008 for reducing high-dimensional data to a lower-dimensional space [33]. The idea is to embed higher dimensional data points into lower dimensions such that the similarity between the data points is reflected. It has the advantage of being able to learn and perform dimensionality reduction even for relationships that cannot be expressed linearly. Specifically, dimensionality reduction is performed by reducing the difference between the probability distribution of the distance of data points in high-dimensional space and the probability distribution of the distance of data points in low-dimensional space. The probability distribution that represents the distance of data points in high-dimensional space uses the normal distribution (Gaussian distribution), while its distribution in low-dimensional space uses the t-distribution (Student-t distribution). The basic steps of t-SNE are as follows:

**Step 1** For each data point, assuming a normal distribution, the probability $p_{ij}$ is calculated from data $x_i$ and $x_j$ using Equations (1) and (2);

**Step 2** Place an equal number of data points to the total number of test data $n$ randomly in a low-dimensional space;

**Step 3** Based on the t-distribution (1 degree of freedom) from the data points $y_i$, $y_j$ in the low-dimensional space corresponding to the data points $x_i$, $x_j$ in the high-dimensional space, the probability distribution $q_{ij}$ after dimensionality reduction is calculated using Equation (3);

**Step 4** The data points in the low-dimensional space are relocated so that the two probability distributions, $p_{ij}$ and $q_{ij}$, become closer to each other;

**Step 5** Repeat steps 3 and 4 until the result converges.

$$p_{j|i} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{\left(-\left\|x_i - x_j\right\|^2\right)}{2\sigma_i^2}}{\sum_{k\neq i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{\left(-\left\|x_i - x_k\right\|^2\right)}{2\sigma_i^2}} = \frac{\exp\frac{\left(-\left\|x_i - x_j\right\|^2\right)}{2\sigma_i^2}}{\sum_{k\neq i}\exp\frac{\left(-\left\|x_i - x_k\right\|^2\right)}{2\sigma_i^2}} \tag{1}$$

$$= \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k\neq i}\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)},$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \tag{2}$$

$$q_{ij} = \frac{\left(1+\left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k\neq i}\left(1+\left\|y_k - y_i\right\|^2\right)^{-1}}, \tag{3}$$

where $i$, $j$, $k$ are sample indices, $p_{i|j}$ is the conditional probability obtained in the calculation process, and $\sigma_i$ is the standard deviation of the normal distribution centered on $x_i$. Note that $p_{ii} = q_{jj} = 0$.

In t-SNE, the probability distribution $p_{ij}$ of data points $x_i$ and $x_j$ in the high-dimensional space is mapped into a probability distribution $q_{ij}$ in the low-dimensional space, while preserving their closeness. In this case, the Kullback-Leibler divergence (KL-divergence) is employed. KL-divergence can be used as a measure of the difference between the distances of data points in high-dimensional space and those in low-dimensional space. Therefore, we aim to minimize the sum of the KL-divergences between the joint probability distributions $p_{ij}$ and $q_{ij}$ in Equation (4). If $p_{ij}$ and $q_{ij}$ have exactly the same distribution, the KL-divergence will be zero, but it increases when the distributions differ greatly.

$$C=\sum_{k\neq i} D_{kL}\left(P_i \parallel Q_i\right)=\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4}$$

Note that the result of Equation (1) varies depending on the standard deviation $\sigma_i$. When the data points around $x_i$ are dense, $\sigma_i$ should be small, and when they are sparse, should be large. Therefore, perplexity is used as a parameter to search for an appropriate $\sigma_i$. Perplexity is a parameter that adjusts the width (variance) of a normal distribution representing the distance between data points in a high-dimensional space. If perplexity is large, $\sigma_i$ also becomes large, and the contribution of neighboring points of $x_i$ located far away to learning becomes significant. In this experiment, perplexity is set to 30.

#### 2.4.3. Calculation of Classification Ambiguity Using Entropy

We measured the ambiguity of the model predictions for each input image. In this study, the entropy of the output values from the softmax layer is used as an indicator of ambiguity. Assuming the data are classified into $n$ classes, let $P_i$ be the probability of belonging to class $i$. The entropy $I$ in information theory is defined by Equation (5).

$$I=\sum_{i=1}^{n} p_i \log_2 \frac{1}{p}=-\sum_{i=1}^{n} p_i \log_2 p_i \left[\text{bit}\right] \tag{5}$$

For example, when comparing the entropy values for two models, a lower entropy indicates that the model is more certain in its predictions and therefore more accurate, while a higher entropy indicates more uncertainty and lower accuracy. In this study, we perform two-class classification of breast tissue into scattered and heterogeneous high-density categories. Therefore, if entropy is close to 1 from Equation (5), it means that the model cannot clearly determine the class to which a particular image belongs, while if it is 0, it means that the image is certainly classified into a specific class. This means that a model still could be used for classifying data, even if the model does not sufficiently understand the data.

