



Breast Cancer Histopathological Image Classification with Convolutional Neural Networks Models

Isil Unaldi¹, Leman Tomak²

^{1,2}Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ondokuz Mayıs University, Samsun, Türkiye

¹isil.unaldi@omu.edu.tr, ²lemant@omu.edu.tr

Abstract

Early diagnosis and treatment can reduce mortality rates by preventing the progression of breast cancer. Owing to convolutional neural networks (CNN), breast cancer diagnosis can be performed faster and more objectively than humans using thousands of histopathological images. This study aimed to evaluate and compare the rapid and effective diagnostic performance of CNN models on breast tumor images, utilizing transfer learning through pre-training and fine-tuning on novel datasets. The study was performed in two ways on BreakHis and BACH datasets. First, fine-tuned VGG16, VGG19, Xception, InceptionV3, ResNet50, and InceptionResNetV2 models were used for classification. Second, these CNN models were used as feature extractors and support vector machines (SVMs) as classifiers. The success of all models in tumor classification was interpreted using performance metrics, such as accuracy, precision, recall, F1 score, and AUC. The models showing the best performance as a result of the analyses were as follows: InceptionResNetV2+SVM model with an accuracy of 99.3%, precision of 99.0%, recall of 100.0%, F1 score of 99.5%, AUC of 98.9% for BreakHis dataset; and InceptionResNetV2 model with accuracy of 96.7%, precision of 93.8%, recall of 100.0%, F1 score of 96.8%, AUC of 96.7% for the BACH dataset. As a conclusion, it has been seen that the CNN methods have good generalization abilities and can respond to clinical needs.

Keywords: breast cancer; histopathological image classification; deep learning; convolutional neural network; support vector machine

How to Cite: I. Unaldi and L. Tomak, "Breast Cancer Histopathological Image Classification with Convolutional Neural Networks Models", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 5, pp. 963 - 974, Sep. 2025.
Permalink/DOI: <https://doi.org/10.29207/resti.v9i5.6420>

Received: February 25, 2025

Accepted: July 23, 2025

Available Online: September 29, 2025

*This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia*

1. Introduction

In the era of big data, the conversion of large amounts of data into useful information has become increasingly important in many areas [1]. The organization and use of data are very important for AI development, and machine learning methods have been adopted to develop human-level AI [2].

Machine learning is used to develop artificial intelligence methods that can learn complex problems such as visual perception, audition perception and natural language processing with human foreknowledge and intervention at least [3]. Feature engineering, which is used to extract data representations and features from data, is one of the fundamental elements of machine learning. However, machine learning algorithms cannot perform feature engineering automatically without human input [4]. To overcome this challenge, deep learning techniques have been developed that can process large amounts of data, while also automatically extracting features from the input [5].

Deep learning architectures have emerged as highly influential tools, demonstrating substantial efficacy in domains such as computer vision, natural language processing, speech analysis, and pattern identification [6]. It is an approach frequently used in the field of health. It is preferred by researchers in many applications such as medical image analysis, disease detection, lung nodule detection, skin cancer detection, EEG pathology detection and breast cancer detection [7]. Convolutional neural networks (CNNs), extensively applied in high-dimensional image recognition problems and increasingly in medical imaging, are regarded as one of the most influential deep learning architectures [8]. The remarkable performance of CNNs in natural image classification has led to their increasing adoption in medical imaging, particularly for classifying histopathological images [9]. Early diagnosis and treatment can reduce mortality rates by preventing the progression of breast cancer. For this reason, women are advised to check themselves by palpation, to have regular screening with

