frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 01 September 2025
DOI 10.3389/fcomp.2025.1569017

OPEN ACCESS

EDITED BY
Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY
Amrutanshu Panigrahi,
Siksha 'O' Anusandhan University, India
Mosiur Rahaman,
Asia University, Taiwan

*CORRESPONDENCE
Indo Intan
✉ indo.intan@undipa.ac.id

RECEIVED 31 January 2025
ACCEPTED 24 June 2025
PUBLISHED 01 September 2025

CITATION
Intan I, Karnyoto AS, Harlina S, Nelwan BJ, Setiawan D, Yamin A and Puspitasari RE (2025) Heterogeneous ensemble learning: modified ConvNextTiny for detecting molecular expression of breast cancer on standard biomarkers. *Front. Comput. Sci.* 7:1569017. doi: 10.3389/fcomp.2025.1569017

Heterogeneous ensemble learning: modified ConvNextTiny for detecting molecular expression of breast cancer on standard biomarkers

Indo Intan[1]*, Andrea Stevens Karnyoto[2], Sitti Harlina[3], Berti Julian Nelwan[4], Devin Setiawan[5], Amalia Yamin[6] and Ririn Endah Puspitasari[4]

[1]Department of Informatics Engineering, Concentration of Intelligent Systems, Dipa Makassar University, Makassar, Indonesia, [2]Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, [3]Department of Informatics Engineering, Concentration of Data Science, Dipa Makassar University, Makassar, Indonesia, [4]Department of Anatomical Pathology, Faculty of Medicine, Hasanuddin University, Makassar, Indonesia, [5]Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, United States, [6]Laboratory of Anatomical Pathology, Wahidin Sudirohusodo Hospital, Makassar, Indonesia

Breast cancer is the highest-ranking type of cancer, with 2.3 million new cases diagnosed each year. Immunohistochemistry (IHC) is the gold standard "examination" for determining the expression of cancer malignancies in patients with the ultimate goal of determining prognosis and therapy. Immunohistochemistry refers to the four WHO standard biomarkers: estrogen receptor, progesterone receptor, human epidermal growth factor receptor-2, and Ki-67. These biomarkers are assessed based on the quantity of cell nuclei and the intensity of brown cell membranes. Our study aims to detect the expression of breast cancer malignancy as an initial step in determining prognosis and therapy. We implemented homogeneous and heterogeneous ensemble learning models. The homogeneous ensemble learning model uses the majority vote technique to select the best performance between the Xception, ResNet50V2, InceptionResNet50V2, and ConvNextTiny models. The heterogeneous ensemble learning model takes the ConvNextTiny model as the best model. Feature engineering in ConvNextTiny combines convolution and cell-quantification features as feature fusion. ConvNextTiny, which applies feature fusion, can detect the expression of cancer malignancy. Heterogeneous ensemble learning outperforms homogeneous ensemble learning. The model performs well for accuracy, precision, recall, F_1-score, and receiver operating characteristic-area under the curve (ROC-AUC) of 0.997, 0.973, 0.991, 0.982, and 0.994, respectively. These results indicate that the model can classify the malignancy expressions of breast cancer well. This model still requires the configuration of the visual laboratory device to test the real-time model capabilities.

KEYWORDS

breast cancer, ConvNextTiny, ensemble learning, Canny, Otsu, IHC, ER/PR/Ki-67, HER-2

1 Introduction

Breast cancer (BC) ranks first in women among all types of cancer in the world (Intan et al., 2024). Approximately 2.3 million cases spread across various countries, and as many as 666,103 deaths (Alismail, 2024). Asian countries have the highest number of 985,817 cases and 315,309 deaths (World Health Organization, 2022). In line with global cases, in Indonesia, breast cancer ranks first in cancer cases in women. Global cancer data in 2022 shows that the total cases in women are 30.1% (66,271 cases) and 19.8% (22,598 cases) (World Health Organization, 2022).

Breast cancer is a type of cancer that occurs when malignant cells grow in breast tissue and is heterogeneous, characterized by various molecular subtypes and genotype profiles (Chen et al., 2024). These cells can form tumors that can be felt on physical examination or detected through mammography. Breast cancer is more common in women but can also occur in men in very small numbers (Kemenkes, 2024). Breast cancer has various presentations with different molecular subtypes, with different biomolecular, pathological, and genetic features, and with different clinical and therapeutic response results, so breast cancer is called a heterogeneous disease. These molecular markers are known to be closely related to oncogenic transformation, cancer cell proliferation, tumor growth, treatment options, and prognosis of breast cancer (Joensuu et al., 2013; Afkari et al., 2021).

An immunohistochemistry (IHC) is an examination to determine the characteristics of breast cancer. The examination involves biomarkers, which are widely used in the process of diagnosis, prognosis, and treatment for patients with breast cancer. Biomarkers are useful for both patients who have recently had breast cancer and those who have experienced a recurrence. There are four types of biomarkers, which are the WHO gold standard that is routinely enforced in the characterization and diagnosis of breast cancer, namely estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor-2 (HER-2), and Ki-67. Especially for ER, PR, and HER-2, implementation is easy because it is effective and inexpensive (Alismail, 2024).

Estrogen receptors (ERs) and progesterone receptors (PRs) are both important for predicting breast cancer pathogenesis and treatment response. Both are hormone receptors in response to estrogen and progesterone; ER and PR contribute to cancer growth in hormone-sensitive breast tissue by facilitating cancer cell proliferation (Alismail, 2024). Estrogen receptors (ERs) have remained the most important biomarker in breast oncology for 60 years after their discovery. ER status is very urgent in clinical decisions and outcome prediction for breast cancer patients, including determining the right therapy for patients. The results can significantly improve clinical outcomes with ER-positive characteristics. As an important predictive biomarker, visualization of its image requires analysis that meets the standard scoring of cell nuclei against stained cell nuclei. The majority of BC are ER-positive. Visually, ER has "positive" and "negative" characteristics. Negative characteristics if ER staining is $\geq$1% of the cell nucleus by IHC (Allison et al., 2021; Reinert et al., 2022; Loggie et al., 2024). Another scoring standard, ER-negative, has a cell nucleus threshold of $\leq$10% (Fei et al., 2021) or ranges from 2 to 7% (Loggie et al., 2024). Progesterone receptor (PR) is a member of the nuclear/steroid hormone receptor (SHR) family of ligand-dependent transcription factors expressed primarily in female reproductive tissues and the central nervous system. In response to the binding of its related steroid hormone, progesterone, PR regulates the expression of a network of genes to control the development, differentiation, and proliferation of target tissues, as well as pathological processes in endocrine-based cancers (Grimm et al., 2016). PR characteristics include "positive" and "negative." Visually, PR is "positive" if the score of stained cell nuclei has a cutoff >1% (Shao et al., 2024) or >10%, conversely if PR is "negative," then the score of stained cell nuclei is <10% (2–7%) (Alismail, 2024; Loggie et al., 2024).

Human epidermal growth factor receptor 2 (HER-2) is overexpressed in approximately 15–30% of breast cancer cases. Thus, it is considered an important prognostic and predictive biomarker for breast cancer. Unfortunately, HER-2 overexpression is associated with a more aggressive tumor phenotype characterized by prone metastasis, poor prognosis, and high recurrence rates. This suggests that HER-2-positive breast cancer is often associated with more advanced stages (Alismail, 2024). Routine determination of HER-2 status is performed using techniques such as immunohistochemistry (IHC) and fluorescence *in situ* hybridization (FISH). FISH detects HER-2 gene amplification, while IHC evaluates HER-2 protein expression levels; both techniques determine eligibility for HER-2-targeted therapy (Lv et al., 2016). Accurate determination of HER-2 status influences therapy choice and prognosis, which are critical for the best patient management (Alismail, 2024). HER-2 has a score that is taken from ER and PR. At the cutoff limit of 10%, the threshold <10% is HER-2 negative, conversely, if the threshold $\geq$10 is HER-2 "positive." HER-2 is divided into three subtypes based on IHC scores: "negative" (IHC 0/1+), equivocal cases (IHC 2+), and "positive" cases (IHC 3+). Equivocal cases are retested with FISH to verify their HER-2 expression more accurately. Positive cases indicate that patients are eligible for anti-HER-2 therapy (Lv et al., 2016).

Ki-67 is a widely used biomarker to measure and monitor tumor proliferation in breast specimens, although there is poor agreement on the analytical approach to its assessment, assessment methods and cutoffs, data handling, and appropriate clinical utility of the biomarker. Ki-67 appears to be a continuously variable type marker, reflecting tumor biology (Penault-Llorca and Radosevic-Robin, 2017). Testing for Ki-67 is performed using different methods, and cutoffs for defining Ki-67 "positive" and "negative" or "high" and "low" populations are not clear. Consequently, the Tumor Marker Guidelines Committee of the American Society of Clinical Oncology (ASCO) determined that the evidence supporting the clinical utility of Ki-67 is insufficient to recommend routine use of this marker for prognosis in patients with newly diagnosed breast cancer. Standardization of Ki-67 assessment is a global standard set by the WHO to improve its reproducibility. The clinical utility of very low and very high Ki-67 indices is good. The 25% threshold has shown significance for predicting overall survival. Multigene testing can provide useful information to guide the management of patients with ER+/HER-2 breast cancer in the "gray zone" Ki-67 index (between 15 and 25%) (Penault-Llorca and Radosevic-Robin, 2017).

Therefore, the four biomarkers visually indicate the expression of cancer malignancy through the number of cells and the extent of brown color between the stained cells. Determining the characteristics of each biomarker will provide appropriate treatment recommendations for breast cancer patients.

Pathologists have difficulty observing tissues with the naked eye and manually analyzing images based on their knowledge and skills. First, they perform fundamental techniques through microscopic

observation of cell morphological structures (nuclei and cell membranes). The method relies heavily on manual naked-eye observation, so it does not save time because IHC images have complex, uneven cell color distribution between normal and cancerous stained cells, overlapping cells, and uncertain cell sizes. Second, the objectivity of the observation results depends on the experience and accommodation of the pathologist's eyes, so the results sometimes differ. Third, there are so many cases of breast cancer in hospitals that it is tiring if the process only relies on manual techniques. On the other hand, the demand for examination results must be released quickly to patients as part of hospital management.

Therefore, this study proposes a modified ConvNextTiny to detect breast cancer malignancy expression, combining cell quantification and convolution features. The cell quantification feature adapts cell morphology as a fundamental pathologist calculation. The convolution feature is a ConvNextTiny feature that utilizes a convolutional neural network architecture and pre-trained transfer learning weights. The cell quantification calculation is a WHO calculation standard in the ConvNextTiny feature that performs well with transfer weights from "imagenet" (transfer learning weights). Combining both is an advantage in finding unique patterns for each breast cancer image that will improve model performance.