#### 2.4.4. Gradient-Weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is a class-discriminative localization method that can generate visual explanations without requiring architecture modification or retraining. It localizes relevant image regions and emphasizes which part of the image has the largest impact on the final prediction probability score using the gradient (derivative) of the feature map of the final convolutional layer of the network. Areas with high gradients have a significant impact on the prediction result. **Figure 4** shows the flowchart of how to implement Grad-CAM:

1) Forward propagate the input images and extract the final convolutional layer and classification result;

2) Perform error backpropagation using the classification results and compute the gradient of the classification output for each element of the final convolutional layer;

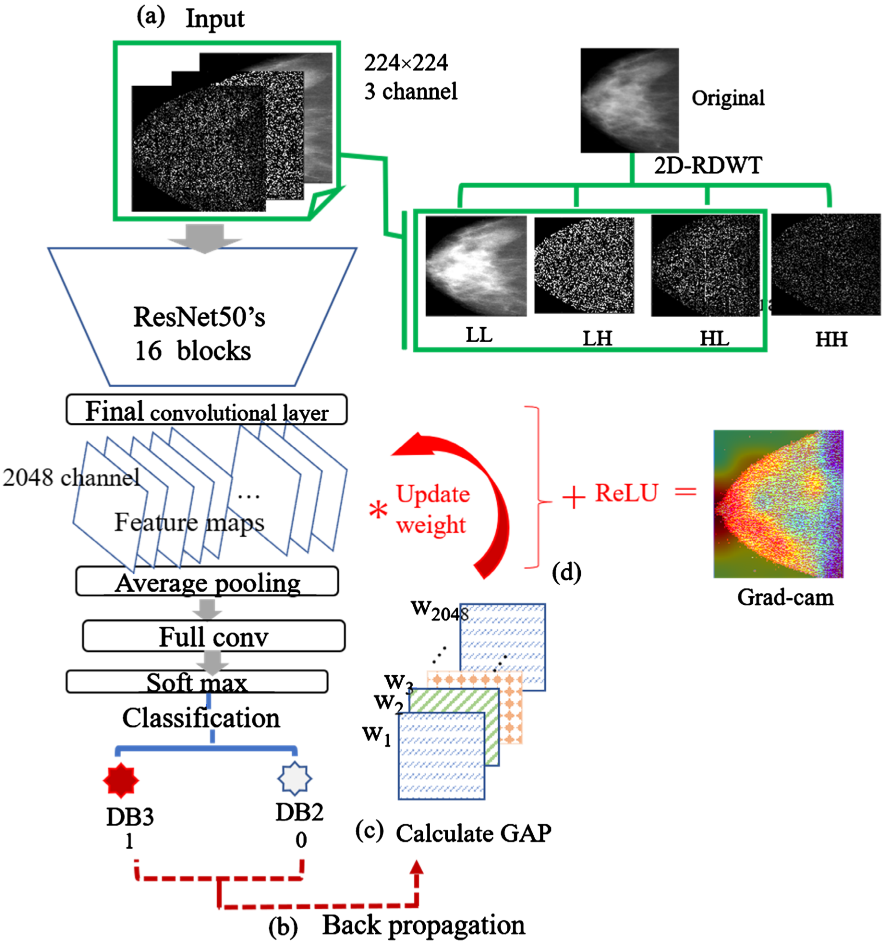


Figure 4. Schematic of the Grad-CAM. (a) Forward propagation; (b) Backpropagation; (c) Calculate GPA for each channel; (d) Gradient calculation process using ReLU function.

3) Calculate the global average pooling (GAP) of the gradients obtained in 2);

4) Take the weighted sum of the convolutional-layer values extracted in 1) with the GAP as weights, and resize to the original image size.

## 3. Results

In this study, we proposed an interpretable breast density classification model using wavelet coefficients (wavelet-model) and constructed an XAI-CAD system that can evaluate the reliability and ambiguity. For comparison, we also constructed a model using the conventional image pixel values (original-model). The accuracy (average accuracy of 10-fold cross-validation) of both models was 0.922 and 0.915, respectively, and the area under an ROC curve (AUC) was 0.974 and 0.977, respectively, with no statistically significant difference. Table 1 shows the results of the performance evaluation of the models.

Figure 5(a) and Figure 5(b) show the validation results for the reliability of the original-model and wavelet-model, respectively, indicating whether the model's score values can be trusted as probability values. The upper graphs display histograms of the mean predicted scores (frequency of occurrence for predicted

Table 1. Model performance evaluation results.

| | Original-model | Wavelet-model |
|---|---|---|
| AUC | 0.974 | 0.977 |
| Accuracy | 0.915 | 0.922 |
| Recall | 0.925 | 0.925 |
| Specificity | 0.905 | 0.920 |
| precision | 0.906 | 0.920 |

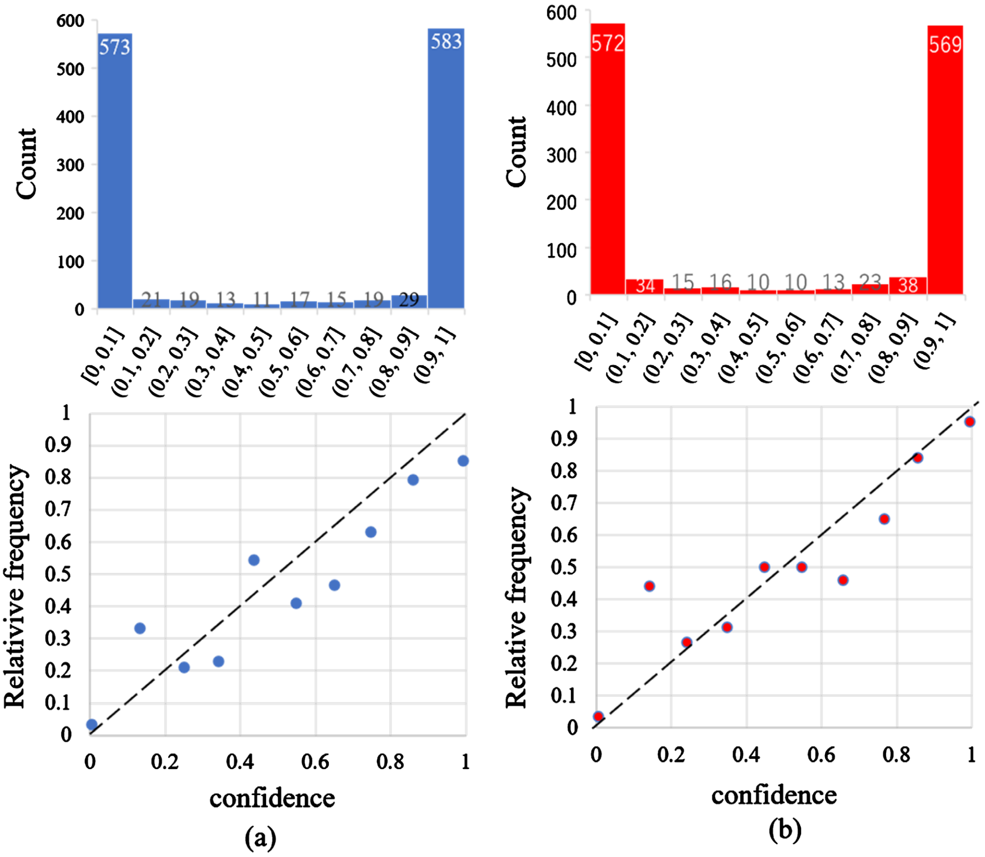
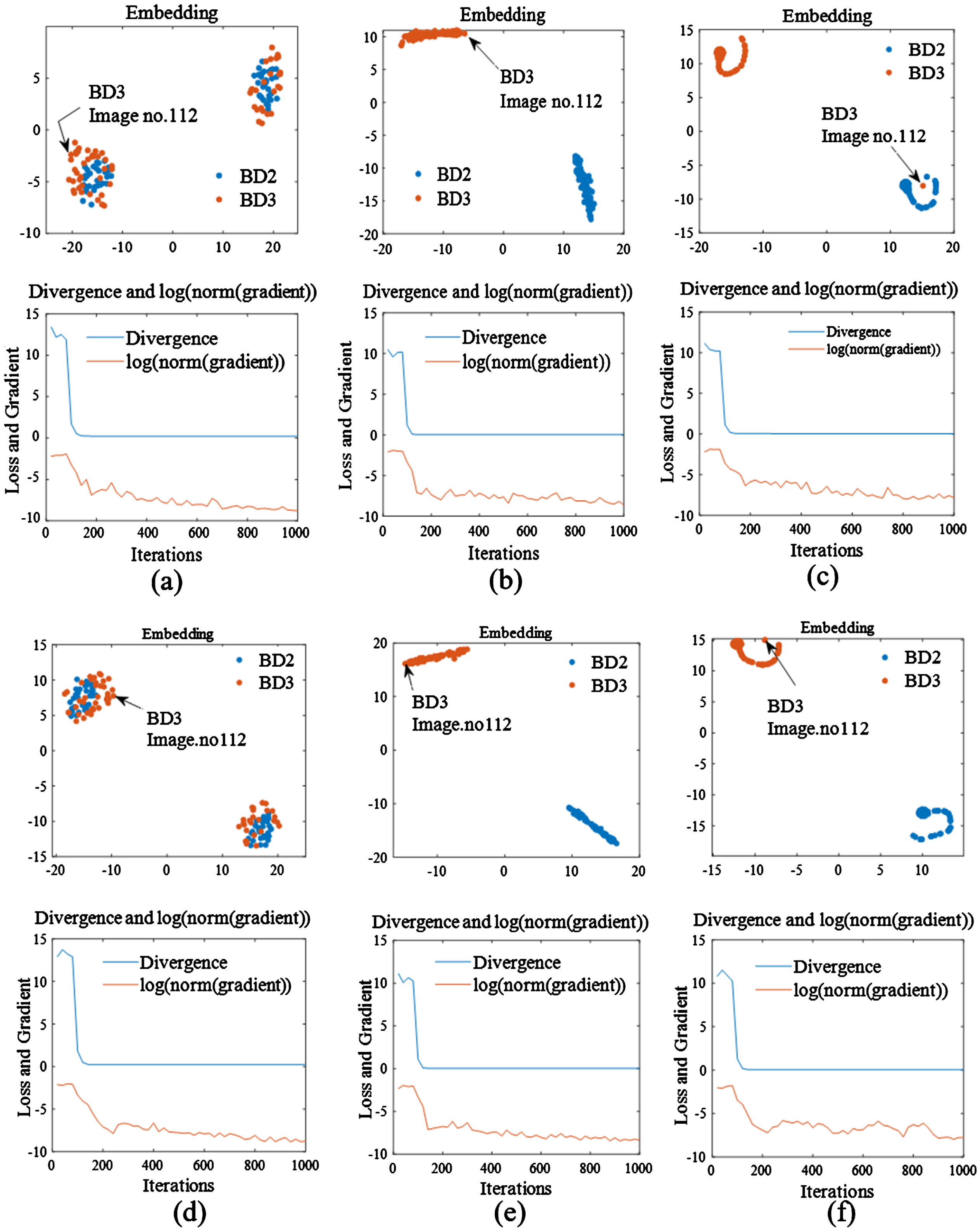


Figure 5. Confidence histograms (top) and reliability diagrams. (a) Original-model; (b) Wavelet-model. The upper graphs show the frequency of occurrence against predicted scores, while the lower graphs are reliability diagrams with dashed lines indicating perfect calibration.

scores), while the lower graphs show reliability diagrams (ratio of BD3 labels per bin to the mean of predictive score values per bin). The number of bins used in this experiment is 10. The closer the plotted point is to the dashed line in the reliability diagrams, the more reliable the prediction score is [32].