In this study, we experimented with two ensemble transfer learning models, namely homogeneous and heterogeneous ensemble learning. The homogeneous ensemble learning model was constructed using a majority voting scheme among four models: Xception, Resnet50V2, InceptionResnet50v2, and ConvNextTiny. We trained each model individually and performed hyperparameter tuning, with particular focus on evaluating the learning rate. ConvNextTiny was selected as it shows dominant performance within the homogeneous ensemble. Combining the ConvNexTiny and cell quantification model features is the final ensemble model by concatenating their features into the ConvNexTiny neural network classifier. The advantage of the homogeneous ensemble learning model is based on transfer learning from CNN, which has high computational feature extraction and classification capabilities, and ConvNexTiny has a short computational time. Moreover, using cell quantifications involves a practical model and simple computation. Combining these two models using feature fusion balances the complexity of the algorithm and delivers better performance and acceleration. The final model classification result is in the form of ER and PR expressions: "positive" and "negative"; HER-2 is +1 (negative) and +3 (positive), while Ki-67 is low and high. The final results show that our proposed model outperforms all single models and as an ensemble result. Our contributions: (1) built the model using feature fusion that contains 768 ConvNextTiny features and one cell quantification feature from (Canny and Otsu); (2) used four biomarkers (ER, PR, HER, and Ki-67) as input for homogeneous and heterogeneous breast cancer classification.

The study presents a systematic approach to detecting four breast cancer biomarkers—ER, PR, HER-2, and Ki-67. Section 1 discusses the biomarkers and highlights the novelty of the research. Section 2 reviews previous studies and emphasizes the contributions made by this work. Section 3 discusses the methodology in detail, including dataset collection, integration, and the application of ensemble learning algorithms with feature fusion techniques to enhance performance. Section 4 demonstrates that feature fusion significantly improves model accuracy in detecting cell nuclei scores and brown intensity in image objects, leveraging feature extraction and combination in both models. Finally, Section 5 concludes that the proposed model meets the needs of ensemble learning while aligning with pathologists' practices, achieving superior performance in accuracy, precision, recall, F_1-score, and ROC-AUC through the combination of cell quantification features and ConvNexTiny features.

2 Related studies

Several previous researchers conducted studies focusing on investigating the status of ER, PR, HER-2, and Ki-67, as well as other biomarkers, using machine learning, deep learning, CNN model adaptation, and stained cell expression scoring.

Fan et al. (2024) presented an intelligent, holistic breast cancer tumor diagnosis system, including an interpretation module and a subtype module. The interpretation module is used to extract and analyze data based on a CNN-based convolutional neural network from HER-2, ER, PR, and Ki-67 images, followed by classification analysis. The subtype module produces holistic detection results of critical tumor markers with diagnostic suggestions for molecular subtypes validated by three pathologists. The model architecture consists of four convolution layers, four pooling layers, fully connected layers, and one output layer. The used dataset consists of 104 HER-2 cases, 198 ER and PR cases, and 60 Ki-67 cases.

Kildal et al. (2024) proposed a model using Mask R-CNN, YOLOv5, and deep learning to detect nuclear, cytoplasmic, and membranous IHC staining patterns in five image objects, namely colon, two prostate, breast, and endometrial. Image objects of the biomarker Ki-67 for colon, prostate, and breast cancer; PMS2 and MSH6 for colon and endometrial cancer; PTEN, CCNB1, CD44, Flotillin1, Mapre2, and β-catenin for prostate cancer; and ER and PR for breast cancer. The models consist of three, namely the nuclear model, the cytoplasmic model, and the membranous model. The nuclear model consists of 69 whole slide imaging (WSI) from the Ki-67 colon set and 23 WSIs from the PMS2-colon set; the cytoplasmic model consists of 34 WSIs from the PTEN-prostate set; and the membranous model consists of 25 WSIs from the β-catenin prostate. The image size is 800 × 800 pixels at 40× magnification, as the feature and the labels are "positive" and "negative" for each biomarker.

Zhao et al. (2024) developed a ResNet-18 model based on the framework and an online clinical application platform to predict molecular features and patient prognosis from triple-negative WSI pathology. The framework architecture consists of a serially working part to compare two separate convolutional networks (CNNs). The first is a tissue type classifier developed based on 20 WSIs' pixel-level tissue type annotations connected to the prediction target. The second is a CNN trained based on sample tiles for different targets. The models were trained and validated using the Fudan University Shanghai Cancer Center Triple Negative Breast Cancer (FUSCCTNBC) cohort through three-fold cross-validation. All three models were applied to the The Cancer Genome Atlas Triple Negative Breast Cancer (TCGATNBC) cohort. Each patient received three prediction scores, and the average was used for the final prediction. Performance metrics were then computed for external validation.

Tafavvoghi et al. (2024) performed two scenarios: *first*, classifying tiles in tumor and non-tumor areas for molecular subtypes using InceptionV3, the tile matrix size is 512 × 512, then decreased in size (1 × 1, 3 × 3, 5 × 5); *second*, using the One-vs-Rest (OvR) strategy to train four binary OvR classifiers and combining the results using the Xtreme Gradient Boosting model. The datasets accessed from The

Cancer Genome Atlas-Breast Cancer Gene (TCGA-BRCA) were 1,175, Breast Cancer Screening System (BRACS) were 129, Clinical Proteomic Tumor Analysis Consortium_Breast Invasive Carcinoma (CPTAC_BRCA) was 382, and HER-2-Warwick was 71.

Solorzano et al. (2024) built a single CNN model and an ensemble model from Inception V3, ResNet50, Inception-ResNet V2, and Xception to determine the presence of invasive carcinoma, IC or not IC. The datasets from Clinseq were 232 WSI and Sos 355 WSI, a total of 2,502,649 tissue tiles of size 598 × 598 pixels (271 × 271 μm) at 20× magnification. To determine the last decision, voting was applied to the model.

Bychkov et al. (2022) detected mitosis, nuclear pleomorphism, and tubule formation images using ResNet CNN. The biomarkers used were BCSS, ER, and ERBB2. The evaluation technique was applied to the model trained on the FinProg test set, which refers to the internal test set, and the FinHer patient series, which was not used for all training. The average output of the five trained models was used for cross-validation to reduce CNN variance and improve prediction accuracy. Validation between prediction scores (CNN output) and real-time sensor readings was conducted using statistical analysis based on Cox PH multivariate regression.

The advantages of our implemented model are as follows: (1) Accommodating conventional techniques that pathologists use to analyze the cell morphology. The quantification and intensification of the staining of the nucleus and membrane of cancer cells determine the expression of cancer malignancy. (2) Homogeneous ensemble learning improves model performance results during training and testing. The majority vote technique selects the best model among the four models, which is more efficient than the average bagging technique. (3) Heterogeneous ensemble learning, through feature fusion of concatenated different features, significantly improves model performance to be visually and medically representative. Feature engineering uses modified ConvNextTiny as a concatenation of convolution and cell quantification features. Computation time is shorter because it uses one ConvNextTiny classifier.

3 Methods

3.1 Datasets

We used datasets from Hasanuddin University Hospital (HUH) and Wahidin Sudirohusodo Hospital (WSH), consisting of 300 WSI: 200 from WSH, and 100 from HUH, which were then sampled into 23,351 samples. The dataset is closed access, and the owner's consent is required. The immunohistochemistry (IHC) biomarkers used for each patient consisted of estrogen receptor (ER), progesterone receptor (PR), HER-2, and Ki-67. However, the condition of biomarkers in the laboratory is not always complete, so data imbalance is an obstacle. The data composition of each biomarker consists of 9,035 ER image samples, and the total dataset used was 1,499 images. As shown in Table 1, the total dataset used was 23,331 images.

Because the image size during capture varies depending on the image cropping area and device resolution, we need to do data preparation. This step is crucial because, besides being thorough, it is also very time-consuming. Therefore, the input data are standardized to 224 × 224 and normalized for computing speed needs. In addition, determining the boundaries of cell morphology and membranes requires careful visual analysis to annotate each cell boundary and its cell membrane boundary. Not all WSIs are normal; some have blurriness during WSI pre-processing, so the cell morphology does not match the actual one.

TABLE 1 Datasets of ER, PR, HER-2, and Ki-67.

Images	Train	Validation	Test	Total
ER	4,790	598	506	8,894
PR	3,268	408	410	4,086
HER-2	5,309	663	665	6,637
Ki-67	5,370	671	673	6,714
Total	18.737	2,340	2,254	23,331

3.2 Pre-processing

We performed physical data acquisition and data preparation that met the required qualification standards. The techniques used included performing the cutting process when capturing the image of each biomarker sample (cropping). The rectangular image dimensions vary according to the ratio of the size and area of the WSI. For instance, an image measuring 1,280 × 613 pixels has a size of approximately 224 kb. WSI scans using KBIO, China Scanner, capture the entire cross-section of the WSI image area. Additionally, we read, enlarge, capture, and crop using Slideviewer 3DHISTECH, Budapest, Hungary, with magnifications of 5×, 10×, 20×, and 40×. The selection of these magnifications is based on the clarity of the colored cell objects, which adapt to the devices of the two hospitals. Physical data acquisition continues with virtual data acquisition if it is already in the programming framework. Acquisition is done by resizing the original image size into patches (224 × 224) to ensure uniform data size, facilitating the arithmetic and geometric operations of the input data.

Here, we focus specifically on the preprocessing of handcrafted features, as preprocessing for convolutional neural networks is handled automatically without manual intervention. The preprocessing techniques applied include Gaussian Blur, Otsu thresholding, and mask inversion. An 11 × 11 Gaussian blur is used to blur the image, remove noise in the image, and make the transition between areas smoother. At the same time, Otsu's method is an automatic technique for determining the optimal threshold to divide grayscale images into two classes: foreground and background. ConvNextTiny was pre-trained using ImageNet weights as transfer learning weights. All handcrafted features use feature standardization for data uniformity.

3.3 Processing

Processing and performing modeling based on the methods used in the feature extractor and classifier.