In this experiment, we aimed to verify the uncertainty of the model in response to changes in the dataset. To achieve this, we used the original images of the 10 subsets of 10-fold cross-validation as a common dataset for both models and visualized the behavior of the models (variation in data distribution). An example of the results is shown in Figure 6. The classification accuracies of the



(a) (b) (c)

(d) (e) (f)

Figure 6. Behavior of the two models. (a)-(c) are original-model, (d)-(f) are wavelet-model. The upper rows for both models are distribution plots of high-dimensional data mapped into two dimensions. The lower rows show the KL-divergence and gradient norm for the two-dimensional mapping. (a) and (d) are the activation outputs of the initial pooling layer; (b) and (e) are the activated outputs of the final convolutional layer; (c) and (f) are the activation outputs of the softmax layer.

original-model and wavelet-model shown in the figures are 0.99 and 1.0, respectively. The arrows in the figure indicate the data that became isolated points in the final distribution of the original-model. The lower part of the figure shows the changes in KL-divergence and gradient norm with iteration counts.

Figure 7 shows the entropy values of the model predictions for the subsets shown in Figure 6. Figure 8 shows the original image of the DB3 Image no. 112 data, which was an isolated point in the original-model, and the regions that serve as the basis for the predictions of both models. A heat map was given to the image data using Grad-CAM for visualization. As the heat map shifts from blue to red, it gives more significant information about the influence of the prediction results.

We compared the reliability of both models using each subset of 10-fold cross-validation. An example of the results is shown in Figure 9. Figure 9(a) and

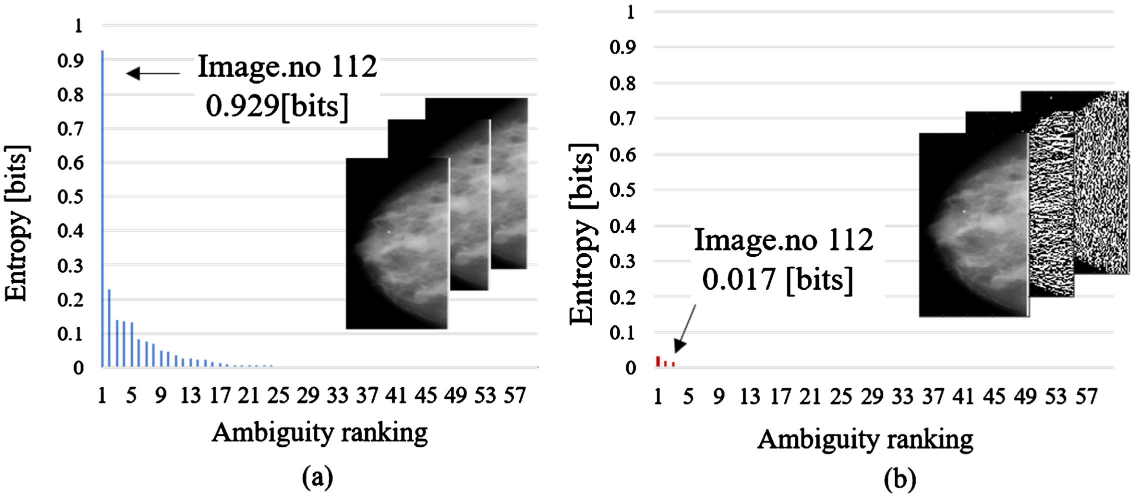


Figure 7. Ambiguity in judgment and extracted data. (a) original-model; (b) wavelet-model. The images in the graphs represent the input data (BD3 Image no.112) that resulted in a false negative for the original-model. The arrows in the graphs indicate the entropy value of BD3 Image no.112.

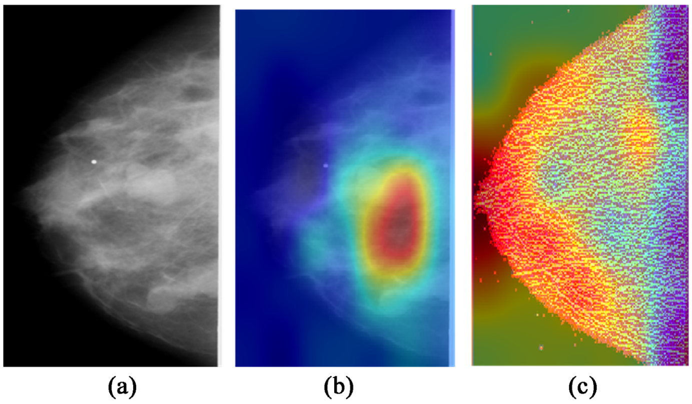


Figure 8. Original image and region of interest for each model's prediction (Grad-cam). (a) Original image of the data (DB3 Image.no112) that became an isolated point in the original-model; (b) Predicted region of interest for BD3 Image.no112 in original-model; (c) Predicted region of interest for DB3 Image.no112 in wavelet-model.

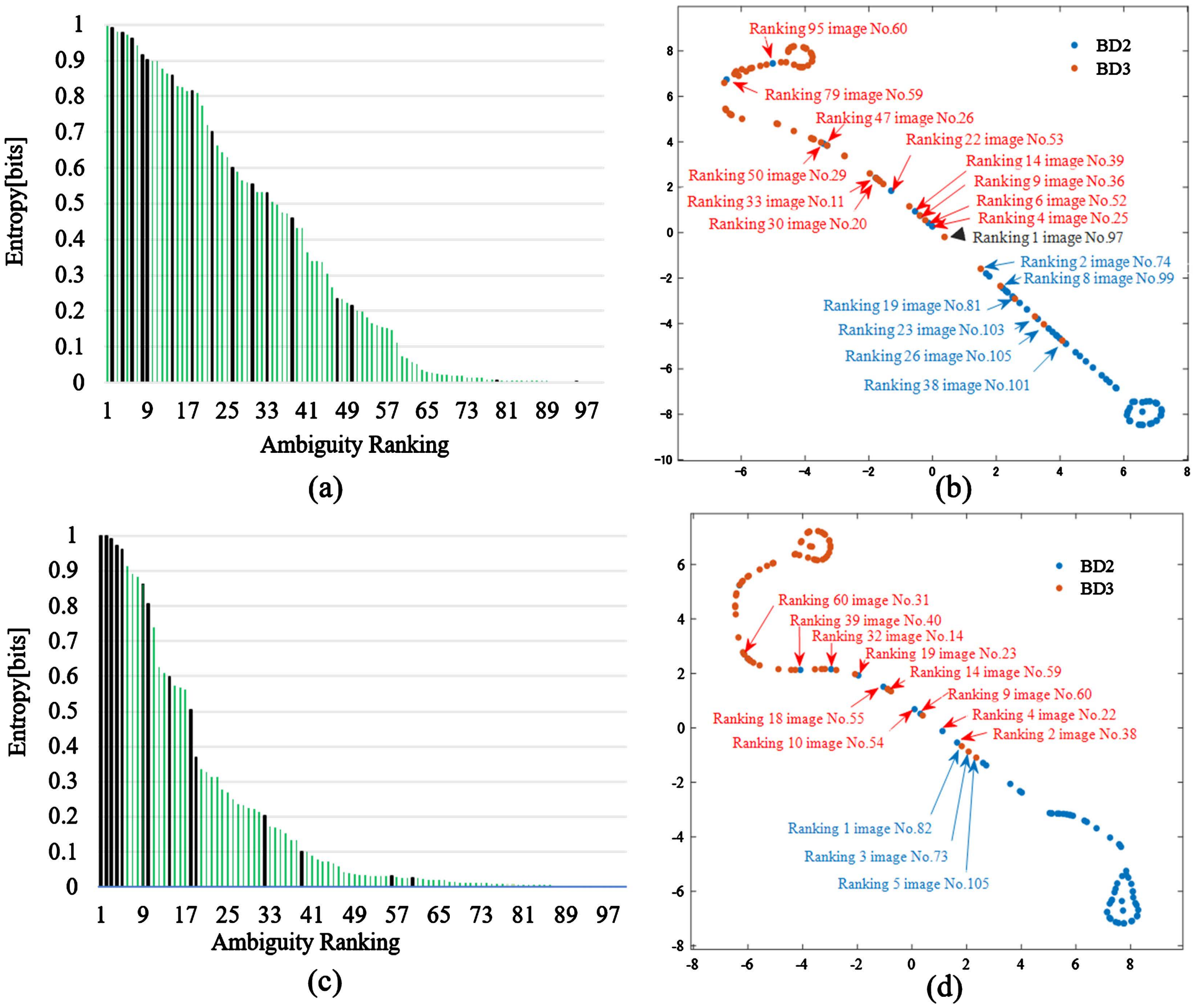


Figure 9. Prediction ambiguity graphs and distributions of prediction. (a) and (c) are the entropy values for the top 100 most ambiguous cases in the original-model and wavelet-model cases, respectively, and the black bars indicate misclassifications. (b) and (d) are the probability distributions of the final predictions; red and blue letters in the figures indicate false positive and false negative data, respectively.

Figure 9(c) show the entropy values for the top 100 most ambiguous images. Black bars indicate misclassified data. Figure 9(b) and Figure 9(d) are two-dimensional mapping by t-SNE. Proper classification results in two clusters, red and blue, but misclassification results in a mixture of red and blue. In Figure 9(b) and Figure 9(d), red text represents false-positive data, while blue text represents false-negative data.