3.3.1 Feature engineering

Feature engineering consists of two main types: cell quantification feature (handcrafted feature) and convolution feature. The cell quantification feature extracts using Canny edge detection and the Otsu thresholding technique. The Canny edge detection

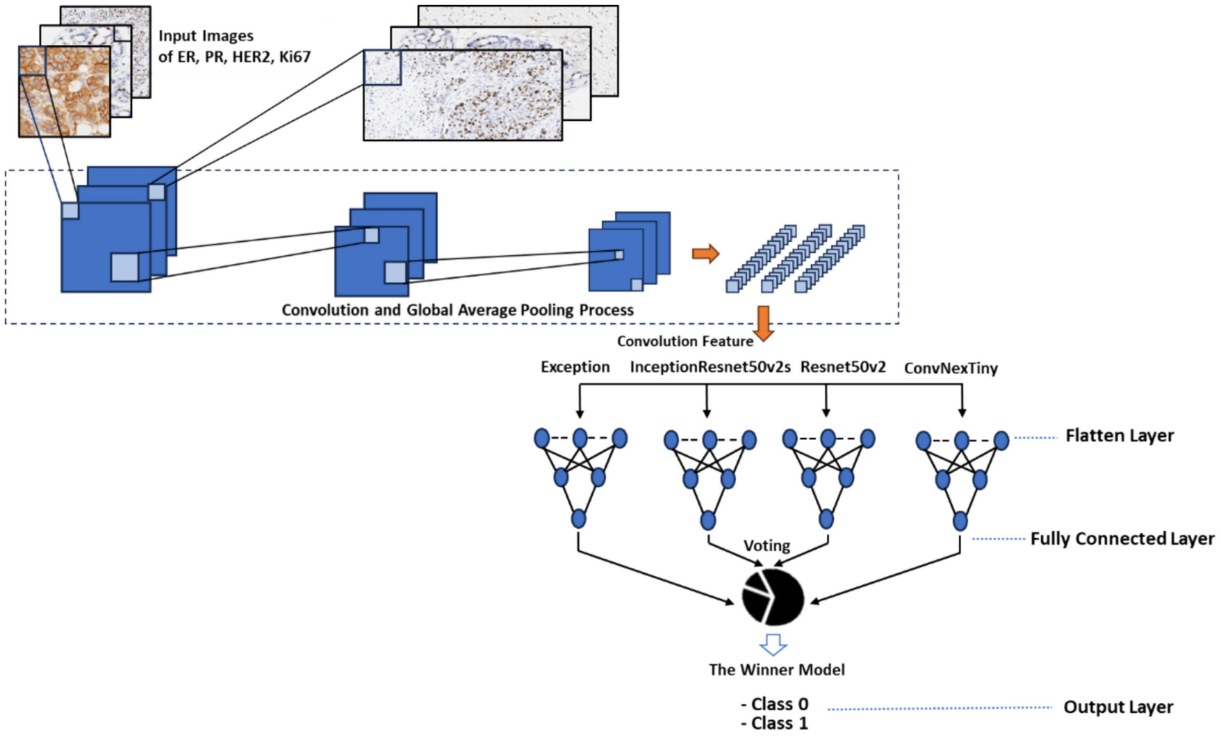

FIGURE 1
Homogeneous ensemble learning.

(Intan et al., 2023) sharpens the edges of cancer cells and colored cell areas while counting the number of cells (Figure 1). Otsu edge thresholding (Chadha et al., 2020) provides a circle boundary on the cell membrane object in a clear circle area through feature extraction (Figure 2) (Aswathy and Jagannath, 2017)—next, automatic convolution feature extraction through the ConvNexTiny model in each of its layers. The third stage of feature extraction is the scored feature. The results of the feature extraction stage are used to extract the number of cells based on their color indications, and then scoring (quantification) is carried out.

In handcrafted image processing, Canny focuses on changes in sharpness between pixels (high gradient) through the following steps: (1) Gaussian blur to reduce noise; (2) Gradient magnitude to detect changes in intensity; (3) Non-maximum suppression to produce thin edges; (4) Hysteresis thresholding to filter strong and weak edges based on pixel intensity; (5) Calculate the average value and standard deviation of the number of cell nuclei. Unlike Canny, Otsu focuses on area segmentation based on pixel intensity. However, Otsu's nature is used to separate between pixel conditions greater than the threshold, where the value is 0, and otherwise, the value is 1. Canny performs edge detection using the hysteresis limit of the image pixel strength. If the pixel is smaller than 30, it is ignored, and if it is above 105, the label is marked as an edge.

The second handcrafted technique uses the Otsu threshold with the following steps: (1) calculate pixel intensity using a binary threshold (8 bits, 0–255), intensity 0–50 is 0, while 100 is 1; (2) calculate the probability of each image intensity (Equation 1); (3) iterate for object and background classes; and (4) calculate the mean and standard deviation of the brown cell membrane area in the HER-2 biomarker.

$$P(i) = \frac{\text{Number of Selected Intensity Pixels}}{\text{Total of Pixels}} \quad (1)$$

3.3.2 Classifier

Our experimental setup involved two ensemble learning models: homogeneous and heterogeneous. For the homogeneous ensemble, we employed a majority voting technique among four distinct models (Xception, ResNet50V2, InceptionResNet50V2, and ConvNextTiny) to determine the best overall performer. Through this process, ConvNextTiny consistently demonstrated superior performance across various metrics, indicating its strength as an individual classifier. Therefore, for the heterogeneous ensemble learning model, ConvNextTiny was selected as the foundational architecture due to its dominant performance within the homogeneous ensemble. This allowed us to integrate additional handcrafted cell quantification features with ConvNextTiny's convolutional features, creating a robust fusion model.

3.3.2.1 Homogeneous ensemble learning

3.3.2.1.1 Model

The classifier used consists of four models, as follows (Figures 1, 2):

First, Xception (Li et al., 2023) has a basic CNN architecture; convolution is used to process the main features against the deep convolution separately from the CNN convolution to achieve feature extraction and successful computation, reducing the number of parameters (Sharma and Kumar, 2022; Krishna et al., 2023). *Second*, Resnet50V2 (He et al., 2016; Rahimzadeh and Attar, 2020), Residual

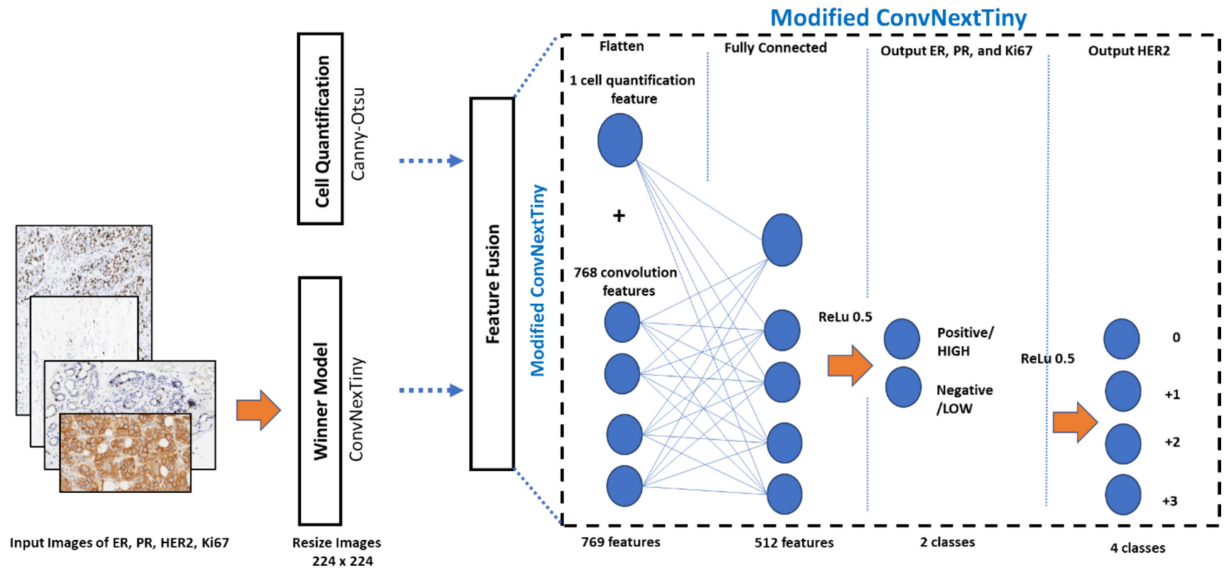

FIGURE 2
Heterogeneous ensemble learning.

Network 50V2, an architecture that has a stack block with the same connection shape (Residual 3 Unit). This base model has the following advantages: (1) ease of optimization and (2) reducing overfitting, unnormalized signals used as input to the next layer, so that all inputs are normalized. *Third*, InceptionNetResnetV2S (Asif et al., 2022; Talukder et al., 2023), each block is followed by an expansion filter layer used to increase the filter bank dimension before augmentation, according to the input thickness. *Fourth*, ConvNexTiny (Tanvir et al., 2024) is a novel convolutional neural network architecture that leverages standard CNN modules and incorporates optimization techniques inspired by the transformer model. ConvNexTiny has a network structure that shows great development potential through comprehensive experimental demonstrations covering macro and micro designs based on ResNet. This model outperforms the Swin Transformer while maintaining the simplicity and efficiency characteristics of standard CNN architectures (Yang et al., 2022).

In the homogeneous model ensemble (Figure 1), first, all models are applied individually to obtain the best weight hyperparameter tuning results. Starting from Exception, Resnet50v2, InceptionResnet50V2, and ConvNextTiny are based on convolution. We perform the ensemble by majority voting among the four models; the results will be the output of the homogeneous ensemble learning model.

Majority voting is a technique used to determine the final decision based on the largest number of labels in the entire model (Equation 2). Suppose there are $M = 4$ classification models, K possible classes; the m-th model gives a prediction. $\hat{y}$ is the class label with the most votes, and y is an indicator function (1 if true, 0 otherwise).

$$\hat{y} = \arg\max \sum_{m=1}^{M} II\left(y^{(m)} = c_j\right) \quad (2)$$

Each image will be plotted into an image grid in pixel form using "imagenet" weights with a learning rate of 10^{-3} and 10^{-4} as its training tuning. The image goes through a feature extraction process, starting from resizing, convolution, and global average pooling 2D to produce 1,024 features in the Exception, Resnet50v2, and InceptionResnet50V2 models and 768 features in ConvNextTiny. This feature is input for a neural network that uses the ReLu activation function and the "softmax" optimizer. The three models have 1,024 features, while ConvNextTiny has 768 features. Those four feature blocks are input to the neural network so that the output produces 512 features and finally produces two classifications, "positive" and "negative," or low and high. Each single model has the same parameter structure initialization. Similarly, parameter tuning is carried out by taking several learning rate scenarios.