## 4. Discussion

In the performance evaluation of the proposed model and the conventional model, as shown in Table 1, both models demonstrated high accuracy exceeding 90%. While the overall results were good, these evaluation methods are insufficient for evaluating error distribution. That is, it does not take into consideration the problem of the model being over-confident, where it outputs large scores despite not being able to make correct predictions. Therefore, in this study, we conducted an investigation into the uncertainty of the prediction scores output obtained by the two models. The results are shown in **Figure 5**. The top row shows the histograms of the mean prediction scores. Regarding the results of the verification using 1300 input images (650 negative (DB2) and 650 positive (BD3)), in the original-model, 573 images were predicted negative with a mean prediction score of 0 to 0.1, while 583 images were predicted positive with a score of 0.9 to 1.0 (**Figure 5(a)**). On the other hand, in the wavelet-model, 572 images were predicted negative with a mean prediction score of 0 to 0.1, while 569 images were predicted positive with a score of 0.9 to 1.0 (**Figure 5(b)**). This means that the original-model assigns higher scores and predicts more positive cases. The bottom row shows reliability diagrams. A reliability diagram visually expressed how well a model is calibrated. When a model is perfectly calibrated, the accuracy (the relative frequency of positives) becomes an identity function of reliability (the mean predicted score) and is plotted on the dashed line in the figure. If the plot is below the perfectly calibrated line, the reliability is greater than the accuracy, which means that the model is overconfident. On the other hand, if the plot is above the line, the accuracy is greater than the reliability, which means that the model is underestimated. Both models are not perfect. In the case of the original-model, the plots in the figure are generally located lower than the perfectly calibrated line (dashed line). On the other hand, the plots of the wavelet-model are closer to the dashed line, indicating that it is well-calibrated. Specifically, when the average predicted score is between 0.9 and 1.0, in the case of the original-model (**Figure 5(a)**), approximately 85% of the predictions are in the positive (BD3) class. As mentioned above, the histograms in the upper row suggest that the original-model assigns high scores and makes positive predictions, but the output of the model cannot be taken as a probabilistic meaning as it is. On the other hand, in the wavelet-model (**Figure 5(b)**), when the average predicted score is between 0.9 to 1.0, approximately 95% of the samples are classified as positive (BD3) class, and the model's score can be treated as a probability. Therefore, it can be said that the results of the wavelet-model are more reliable.

**Figure 6** visualizes the behavior of input data as it passes through the layers of the network. The top rows of **Figures 6(a)-(c)** and **Figures 6(d)-(f)** show the distribution changes of a certain subset using t-SNE for the original-model and wavelet-model, respectively. The bottom rows show the KL-divergence and gradient norm. As the number of iterations increases, the values decrease, indicating that high-dimensional probability distributions are effectively mapped to lower dimensions. Dense clusters in the t-SNE plots correspond to classes that are inherently classified correctly. Generally, when pixel information of an image is input to a CNN, the initial layers tend to act on low-level features such as edges and luminance. Therefore, as shown in **Figure 6(a)**, the original-model does not show the clustering of the target BD2 and BD3 classes. From **Figure 6(d)**, it can be seen that the wavelet-model, which takes image spectrum information as input, exhibits a similar tendency. Moreover, as the layers become deeper in both models, clustering for each class is shown. However, in Figure 6(c) and Figure 6(f), in the original-model, the BD3 (image No.112) is displayed as an isolated point in the BD2 (blue) cluster. On the other hand, in the wavelet-model, both BD2 (red) and BD3 (blue) are appropriately clustered, indicating correct classification. In both models, the output of the final convolutional layer (Figure 6(b) and Figure 6(e)) is appropriately clustered. This suggests model uncertainty rather than data uncertainty. In other words, it can be inferred that the original-model has higher ambiguity in the final decision compared to the wavelet-model.

In this proposed method, the information entropy is introduced as an algorithm to calculate the level of ambiguity in the softmax layer. As this experiment is a binary classification, the highest ambiguity is 1 bit, and the lowest is 0 bit when there is no ambiguity. As shown in Figure 7, the ambiguity of image No.112 (an isolated data point in Figure 6(c)) is high at 0.929. This means that the original-model cannot clearly determine the class it should be classified into. The reason may be that the original-model has not learned the differences in similar features or that the image contains elements of both classes, causing confusion in the model. On the other hand, wavelet-model has a very low entropy value of 0.017, which suggests that it is making a confident decision. Figure 8 shows the regions of interest that the models focused on when making the prediction for Image No.112. The red regions had the most influence on the prediction, while the blue regions had little influence. In the original-model (Figure 8(b)), it can be seen that the model focused more on the tumor area rather than the mammary region of the image. This may have caused confusion in the model. On the other hand, the wavelet-model (Figure 8(c)) is correctly classified by focusing on the mammary gland region and has high certainty. As shown in Figure 7, the wavelet-model is generally less ambiguous. This suggests that the wavelet-model provides a clearer judgment than the original-model prediction. Figure 9 shows the measurement results of ambiguity for a certain subset and the probability distribution of the final output. From Figure 9(a) and Figure 9(c), it can be seen that the wavelet-model has overall lower entropy values and lower ambiguity than the original-model. However, despite low ambiguity in both models, some data are misclassified (black bars). It is considered that this is caused by confusion in judgment of similar images. However, for the wavelet-model, it turns out that over-confidence does not occur in BD2. From this result, it can be said that the proposed wavelet-model is better in terms of reliability.

In this study, we constructed a breast density classification XAI-CAD system using spectral information from mammograms. In addition to conventional accuracy validation, we verified the reliability of the model from the histogram of the mean prediction score and reliability diagram, and demonstrated that the model's prediction scores are reliable. In the proposed model, t-SNE was used to visualize the behavior of input data passing through the layers of the network as probability distributions. Through visualization of these probability distributions, we were able to detect isolated data points, misclassified data, and data points near the classification boundary. In addition to the explainability of the model, we believe that it has a wide range of generalizations, such as data cleaning, model retraining, and reviewing data-label reviewing. In the proposed model, we also introduced an algorithm for calculating information entropy as a quantitative measure of uncertainty. In the present study, it was easy to determine the ambiguity from the prediction score because of the two-class classification. However, in multi-class classification, since various scores are assigned, we believe that evaluating uncertainty using information entropy is a useful evaluation method for detecting ambiguous data. In the final prediction, we visualized the model's reasoning using Grad-CAM, which not only allowed us to identify the causes of misclassification and adjust the training data but also enhanced the transparency and interpretability of the model. However, there are limitations to this study. If the model is properly calibrated, the accuracy (relative frequency of positives) should be an identity function of the confidence (mean predicted score). Nevertheless, our model was not able to achieve this result. Additionally, the shape of clusters in t-SNE can be altered by parameter tuning. These considerations should be addressed as future tasks. Moreover, how to apply the proposed model in a clinical setting will be one of our future work.