The results of turning parameters with a learning rate of 10^{-3} then become weights for pre-trained. The results of these weights become the initialization when doing the second training to obtain the best weights. Pre-trained has a learning rate of 10^{-4} to load parameters with a learning rate of 10^{-3}, down from before, to obtain a smaller gradient descent, so that the loss decreases and the accuracy improves during training (Equations 6, 7).

3.3.2.2 Heterogeneous ensemble learning model

Medical record data images in ER, PR, Ki-67, and HER-2 images were taken from the examination results released by the pathologist. The pathologist selected the threshold for cell quantification (nucleus and cell membrane) based on WHO standards. The model determines the classification of ER, PR, and Ki-67, focusing on the number of dark brown cell nuclei, while HER-2 focuses on the area of brown cell membranes. The ER and PR use a threshold of 1%. The "negative" class has several cells ≤1% in stained cells; conversely, if the number of cells is>1%, then the class is "positive." In HER-2, it does not count the number of cells but computes the intensity of the brown color in the image. The HER-2 threshold is at 10%; if the intensity of the dark brown color is greater, then it is "positive," and if not, then it is "negative." Unlike the three previous biomarkers, Ki-67 has a higher threshold of 20%. The classification results are the labeling of the

images loaded into the model, and also serve as learning data for the model to recognize classification characteristic patterns: positive, negative, high, and low.

A heterogeneous ensemble model is a concatenation of homogeneous ensemble learning and cell quantification. Cell quantification is handcrafted from Canny edge detection and Otsu segmentation. Cell quantification, as a Canny edge detection model, counts the number of stained cells and Otsu counts segmentation of the area of dark brown cell membranes. Canny (Intan et al., 2023) cell quantification is an edge detection technique to reduce noise, preventing fake edge detection. The image $I(x, y)$ is filtered with a Gaussian Kernel to produce a convolution image, $I_S(x, y)$, as shown in Equations 3, 4.

$$G(x,y) = \frac{1}{2\pi\sigma^2}.e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

$$I_s(x,y) = I(x,y) * G(x,y) \tag{4}$$

To determine the feature map of edge boundaries using binary thresholding, 0 and 1. If $T_{\min}$ then 0 (not an edge boundary) and $T_{\max}$ then 1 (edge boundary) (Equation 5). The morphological kernel is rectangular 1 × 1 and produces the number of contours as the number of cells for classification.

$$g(x,y) = \begin{cases} 1 & jika\ T_{\min} \le f(x,y) \le T_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Otsu determines the binary threshold using a 3 × 3 elliptical kernel to segment the brown cell area. The cell quantification feature (Table 2), based on average values and standard deviations, indicates that ER, PR, and Ki-67 show average cell nucleus counts and standard deviations in the training data. These range from 1,600 to 2,400 and 1,250 to 1,650, respectively. On the other hand, HER-2 has an average area percentage of 0.32886 with a standard deviation of 0.17146. These average values are the cell quantification features used to train the model and determine the best final parameters. They are also key parameters for classifying conditions as "positive" or "negative" and "high" or "low."

The cell quantification model computes the number of cells resulting from extracting Canny features, as explained in Intan et al. (2023). It then computes the average number of cells in the training data and its standard deviation value. The image size, a sample patch for each image, is 224 × 224. Patches are partitions of each image into smaller square sizes as two-dimensional images that will be converted into n-dimensional features according to the layer's dimensions. This number of cells only has one feature to be input to the ConvNextTiny classifier. The layer structure of ConvNextTiny consists of four-layer blocks: the first block has 96 features; the second block has 192 features; the third block has 384 features; and the last block has 768 features and a ReLu activation function. The last block consists of 768 connecting features combined with one cell quantification feature to produce 769 convolution features, also called feature fusion. The process results in a modified ConvNextTiny. Feature fusion will simplify the features of both techniques that were initially separate, aiming to simplify the feature layer while improving the performance of the convolution model from 768 features. The 769 features are input to the neural network to be passed through the RELU activation function, and then 512 features are produced at its dropout (0.5), which are classified into two classes. The ER, PR, and HER-2 produce "negative" (0 and +1) and "positive" (+2 and +3) classes, while Ki-67 produces "low" and "high" classes.

TABLE 2 Cell quantification features.

	ER	PR	Ki-67	HER-2
Mean	1610.3472	2343.6925	2048.1196	0.32886
Deviation standard	1454.9679	1610.3472	1252.7431	0.17146

Figure 3 is a general concatenation of ensemble learning. The result of homogeneous ensemble learning in the form of the best model is ConvNextTiny, then the features of the ConvNextTiny head layer consist of 768 convolution features. Cell quantification is a manual feature extraction (handcrafted feature extraction) consisting of one Canny feature and one Otsu feature, each combined into the ConvNextTiny head to form modified ConvNextTiny. Modified ConvNextTiny is the last classifier to determine the final detection and prediction.

3.3.3 Evaluation

Model performance is an indicator of the success of building a new model. A good model will have improved performance with its reference model. It requires parameter tuning to obtain better results if it does not improve. Model classification requires validation of its output; if the probability of correct validation is high, then it is confirmed to be a good model; conversely, if the probability is low, it will reduce model performance. It is also a factor in the feasibility of a model being implemented. This study model uses a confusion matrix (Wang et al., 2025; Chicco and Jurman, 2023), which is an element of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Some of the parameters measured include:

1 Loss: Computes the predicted and actual values of the model. y_i is the true label, p_i is the predicted probability for some samples N (Terven et al., 2025).

$$\log\text{loss} = \frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i) \tag{6}$$

2 Accuracy: Measures the proportion of correct data predictions to the overall model predictions (Wang et al., 2025; Chicco and Jurman, 2023). Higher accuracy values indicate better model performance, indicating that the majority of the data have correct classifications.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

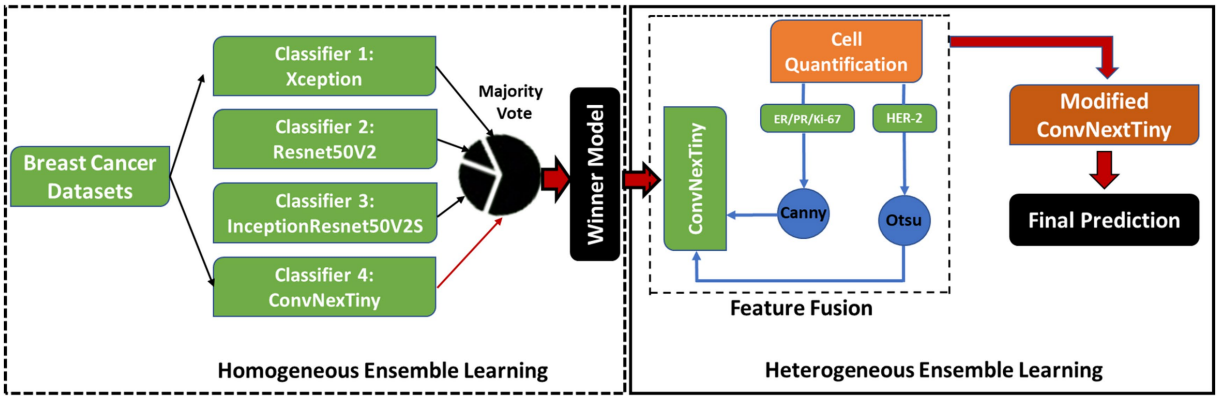

FIGURE 3
The concatenation of ensemble learning.

3 Precision: This metric assesses the proportion of true positive predictions to the total positive predictions (Wang et al., 2025; Chicco and Jurman, 2023) (Equation 8). A higher precision value indicates false positive errors of the minor data, focusing on the model's ability to classify samples accurately.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

4 Recall: This metric measures the proportion of true positive predictions from the total number of positive samples (Wang et al., 2025; Chicco and Jurman, 2023) (Equation 9). A higher recall value indicates fewer false negative errors, reflecting the model's ability to identify positive cases correctly.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

5 F_1-score: The harmonic mean of precision and recall measures the model's performance (Wang et al., 2025; Chicco and Jurman, 2023) (Equation 10). A higher F_1-score indicates better model performance, balancing precision and recall.

$$F_1 - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} * \text{Recall}} \tag{10}$$

6 ROC Curve is a graph that shows the performance of a binary classification model at various threshold values by plotting (Martínez Pérez and Pérez Martin, 2023; Carrington et al., 2023): (1) true positive rate (TPR) on the y-axis (also called sensitivity), and false positive rate (FPR) on the x-axis (which is 1—specificity) (Equations 11, 12).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{12}$$

4 Results and discussion

4.1 Homogeneous ensemble learning

A single model is the original model of each model element combined. Simulation of each model has different characteristics, including shorter computation speed, epoch, accuracy, and loss, which are parameters owned by each model.

In Tables 3–6, all models' performance and classification results show that ConvNexTiny has the highest accuracy and the smallest loss for all types of biomarkers, namely ER, PR, HER-2, and Ki-67. The model achieved the highest accuracy of 0.9933 and the smallest loss of 0.0078. These results indicate that the ConvNexTiny model wins the majority voting results on these metrics. Moreover, the classification results show that its valid data outperforms other models; even its invalid data has the smallest data, so it is very appropriate that ConvNexTiny is the best model among other single models.

Table 2 summarizes the performance of the ConvNexTiny model on the four types of images in detail. The experiments conducted on the training model used an initial learning rate of 10^{-3}, and then fine-tuning was performed at a learning rate of 10^{-4} until the model obtained optimum weights. These optimum weights are used as a reference for testing each biomarker. The lowest learning rate produces higher accuracy and lower loss due to the descent of errors through fine-tuning, which aims to reduce errors in gradient descent (Equation 6).

Table 4 shows the performance capabilities of the four models. ConvNexTiny outperforms the other three models, which can only recognize approximately 925 valid images, while ConvNexTiny can classify as many as 933 valid images and only four invalid images. The same thing is also shown in Table 5; ConvNexTiny outperforms the accuracy of Exception, Resnet50V2, and InceptionResnet50V2 with a value of 0.9974 and the smallest loss

TABLE 3 Model performance of ER and PR.

Models	ER				PR			
	Lr = 10^{-4}		Lr = 10^{-3}		Lr = 10^{-4}		Lr = 10^{-3}	
	Test performance							
	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss
Exception	0.985	0.0449	0.9658	0.1051	0.9707	0.0577	0.9512	0.1318
Resnet50V2	0.985	0.0538	0.9573	0.1317	0.9585	0.1133	0.9512	0.1096
InceptionResnet50V2	0.979	0.0479	0.9712	0.0973	0.9732	0.0887	0.9512	0.1369
ConvNextTiny	0.9933	0.0078	0.9916	0.0394	0.9800	0.05	0.9636	0.1105

TABLE 4 Data distribution of ensemble model classification of ER, PR, HER-2, and Ki-67.