## 5. Conclusion

We used spectral information from mammograms to construct an interpretable CNN-based system for breast density classification. We evaluated whether the prediction score of the classification model is a highly reliable probability value using a reliability diagram, and visualized the basis for the final prediction using Grad-CAM. In constructing the classification model, we modified ResNet50 and introduced algorithms for extracting and inputting image spectra, visualizing network behavior using t-SNE, and quantifying prediction ambiguity using information entropy. Experimental results show that the proposed model can demonstrate not only high classification accuracy but also high reliability and interpretability compared to conventional CNN models that use image pixel information. Furthermore, the proposed model can detect misclassified data and explicitly indicate the basis of prediction. The results demonstrated the effectiveness and usefulness of our proposed model from the perspective of credibility and transparency.

## Acknowledgements

This work was supported in part by JSPS KAKENHI (Grant-in-Aid for Scientific Research) Grant Number 18K15641.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Breast Cancer Facts and Statistics. https://www.breastcancer.org/facts-statistics

[2] Early Detection: Breast Health Awareness and Early Strategies. The Global Breast Cancer Initiative. https://www.paho.org/hq/dmdocuments/2016/KNOWLEDGE-SUMMARY---EARLY-DETECTION.pdf

[3] Broeders, M., Moss, S., Nyström, L., *et al.* (2012) The Impact of Mammographic Screening on Breast Cancer Mortality in Europe: A Review of Observational Studies. *Journal of Medical Screening*, **19**, 14-25. https://doi.org/10.1258/jms.2012.012078

[4] Boyd, N.F., Lockwood, G.A., Martin, L.J., Knight, J.A., Jong, R.A., *et al.* (1999) Mammographic Densities and Risk of Breast Cancer among Subjects with a Family History of This Disease. *Journal of the National Cancer Institute*, **91**, 1404-1408. https://doi.org/10.1093/jnci/91.16.1404

[5] Ursin, G., Ma, H., Wu, A.H., Bernstein, L., Salane, M., Parisky, *et al.* (2003) Mammographic Density and Breast Cancer in Three Ethnic Groups. *Cancer Epidemiology, Biomarkers & Prevention*, **12**, 332-338. https://pubmed.ncbi.nlm.nih.gov/12692108/

[6] Boyd, N.F., Rommens, J.M., Vogt, K., Lee, V., Hopper, J.L., *et al.* (2005) Mammographic Breast Density as an Intermediate Phenotype for Breast Cancer. *The Lancet Oncology*, **6**, 798-808. https://doi.org/10.1016/S1470-2045(05)70390-9

[7] Breast Composition Is an Important Factor in Managing Breast-Cancer Risk. https://www.uclahealth.org/Workfiles/clinical_updates/oncology/16v3-11-BreastDensity.pdf

[8] Dense Breasts: Answers to Commonly Asked Questions. https://www.cancer.gov/types/breast/breast-changes/dense-breasts

[9] America College of Radiology (2019) ACR BI-RADS Atlas 5th Edition. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-RadsDeep

[10] Bodewes, F.T.H., van Asselt, A.A., Dorrius, M.D., Greuter, M.J.W. and de Bock, G.H. (2022) Mammographic Breast Density and the Risk of Breast Cancer: A Systematic Review and Meta-Analysis. *Breast*, **66**, 62-68. https://doi.org/10.1016/j.breast.2022.09.007

[11] Mohamed, A.A., Berg, W.A., Peng, H., *et al.* (2018) A Deep Learning Method for Classifying Mammographic Breast Density Categories. *Medical Physics*, **45**, 314-321. https://doi.org/10.1002/mp.12683

[12] Berg, W.A., Campassi, C., Langenberg, P. and Sexton, M.J. (2000) Breast Imaging Reporting and Data System: Inter- and Intraobserver Variability in Feature Analysis and Final Assessment. *American Journal of Roentgenology*, **174**, 1769-1777. https://doi.org/10.2214/ajr.174.6.1741769

[13] Winkler, N.S., Raza, S., Mackesy, M. and Birdwell, R.L. (2015) Breast Density: Clinical Implications and Assessment Methods. *RadioGraphics*, **35**, 316-324. https://doi.org/10.1148/rg.352140134

[14] Chan, H.-P. and Helvie, M.A. (2019) Deep Learning for Mammographic Breast Density Assessment and Beyond. *Radiology*, **290**, 59-60. https://doi.org/10.1148/radiol.2018182116

[15] Lawrence, S., Giles, C.L., Tsoi, A.C. and Back, A.D. (1997) Face Recognition: A Convolutional Neural-Network Approach. *IEEE Transactions on Neural Networks*, **8**, 98-113.