Models	ER				PR			
	Lr = 10^{-4}		Lr = 10^{-3}		Lr = 10^{-4}		Lr = 10^{-3}	
	Numbers of data							
	Valid	Invalid	Valid	Invalid	Valid	Invalid	Invalid	Valid
Exception	925	12	905	32	401	9	390	20
Resnet50V2	925	12	897	40	399	11	390	20
InceptionResnet50V2	925	12	910	27	399	11	390	20
ConvNextTiny	933	4			403	7		

TABLE 5 Model performance of HER-2 and Ki-67.

Models	HER-2				Ki-67			
	Lr = 10^{-4}		Lr = 10^{-3}		Lr = 10^{-4}		Lr = 10^{-3}	
	Test performance							
	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss
Exception	0.9955	0.0185	0.9835	0.0631	0.985	0.0449	0.9658	0.1051
Resnet50V2	0.9925	0.0129	0.9865	0.0439	0.985	0.0538	0.9573	0.1317
InceptionResnet50V2	0.9925	0.0336	0.9835	0.0477	0.979	0.0479	0.9712	0.0973
ConvNextTiny	0.9964	0.0112	0.9960	0.0157	0.990	0.0419	0.9797	0.038

TABLE 6 Model classification of HER-2 and Ki-67.

Models	HER-2				Ki-67			
	Lr = 10^{-6}		Lr = 10^{-3}		Lr = 10^{-6}		Lr = 10^{-3}	
	Numbers of data							
	Valid	Invalid	Valid	Invalid	Valid	Invalid	Invalid	Valid
Exception	662	3	654	11	923	14	905	32
Resnet50V2	661	4	656	9	923	14	897	40
InceptionResnet50V2	555	10	654	11	918	19	910	27
ConvNextTiny	663	2			928	9		

value of 0.0112 on HER-2, as well as on Ki-67. The value has an impact on the data that is validated correctly (Table 6), with as many as 663 valid data and only two invalid data on HER-2. Similarly, on Ki-67, there are 928 valid data and nine invalid data. The more valid data depends on the higher accuracy and the lower loss. Conversely, if the accuracy is lower and the loss is higher, it will affect the number of correctly validated data. Good performance is obtained from the fine-tuning process to obtain the smallest error and high accuracy, even though the training time is longer.

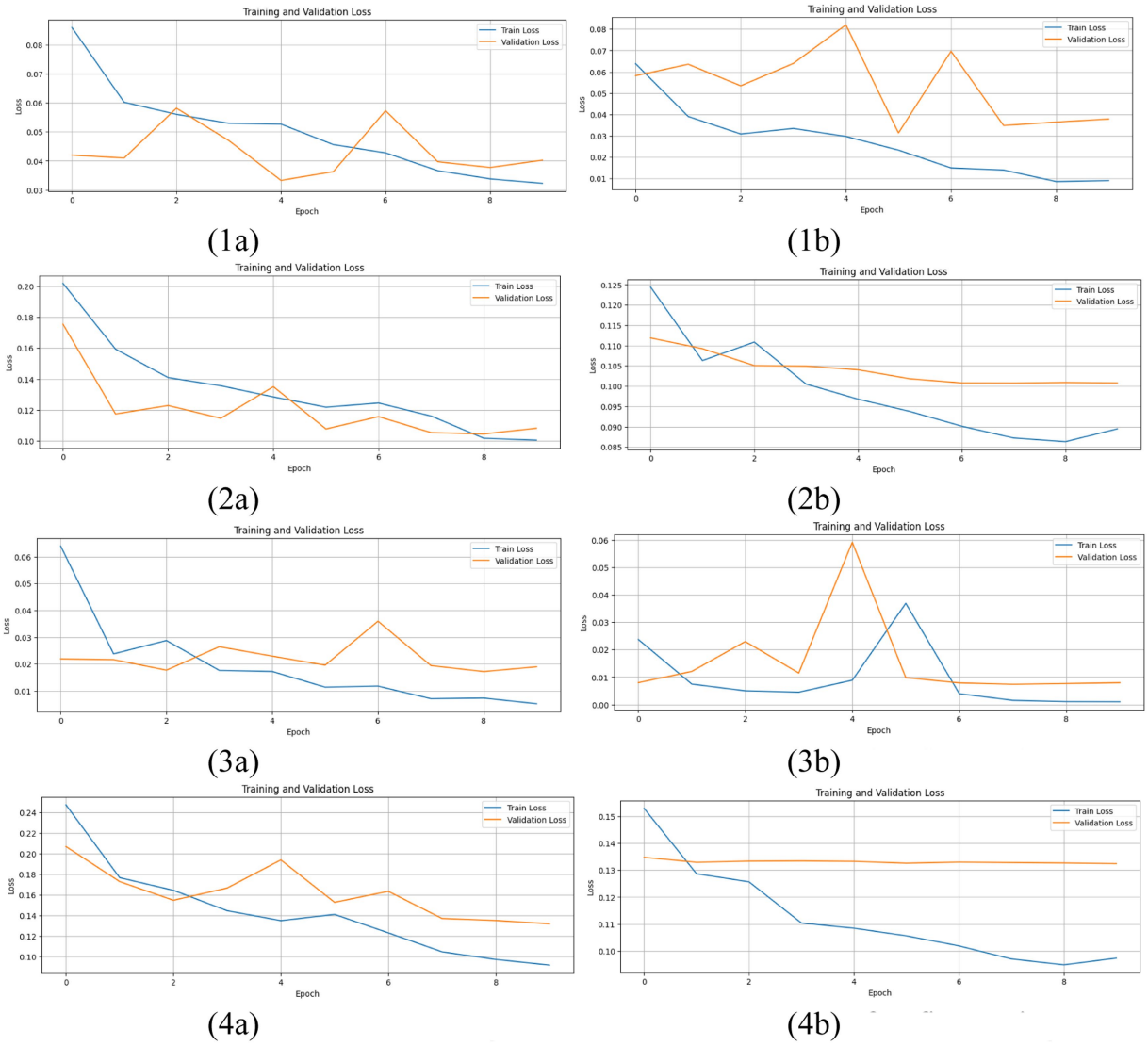

FIGURE 4
Results of train loss with fine-tuning at a heterogeneous ensemble model. (1a), (1b) ER; (2a), (2b) PR; (3a), (3b) HER-2; and (4a), (4b) Ki-67.

4.2 Heterogeneous ensemble learning

Figure 4 shows that the fine-tuning results provide better loss conditions. (1a), (2a), (3a), and (4a) show overlapping between train and validation at a learning rate of 10^{-3}, while (1b), (2b), (3b), and (4b) show that fine-tuning at a learning rate of 10^{-4} successfully separates the training curve and the validation curve so that overlapping is resolved. The error is getting smaller, indicating that the model's ability to distinguish between its two classification classes is improving during training and testing. To prove it, Table 7 shows the amount of data and the percentage of model classification on ER, PR, HER-2, and Ki-67.

Heterogeneous ensemble learning is an ensemble model between cell quantification and ConvNextTiny. Cell quantification uses grayscale, binary, blurred image, canny edge detection, dilated image, and contour techniques to determine the radius of the cell circle and compute the number of circles resulting from contours. This circle is a colored cell observed using a microscope display (Figure 5).

TABLE 7 Data distribution of ensemble model classification of ER, PR, HER-2, and Ki-67.

Images	Train	Val	Test	% Valid	% Invalid
ER	4,790	598	506	99.16	0.84
PR	3,268	408	410	97.06	2.94
HER-2	5,309	663	665	99.69	0.30
Ki-67	5,370	671	673	99.25	0.74

In addition to features learned via convolutional neural networks (CNNs), we incorporated a handcrafted feature to quantify cell density or stained area, depending on the biomarker type. We estimated cell quantification using Canny edge detection for ER, PR, and Ki-67 datasets. Images were first converted to grayscale and binarized using Otsu's thresholding in inverse mode to isolate foreground structures. A Gaussian blur (kernel size: 11×11) was applied to reduce noise, followed by Canny edge detection with thresholds of 30 and 105. The

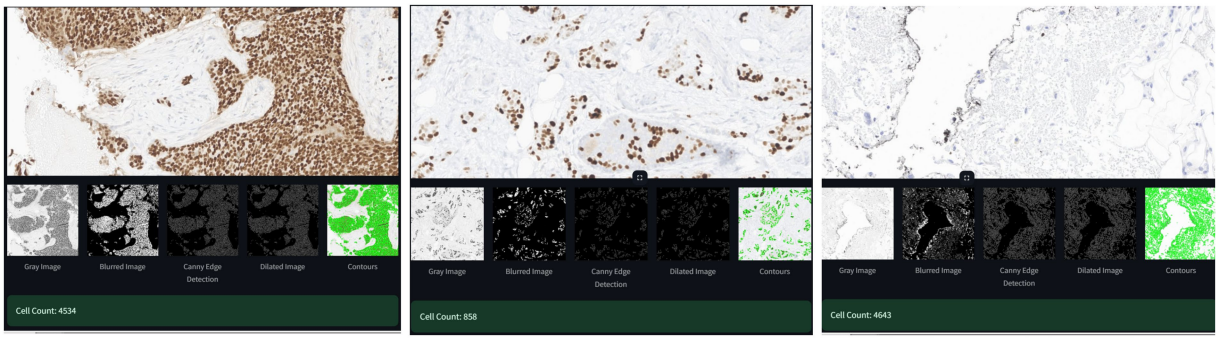

FIGURE 5
Cell quantification as results of feature extraction of Canny edge detection: ER (left), PR (center), and Ki-67 (right).

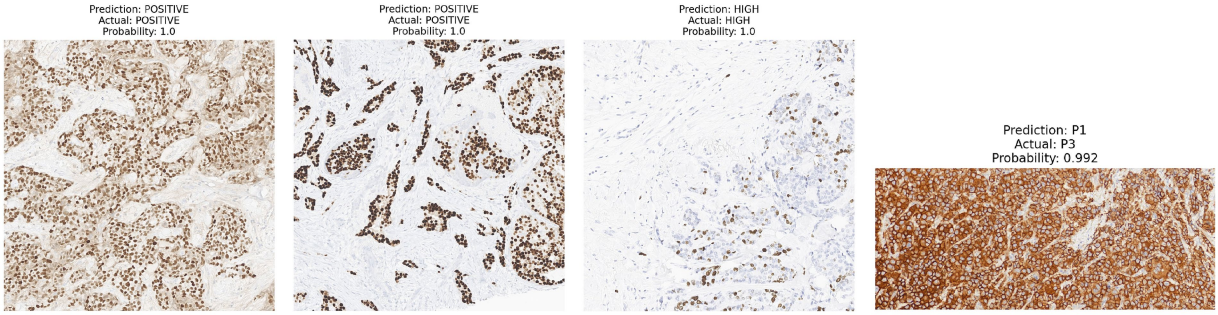

FIGURE 6
Model classification results with labels of ER, PR, Ki-67, and HER-2 (from left to right).

resulting edges were dilated using a 1 × 1 rectangular kernel for two iterations, and external contours were counted as a proxy for cell nuclei. These threshold values were selected empirically by evaluating multiple samples and identifying the settings that produced the most accurate and consistent contour quantification relative to visual inspection. The extraction of contour cells in ER, PR, and Ki-67 focuses on the brown nucleus circle of cells, totaling 4,534, 858, and 4,643, respectively (Figure 5).

For HER-2 images, we computed the proportion of stained (brown) regions to the total image area. Grayscale images were binarized using Otsu's method and inverted to highlight dark-stained regions. The morphological opening with an elliptical 3 × 3 kernel removed minor artifacts, and the stained ratio was computed as the fraction of non-zero pixels in the mask. All extracted features were standardized using the mean and standard deviation computed from the training set before being concatenated with CNN outputs for final classification. The model in Figure 6 does not compute the number of cells as circles, but it computes the brown cell area, indicating that the cell membrane is 0.45367 or 45.367%. This area is already within the threshold limit, the area of brown cells for HER-2 (Figure 7).

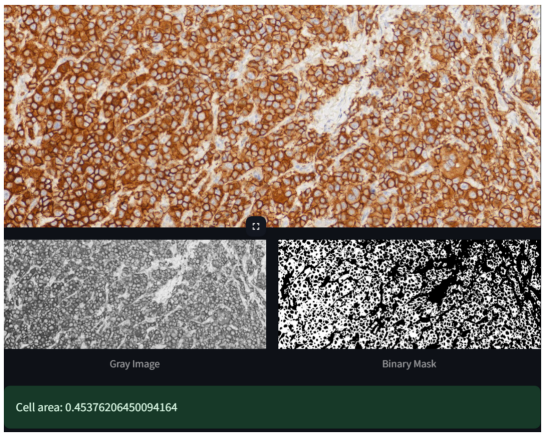

FIGURE 7
Cell quantifications as a result of feature extraction of Otsu edge detection (HER-2).

Table 8 shows that the results of the heterogeneous ensemble model have improved performance beyond the homogeneous ensemble model (Table 2). It shows that the ensemble technique improves model performance by adding one extraction feature (Canny and Otsu). Table 8 shows a fairly significant value of the heterogeneous ensemble model, exceeding the capabilities of the homogeneous ensemble model compared to the homogeneous values found in Tables 3–5. Technically, the homogeneous ensemble results of the single ConvNextTiny model show quite good performance, but assigning image labels to one technique has not accommodated the needs of pathologists based on their fields of science, which require observation of cell morphology. These observations are based on the

threshold of the number of cells and the brown area on the cell membrane. This is very important to provide confidence in determining the status characteristics of each sample or patch from WSI.

The heterogeneous ensemble model performs classification of ER, PR, Ki-67, and HER-2 with labeling and validation between the prediction and actual. The image shows that the characteristics of the number of cells will be identified based on the brown cycle, as the cell nucleus (the first three images from the left). In contrast, the rightmost image is the HER-2 image, marked as the distribution of cell membranes that dominate the patch area (square images), brown, and focuses on its area rather than on the nucleus. The model will visually compute the number of cell nuclei, which can be computed manually, although it takes an inefficient time. It is unlike the cell membrane, which is observed around the brown area and not the cycle contained in the patches.

The superiority of the heterogeneous ensemble model is also seen in the proportion of correctly recognized image samples. The higher the proportion, the higher the confidence in its classification performance, as shown in Table 8. Indeed, of course, it depends on the train's performance on all types of biomarkers, which greatly determine the performance of the test. Based on the values in Table 8, HER-2 outperforms the other three biomarkers because feature extraction of the brown cell area makes it easy to determine its presence with certainty. In comparison, the other three biomarkers, ER, PR, and Ki-67, still need to compute the number of cells, which is not always certain for each image sample. Automated cell quantification makes it more difficult for the model to classify it, even though it already has a threshold number. ER and PR have a threshold of 1%, while HER-2 and Ki-67 have value thresholds of 10 and 20%, respectively. This superiority of HER-2 is indeed unique, even though each type has the same number of image samples. However, HER-2 is still superior for the previous reason.

This study is in line with previous studies using other model ensembles, in that the output of the ensemble model will reinforce the model before the ensemble so that the existence of this technique is significant enough to present a model applied to public service access in the medical field. It is shown in previous studies, including in Table 7.

Figure 8 presents the metrics for evaluating the model using a confusion matrix, including accuracy, precision, recall, F_1-score, and ROC-AUC. The model achieved its highest accuracy of 0.99699 on HER-2, recall of 0.998022 on ER, F_1-score of 0.99505 on ER, ROC-AUC of 0.99426 on HER-2, and precision of 0.9921 on ER. HER-2s get superior accuracy when they focus on the cell membrane area, which is influenced by the portion of the brown area in the image. HER-2 also has the highest number of valid predictions. However, accurately counting cancer cells remains challenging due to

TABLE 8 Performance of a heterogeneous ensemble model of ER, PR, HER-2, and Ki-67.

Images	Acc	Precision	Rec	F_1-score	ROC-AUC
ER	0.99164	0.99211	0.99802	0.99505	0.97773
PR	0.98049	0.97436	0.97436	0.97436	0.97931
HER-2	0.99699	0.97345	0.99099	0.98214	0.99426
Ki-67	0.98662	0.97787	0.98222	0.98004	0.98553

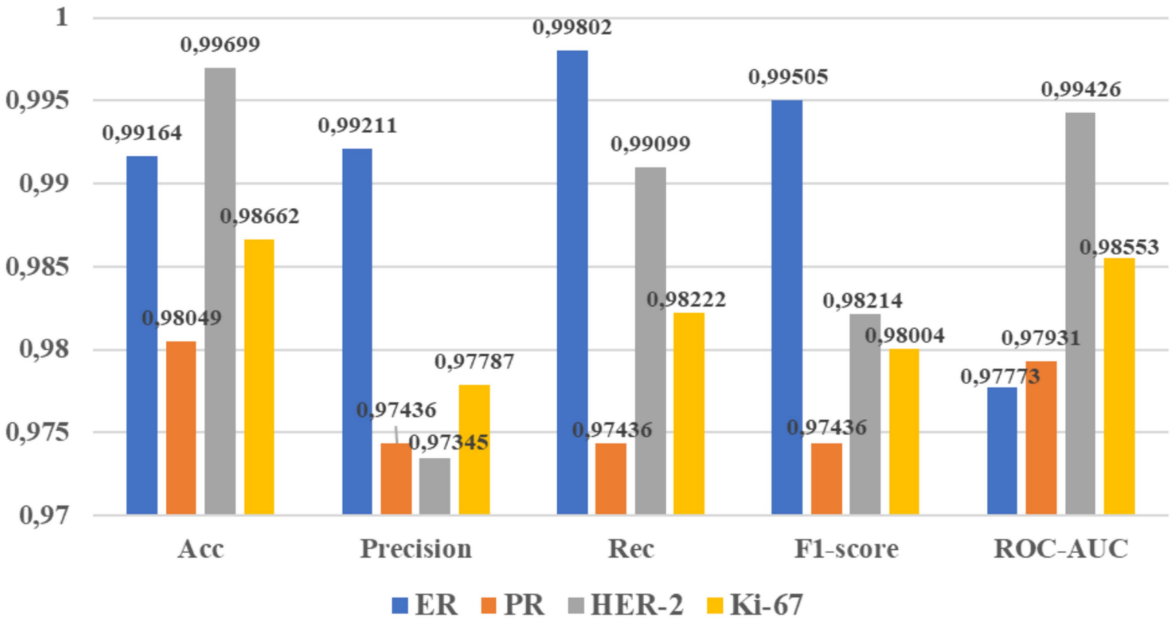

FIGURE 8
Results of the confusion matrix.

the uncertain distribution of these cells within the image. Despite Ki-67 having the most extensive dataset, in our case, it does not surpass HER-2 for prediction percentage. The model struggles to differentiate between light, regular, and dark brown shades, leading it to rely on edge detection results, which often produce invalid classifications. Additionally, the smallest dataset size for PR contributes to its lower prediction values.

4.3 Research achievement

We conducted multiple scenarios to enhance the model we developed. As expected, these results were also achieved by previous researchers employing different methodologies. Table 9 compares our research findings with those of previous studies.

The performance comparison involves various variables, including different data sources, datasets, feature fusion, ensemble techniques, and models (Table 7). We analyze results across smaller and larger data clusters relative to our dataset, including datasets with restricted access. Our model demonstrates superior performance compared to studies with smaller datasets (Mudeng et al., 2023; Zheng et al., 2023; Abdullakutty et al., 2024; Alam et al., 2024; Islam et al., 2024; Qasrawi et al., 2024; Solorzano et al., 2024; Sreelekshmi and Nair, 2024). It also outperforms those with larger datasets (Khan et al., 2023; Kumari and Ghosh, 2023; Prezja et al., 2024). However, research by Ahmad and Alqurashi (2024) and Rahaman et al. (2024) achieves better results using ensembles of older models. Our approach leverages a new model that integrates feature fusion from handcrafted with ConvNextTiny features in the head block, offering efficient computation and simple edge detection.

Here, our feature fusion model has an average performance above 99%, including previous models (Alam et al., 2024). These features of each class will find their unique patterns and then be combined with the convolution features in the transfer learning model, so that it will produce a unique pattern if only using the convolution model. Feature fusion has provided a unique pattern to the model that produces better performance than without feature fusion, even though homogeneous ensemble learning is carried out. Feature fusion improves model performance on adequate datasets, bagging techniques, and robust models. The selection of models and feature techniques greatly determines the final performance of the model, so that the classification ability becomes a reliable and final result.

This research is an ongoing process of interpreting the molecular expression of breast cancer patients in immunohistochemistry examination. The expression still requires the characteristics of all biomarkers to determine the molecular subtypes that play a role in determining the enforced prognosis and type of therapy. Research by Fan et al. (2024) becomes a reference for the development of this research in the future with various feature engineering and models to obtain robust performance and contribute to efficient architecture in its deployment.