[16] Pan, S.J. and Yang, Q. (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[17] Boyd, N.F., Guo, H., Martin, L.J., *et al.* (2007) Mammographic Density and the Risk and Detection of Breast Cancer. *The New England Journal of Medicine*, **56**, 227-236. https://doi.org/10.1056/NEJMoa062790

[18] McCormack, V.A. and dos Santos Silva, I. (2006) Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-Analysis. *Cancer Epidemiology, Biomarkers & Prevention*, **15**, 1159-1169. https://doi.org/10.1158/1055-9965.EPI-06-0034

[19] Mahmood, T., Li, J., Pei, Y., Akhtar, F., Rehman, M.U. and Wast, S.H. (2022) Breast Lesions Classifications of Mammographic Images Using a Deep Convolutional Neural Network-Based Approach. *PLOS ONE*, **17**, 1-25. https://doi.org/10.1371/journal.pone.0263126.

[20] Li, S., Wei, J., Chan,H.-P., Helvie, M.A., Roubidoux, M.A., *et al.* (2018) Computer-Aided Assessment of Breast Density: Comparison of Supervised Deep Learning and Feature-Based Statistical Learning. *Physics in Medicine & Biology*, **63**, Article ID: 025005. https://doi.org/10.1088/1361-6560/aa9f87

[21] Dağlarli, E. (2020) Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In: Aceves-Fernandez, M.A., Ed., *Advances and Applications in Deep Learning*, IntechOpen, London, 1-17. https://www.intechopen.com/chapters/72398 https://doi.org/10.5772/intechopen.92172

[22] Singh, A., Sourya, S. and Lakshminarayanan, V. (2020) Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*, **6**, 52. https://doi.org/10.3390/jimaging6060052

[23] van der Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A. and Viergever, M.A. (2022) Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis. *Medical Image Analysis*, **79**, Article ID: 102470. https://doi.org/10.1016/j.media.2022.102470

[24] Brunese, L., Mercaldo, F., Reginelli, A. and Santone, A. (2020) Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-Rays. *Computer Methods and Programs in Biomedicine*, **196**, Article ID: 105608. https://doi.org/10.1016/j.cmpb.2020.105608

[25] Cheng, C.-T., Ho, T.-Y., Lee, T.-Y., Chang, C.-C., *et al.* (2019) Application of a Deep Learning Algorithm for Detection and Visualization of Hip Fractures on Plain Pelvic Radiographs. *European Radiology*, **29**, 5469-5477. https://doi.org/10.1007/s00330-019-06167-y

[26] Huff, D.T., Weisman, A.J. and Jeraj, R. (2021) Interpretation and Visualization Techniques for Deep Learning Models in Medical Imaging. *Physics in Medicine and Biology*, **66**, 04TR01. https://doi.org/10.1088/1361-6560/abcd17

[27] Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S.H., *et al.* (2019) An Explainable Deep-Learning Algorithm for the Detection of Acute Intracranial Haemorrhage from Small Datasets. *Nature Biomedical Engineering*, **3**, 173-182. https://www.nature.com/articles/s41551-018-0324-9#citeas https://doi.org/10.1038/s41551-018-0324-9

[28] Liao, W., Zou, B., Zhao, R., Chen, Y., He, Z. and Zhou, M. (2019) Clinical Interpretable Deep Learning Model for Glaucoma Diagnosis. *IEEE Journal of Biomedical and Health Informatics*, **24**, 1405-1412. https://doi.org/10.1109/JBHI.2019.2949075

[29] Lei, Y., Tian, Y., Shan, H., Zhang, J., Wang, G. and Kalra, M.K. (2020) Shape and Margin-Aware Lung Nodule Classification in Low-Dose CT Images via Soft Activation Mapping. *Medical Image Analysis*, **60**, Article ID: 101628. https://doi.org/10.1016/j.media.2019.101628

[30] Matsuyama, E., Takehara, M. and Tsai, D.-Y. (2020) Using a Wavelet-Based and Fine-Tuned Convolutional Neural Network for Classification of Breast Density in Mammographic Images. *Open Journal of Medical Imaging*, **10**, 17-29. https://doi.org/10.4236/ojmi.2020.101002

[31] Mohamed, A.A., Berg, W.A., Peng, H., *et al.* (2018) A Deep Learning Method for Classifying Mammographic Breast Density Categories. *Medical Physics*, **45**, 314-321. https://doi.org/10.1002/mp.12683

[32] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) On Calibration of Modern Neural Networks. *Proceedings of the* 34*th International Conference on Machine Learning*, *PMLR*, Vol. 70, 1321-1330. https://arxiv.org/pdf/1706.04599.pdf

[33] van der Maaten, L. and Hinton, G. (2008) Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2579-2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html

[34] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/ICCV.2017.74

[35] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[36] ImageNet. http://www.image-net.org

[37] The Cancer Imaging Archive (TCIA). https://www.cancerimagingarchive.net

[38] Matsuyama, E., Tsai, D.-Y., Lee, Y., *et al.* (2013) A Modified Undecimated Discrete Wavelet Transform Based Approach to Mammographic Image Denoising. *Journal of Digital Imaging*, **26**, 748-758. https://doi.org/10.1007/s10278-012-9555-6 https://link.springer.com/article/10.1007/s10278-012-9555-6

[39] Daubechies, I. (1992) Ten Lectures on Wavelets. The Society for Industrial and Applied Mathematics, Philadelphia, US Patent No. 821393. https://doi.org/10.1137/1.9781611970104