4.4 Limitations

This research endeavors to enhance the model performance utilized by pathologists during immunohistochemistry examinations, which are concerned with the determination of cell morphology and cell proliferation, explicitly focusing on cell nuclei and membranes. This investigation leverages computer vision to ensure precise and accurate results, thereby facilitating reliable diagnostic outcomes. Nevertheless, various limitations must be acknowledged and addressed in this research context, including:

The fixation of the number of samples for each WSI does not represent the entire area because only three to four samples are taken for each WSI for each magnification. It will compute the number of cells in each WSI accurately and determine the average value and standard deviation of each WSI, not only based on each sample but also focusing on the mean and standard deviation of each WSI. It also overcomes overshooting and overlapping during training, which are still relatively high. It will certainly provide strong confidence in the results released by pathologists for hospitals.

The real-time model integration with the WSI display and scanner has not been configured, so the pointer changes or shifts can provide detection results for each WSI sample and magnification. The real-time configuration between the application and the microscope display hardware makes it easy to determine the final results of the examination, which are released visually.

5 Conclusion

We successfully implemented a heterogeneous ensemble learning model to address the problem of feature classification in anatomical pathology cases. Pathologists typically observe cell morphology to determine the malignant status of cancer, with cell quantification being the gold standard in their practice. This method computes the number of cells and the area of brown-stained cells to identify the cell nucleus and cell membrane. Our study selected the CNN types, ConvNextTiny model based on the majority voting results from four models experimented with cell quantification. Both ensembles demonstrated performance that significantly outperformed the other three single models. The result indicates that our approach, which concatenated a single feature into ConvNextTiny with the simplest structure, achieved performance superior to the more complex structures of the three models with a more significant number of features. Heterogeneous ensemble learning, which is a feature fusion, has significantly better performance on adequate datasets, ensemble techniques, and robust models, so its performance is better than that of homogeneous ensemble learning.

Further development of this study is necessary through more extensive patch exploration for each WSI image sample to achieve a more accurate and precise average. Each WSI sample has a fixed number of samples, and each WSI has its average and standard deviation identified to match the annotation results manually performed by pathologists. Additionally, the WSI shift results have been read, and the WSI status has been detected for each patch shift angle in real-time. This approach will undoubtedly produce an accurate analysis that matches the needs of pathologists and hospitals. The study aims to provide immunohistochemistry examination results that can effectively map the appropriate subtype of malignancy expression of breast cancer.

TABLE 9 Performance of the previous work.

Author	Dataset source	Number of data	Methods	Results
Mudeng et al. (2023)	BreakHis DataBiox	7,909 922	InceptionResNetV2; InceptionV3; NASNet-Large; ResNet50; ResNet101; VGG19; and Xception as single model Majority voting	Accuracy: 97.67% F_1-score: 97.60%
Zheng et al. (2023)	P&D Laboratory	7,909	VGG16 + Xception + ResNet50 + DenseNet201 Weighted voting strategy	Accuracy: 98.90%
Khan et al. (2023)	ITMP; University of Bern RUMC	53,814	U-Net + ViT Segmentation and classifier	F_1-score: 97.4% Sensitivity: 99.5% Specificity: 96.7%
Analysis: concatenate				
Kumari and Ghosh (2023)	IDC BreaKHis	277,524 7,909	VGG-16 + Xception + DenseNet201	Accuracy: 94. 2%
Sreelekshmi and Nair, 2024	MIAS	332	U-Net + Auto Encoder	Accuracy: 75.3%
Prezja et al. (2024)	NCT UMM	100,000	EfficientNet + Vision Transformer + Random Forest	Accuracy: 96.74%
Abdullakutty et al. (2024)	Electronic Medical Record	3,764	PCA + auto encoder; VGG-16; ViT; and ResNet-50	Accuracy: 78.84%
Parshionikar and Bhattacharyya (2024)	BreakHis IR Thermal Image Dataset	9,713 1,279	Inception + CapsNet	Accuracy: 99.74%
Rahaman et al. (2024)	In-House	12,156	EfficientnetB3 + ResNet50 + SCL	Accuracy: 99.92% Precision: 99.88% Recall: 99.90% F_1-score: 99.89%
Ahmad and Alqurashi (2024)	American Oncology Institute at Shrimann Hospital	1,935	ResNet50 + InceptionV3	Accuracy: 99.80% F_1-score: 99% Sensitivity: 99% Specificity: 99%
Solorzano et al. (2024)	Clinseq Sos	355 284; total 2, 502,649 tiles	Majority vote of Inc. Xception V3, ResNet50, Inception-ResNet V2 and Xception	Accuracy: 91.2% Dice: 86.2% Specificity: 85.9% Precision: 83.7%
Qasrawi et al. (2024)	HMUH	20,000	YOLO; VGG-16, DenseNet121	Accuracy: 88.9% Precision: 88.9% Recall: 88.7% F_1-score: 88.8% AUC: 89.4%
Karuppasamy et al. (2024)	SQUH BreaskHis	158 7,909	AlexNet + VggNet	AUC: 95%
Islam et al. (2024)	BUSI UDAIT	780 163	MobilleNet + Xception	Accuracy: 87.82% Precision: 87.33% Recall: 85.33% F_1-score: 86.00%
Alam et al. (2024)	BUSI	1,312	GAN + SVM + U-Net + VGG-19	Accuracy: 99.48% Sensitivity: 99.40% Specificity: 99.55%
Proposed model	Hasanuddin University Hospital Wahidin Sudirohusodo Hospital	23,154	Majority voting: Exception, ResNet50V2; InceptionResnet50V2; ConvNextTiny Ensemble (feature fusion): Canny/Otsu + ConvNextTiny for every Biomarker	Accuracy: 99.7% Precision: 97.35% Recall: 99.1% F_1-score: 98.21% ROC-AUC: 99.43%

Data availability statement

Datasets from Hasanuddin University Hospital, Dr. Wahidin Sudirohusodo Hospital, and https://www.kaggle.com/datasets/akbarnejad1991/ihc4bc-compressed/ were used for the study. The datasets from the two hospitals are private, while the third dataset is publicly available.

Ethics statement

The studies involving humans were approved by Prof. Dr. Muh Nasrum Massi, Ph.D., Sp.MK(K)Komite Etik Penelitian Kesehatan RSPTN Universitas Hasanuddin RSUP Dr. Wahidin Sudirohusodo Makassar. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

II: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. AK: Formal analysis, Project administration, Software, Visualization, Writing – review & editing. SH: Data curation, Formal analysis, Investigation, Project administration, Visualization, Writing – review & editing. BN: Conceptualization, Data curation, Formal analysis, Project administration, Supervision, Writing – review & editing. DS: Conceptualization, Formal analysis, Software, Validation, Writing – review & editing. AY: Data curation, Formal analysis, Supervision, Project administration, Writing – review & editing. RP: Data curation, Formal analysis, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Key Research and Development Project, Ministry of Higher Education, Science and Technology of the Republic of Indonesia (111/E5/PG.02.00.PL/2024), Special Funds at the Fundamental Research Scheme to Dipa Makassar University (103/UNDIPA/G.4/VI/2024).

Acknowledgments

The authors would like to thank the Ministry of Higher Education, Science and Technology of the Republic of Indonesia for funding, the Rector and Head of the Research and Development Agency of Dipa Makassar University for their support in this research until it could be published, and also to Hasanuddin University Hospital and Dr. Wahidin Sudirohusodo Hospital for their data support and expert consultation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2025.1569017/full#supplementary-material

References

Abdullakutty, F., Akbari, Y., Al-Maadeed, S., Bouridane, A., Talaat, I. M., and Hamoudi, R. (2024). Towards improved breast cancer detection via multi-modal fusion and dimensionality adjustment. *Comput. Struct. Biotechnol. J.* 1:100019. doi: 10.1016/j.csbr.2024.100019

Afkari, H., Makrufardi, F., Hidayat, B., Budiawan, H., and Sundawa Kartamihardja, A. H. (2021). Correlation between ER, PR, HER-2, and Ki-67 with the risk of bone metastases detected by bone scintigraphy in breast cancer patients: a cross sectional study. *Ann. Med. Surg.* 67:102532. doi: 10.1016/j.amsu.2021.102532

Ahmad, I., and Alqurashi, F. (2024). Early cancer detection using deep learning and medical imaging: a survey. *Crit. Rev. Oncol. Hematol.* 204:104528. doi: 10.1016/j.critrevonc.2024.104528

Alam, M. N. A., Mohi Uddin, K. M., Rahman, M. M., Manu, M. M. R., and Nasir, M. K. (2024). A novel automated system to detect breast cancer from ultrasound images using deep fused features with super resolution. *Intell. Based Med.* 10:100149. doi: 10.1016/j.ibmed.2024.100149

Alismail, H. (2024). Review: merging from traditional to potential novel breast cancer biomarkers. *J. King Saud Univ. Sci.* 36:103551. doi: 10.1016/j.jksus.2024.103551

Allison, K. H., Hammond, M. E. H., Dowsett, M., McKernin, S. E., Carey, L. A., Fitzgibbons, P. L., et al. (2021). Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update special articles abstract. *J. Clin. Oncol.* 38, 1346–1366. doi: 10.1200/JCO.19.02309

Asif, S., Yi, W., Ain, Q. U., Hou, J., Yi, T., and Si, J. (2022). Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from MR images. *IEEE Access* 10, 34716–34730. doi: 10.1109/ACCESS.2022.3153306

Aswathy, M. A., and Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: present status and future possibilities. *Inform. Med. Unlocked* 8, 74–79. doi: 10.1016/j.imu.2016.11.001

Bychkov, D., Joensuu, H., Nordling, S., Tiulpin, A., Kücükel, H., Lundin, M., et al. (2022). Outcome and biomarker supervised deep learning for survival prediction in two multicenter breast cancer series. *J. Pathol. Inform.* 13:100171. doi: 10.4103/jpi.jpi_29_21

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., et al (2023). Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 329–341. doi: 10.1109/TPAMI.2022.3145392

Chadha, G. K., Srivastava, A., Singh, A., Gupta, R., and Singla, D. (2020). An automated method for counting red blood cells using image processing. *Procedia Comput. Sci.* 167, 769–778. doi: 10.1016/j.procs.2020.03.408

Chen, H., Gui, X., Zhou, Z., Su, F., Gong, C., Li, S., et al. (2024). Distinct ER and PR expression patterns significantly affect the clinical outcomes of early HER2-positive breast cancer: a real-world analysis of 871 patients treated with neoadjuvant therapy. *Breast* 75:103733. doi: 10.1016/j.breast.2024.103733

Chicco, D., and Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16, 1–23. doi: 10.1186/s13040-023-00322-4

Fan, L., Liu, J., Ju, B., Lou, D., and Tian, Y. (2024). A deep learning based holistic diagnosis system for immunohistochemistry interpretation and molecular subtyping. *Neoplasia* 50:100976. doi: 10.1016/j.neo.2024.100976

Fei, F., Siegal, G. P., and Wei, S. (2021). Characterization of estrogen receptor-low-positive breast cancer. *Breast Cancer Res. Treat.* 188, 225–235. doi: 10.1007/s10549-021-06148-0

Grimm, S. L., Hartig, S. M., and Edwards, D. P. (2016). Progesterone receptor signaling mechanisms. *J. Mol. Biol.* 428, 3831–3849. doi: 10.1016/j.jmb.2016.06.020

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks, European Conference on Computer Vision. Springer: Cham. 630–645.

Intan, I., Nelwan, B. J., Henry, M. M., Karnyoto, A. S., Puspitasari, R. E., and Pardamean, B. (2024). Deep ensemble transfer learning for detecting breast cancer in histopathological images. *Commun. Math. Biol. Neurosci.* 2024:101. doi: 10.28919/cmbn/8823

Intan, I., Nurdin, and Pangerang, F. (2023). Facial recognition using multi edge detection and distance measure. *IAES Int. J. Artif. Intell.* 12, 1330–1342. doi: 10.11591/ijai.v12.i3.pp1330-1342

Islam, M. R., Rahman, M. M., Ali, M. S., Nafi, A. A. N., Alam, M. S., Godder, T. K., et al. (2024). Enhancing breast cancer segmentation and classification: an ensemble deep convolutional neural network and U-Net approach on ultrasound images. *Mach. Learn. Appl.* 16:100555. doi: 10.1016/j.mlwa.2024.100555

Joensuu, K., Leidenius, M., Kero, M., Andersson, L. C., Horwitz, K. B., and Heikkilä, P. (2013). ER, PR, HER2, Ki-67 and CK5 in early and late relapsing breast cancer-reduced CK5 expression in metastases. *Breast Cancer* 7, 23–34. doi: 10.4137/BCBCR.S10701

Karuppasamy, A. D., Abdesselam, A., zidoum, H., Hedjam, R., and al-Bahri, M. (2024). Combining a forward supervised filter learning with a sparse NMF for breast cancer histopathological image classification. *Intell. Based Med.* 10:100174. doi: 10.1016/j.ibmed.2024.100174

Kemenkes. (2024). Kanker Payudara, Kemenkes. Available online at: https://ayosehat.kemkes.go.id/topik-penyakit/neoplasma/kanker-payudara. (Accessed December 22, 2024)

Khan, A., Brouwer, N., Blank, A., Müller, F., Soldini, D., Noske, A., et al. (2023). Computer-assisted diagnosis of lymph node metastases in colorectal cancers using transfer learning with an ensemble model. *Mod. Pathol.* 36:100118. doi: 10.1016/j.modpat.2023.100118

Kildal, W., Cyll, K., Kalsnes, J., Islam, R., Julbø, F. M., Pradhan, M., et al. (2024). Deep learning for automated scoring of immunohistochemically stained tumour tissue sections—validation across tumour types based on patient outcomes. *Heliyon* 10:e32529. doi: 10.1016/j.heliyon.2024.e32529

Krishna, S., Suganthi, S. S., Bhavsar, A., Yesodharan, J., and Krishnamoorthy, S. (2023). An interpretable decision-support model for breast cancer diagnosis using histopathology images. *J. Pathol. Inform.* 14:100319. doi: 10.1016/j.jpi.2023.100319

Kumari, V., and Ghosh, R. (2023). A magnification-independent method for breast cancer classification using transfer learning. *Healthc. Anal.* 3:100207. doi: 10.1016/j.health.2023.100207

Li, L., Yang, Z., Yang, X., Li, J., and Zhou, Q. (2023). PV resource evaluation based on Xception and VGG19 two-layer network algorithm. *Heliyon* 9:e21450. doi: 10.1016/j.heliyon.2023.e21450

Loggie, J., Barnes, P. J., Carter, M. D., Rayson, D., and Bethune, G. C. (2024). Is Oncotype DX testing informative for breast cancers with low ER expression? A retrospective review from a biomarker testing referral center. *Breast* 75:103715. doi: 10.1016/j.breast.2024.103715

Lv, Q., Meng, Z., Yu, Y., Jiang, F., Guan, D., Liang, C., et al. (2016). Molecular mechanisms and translational therapies for human epidermal receptor 2 positive breast cancer. *Int. J. Mol. Sci.* 17:2095. doi: 10.3390/ijms17122095

Martínez Pérez, J. A., and Pérez Martin, P. S. (2023). La curva ROC. *Medicina de Familia. SEMERGEN*, 49:101821. doi: 10.1016/j.semerg.2022.101821

Mudeng, V., Farid, M. N., Ayana, G., and Choe, S. W. (2023). Domain and histopathology adaptations–based classification for malignancy grading system. *Am. J. Pathol.* 193, 2080–2098. doi: 10.1016/j.ajpath.2023.07.007

Parshionikar, S., and Bhattacharyya, D. (2024). An enhanced multi-scale deep convolutional orchard capsule neural network for multi-modal breast cancer detection. *Healthc. Anal.* 5:100298. doi: 10.1016/j.health.2023.100298

Penault-Llorca, F., and Radosevic-Robin, N. (2017). Ki67 assessment in breast cancer: an update. *Pathology* 49, 166–171. doi: 10.1016/j.pathol.2016.11.006

Prezja, F., Annala, L., Kiiskinen, S., Lahtinen, S., Ojala, T., Ruusuvuori, P., et al. (2024). Improving performance in colorectal cancer histology decomposition using deep and ensemble machine learning. *Heliyon* 10:e37561. doi: 10.1016/j.heliyon.2024.e37561

Qasrawi, R., Daraghmeh, O., Qdaih, I., Thwib, S., Vicuna Polo, S., Owienah, H., et al. (2024). Hybrid ensemble deep learning model for advancing breast cancer detection and classification in clinical applications. *Heliyon* 10:e38374. doi: 10.1016/j.heliyon.2024.e38374

Rahaman, M. M., Millar, E. K. A., and Meijering, E. (2024). Generalized deep learning for histopathology image classification using supervised contrastive learning. *J. Adv. Res.* 1–16. doi: 10.1016/j.jare.2024.11.013

Rahimzadeh, M., and Attar, A. (2020). A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. *Inform. Med. Unlocked* 19:100360. doi: 10.1016/j.imu.2020.100360

Reinert, T., Cascelli, F., Resende, C. A. A., Gonçalves, A. C., Godo, V. S. P., and Barrios, C. H. (2022). Clinical implication of low estrogen receptor (ER-low) expression in breast cancer. *Front. Endocrinol.* 13:1015388. doi: 10.3389/fendo.2022.1015388

Shao, Y., Guan, H., Luo, Z., Yu, Y., He, Y., Chen, Q., et al. (2024). Clinicopathological characteristics and value of HER2-low expression evolution in breast cancer receiving neoadjuvant chemotherapy. *Breast* 73:103666. doi: 10.1016/j.breast.2023.103666

Sharma, S., and Kumar, S. (2022). The Xception model: a potential feature extractor in breast cancer histology images classification. *ICT Express* 8, 101–108. doi: 10.1016/j.icte.2021.11.010

Solorzano, L., Robertson, S., Acs, B., Hartman, J., and Rantalainen, M. (2024). Ensemble-based deep learning improves detection of invasive breast cancer in routine histopathology images. *Heliyon* 10:e32892. doi: 10.1016/j.heliyon.2024.e32892

Sreelekshmi, V., and Nair, J. J. (2024). Variational auto encoders for improved breast cancer classification. *Procedia Comput. Sci.* 233, 801–811. doi: 10.1016/j.procs.2024.03.269

Tafavvoghi, M., Sildnes, A., Rakaee, M., Shvetsov, N., Bongo, L. A., Busund, L. R., et al. (2024). Deep learning-based classification of breast cancer molecular subtypes from H & E whole-slide images. *J. Pathol. Inform.* 16:100410. doi: 10.1016/j.jpi.2024.100410

Talukder, M. A., Islam, M. M., Uddin, M. A., Akhter, A., Pramanik, M. A. J., Aryal, S., et al. (2023). An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. *Expert Syst. Appl.* 230:120534. doi: 10.1016/j.eswa.2023.120534

Tanvir, J., Mehedi, S. T., Paul, B. K., and Morshed, M. (2024). TrashNeXt: classification of recyclable water pollutants using deep transfer learning method. *Case Stud. Chem. Environ. Eng.* 11:101073. doi: 10.1016/j.cscee.2024.101073

Terven, J., Cordova-Esparza, D. M., Romero-González, J. A., Ramírez-Pedraza, A., and Chávez-Urbiola, E. A. (2025). A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58. doi: 10.1007/s10462-025-11198-7

Wang, Z., Yin, Y., Xu, W., Mo, Y. K., Yang, H., Xiong, J., et al. (2025). Illuminating breast cancer malignancies: Lightweight histopathology computer vision classifier for precise breast cancer screening. *Eng Medicine*, 2:100053. doi: 10.1016/j.engmed.2024.100053

World Health Organization. (2022). Cancer today. Available online at: https://gco.iarc.who.int/today/en/dataviz/bars?mode=cancer&key=total&group_populations=1&types=1&sort_by=value0&populations=900&multiple_populations=0&values_position=out&cancers_h=39&sexes=2 (Accessed February, 2024).

Yang, X., Zhao, J., Zhang, H., Dai, C., Zhao, L., Ji, Z., et al. (2022). Remote sensing image detection based on YOLOv4 improvements. *IEEE Access* 10, 95527–95538. doi: 10.1109/ACCESS.2022.3204053

Zhao, S., Yan, C. Y., Lv, H., Yang, J. C., You, C., Li, Z. A., et al. (2024). Deep learning framework for comprehensive molecular and prognostic stratifications of triple-negative breast cancer. *Fundam. Res.* 4, 678–689. doi: 10.1016/j.fmre.2022.06.008

Zheng, Y., Li, C., Zhou, X., Chen, H., Xu, H., Li, Y., et al. (2023). Application of transfer learning and ensemble learning in image-level classification for breast histopathology. *Intell. Med.* 3, 115–128. doi: 10.1016/j.imed.2022.05.004