ultrasonography or mammography, and in case of abnormality, breast tissue biopsy is performed. In the evaluation phase, cancer findings are sought in microscopic parts by analyzing the histological features of the tissue. This is a very laborious and subjective process that requires experience, and by the advancement of computer-aided diagnostic systems, these systems, which offer a complementary and more objective assessment, help the diagnosis process [10]. In histopathology, the analysis of the obtained tissue image is performed manually by the pathologist, based entirely on her or his own experience. Examination of manual requires time and expertise. It can lead to erroneous results depending on the psychological and physical fatigue of the pathologist [11]. Manual analysis of histopathological images presents a significant challenge due to the complex structural patterns and the irregular morphology of benign and malignant tumors. Some of the challenges associated with manual analysis are the limited availability of experts in this field in hospitals and the requirement for significant labor from pathologists. In addition, this analysis cannot be performed quickly and always objectively because there are hundreds and thousands of images. Therefore, breast cancer classification using histopathologic images may not always be reliable and efficient [12]. Computer-aided diagnostic systems play a crucial role in enhancing diagnostic accuracy and efficiency, addressing the time-consuming nature of diagnosis and the variability in interpretations among pathologists. Conventional machine learning methods in histopathological imaging typically depend on manually engineered feature extraction, demanding considerable domain-specific expertise. Deep learning techniques enable automatic extraction of meaningful features from histopathological images of breast cancer, which frequently enhances classification performance. In this framework, CNNs are frequently utilized to automatically extract features for tumor classification, with the final categorization into benign or malignant carried out through support vector machines (SVMs). Alternatively, CNN models can perform both feature extraction and classification tasks simultaneously [13].

Many studies in the literature have used CNN models. Various studies have been conducted, such as different split ratios, different performance metrics, binary classification of data as a whole, and multi classification with tumor subgroups. Spanhol et al. [14] divided the BreakHis (Breast Cancer Histopathological Database) dataset by 70%-30% for training and testing, attempted different kernel methods for the SVM classifier, and obtained the best result with the Gaussian kernel. They used six different methods for classifier training, and achieved an approximate accuracy of 85% at the patient level. Spanhol et al. [15] proposed a method called the fusion rule (AlexNet + fusion rule) based on the extraction of image patches for CNN training and the assembly of these patches for the final classification. They reported medical images at two levels: the recognition rate was calculated at the patient and image

levels. Using the proposed method, they obtained optimal accuracies of 80.8% and 85.6% at the image level. Chang et al. [16] conducted a pilot study using BreakHis dataset. The InceptionV3 model showed a classification accuracy of 83% for the benign label and 89% for the malignant label. Mehra [17] trained the network with the VGG16, VGG19, and ResNet50 models, used logistic regression for classification, and performed a performance evaluation independent of magnification. With the CNN models considered, accuracy between 78.73% and 91.73% and a F1 score between 79% and 92% were obtained. Nathan et al. [18] applied two distinct transfer learning strategies utilizing ResNet, Inception, and DenseNet architectures for breast cancer classification. By extracting features with ResNet50, they achieved classification accuracies ranging from 93.34% to 95.04%, along with F1 scores between 92% and 94%, when logistic regression was used for classification. Kassani et al. [19] assessed the performance of binary classification on histopathological images by employing transfer learning and a fine-tuned ensemble model. The multi-model ensemble method they presented showed an accuracy of 98.13% for the BreakHis dataset and 95.00% for the BACH (BreAst Cancer Histology) dataset. InceptionV3, Xception, ResNet50, InceptionResNetV2, VGG16, and VGG19 models achieved accuracy values ranging from 79.48% to 96.41% for BreakHis, BACH, PatchCamelyon, and Bioimaging datasets, respectively. Zhang et al. [20] used VGGNet and ResNet on the BreakHis dataset, they examined the recognition rates at the patient level and image levels, and evaluated classification accuracies. Optimal accuracies were obtained with the ResNet50 model at 93.14% at the patient level and 92.68% at the image level. Gupta and Chawla [21] classified histopathology images using a two-stage model. First, features were extracted using the VGG16, VGG19, Xception, and ResNet50 models and transferred to SVM and logistic regression models for classification. The ResNet50 + SVM model showed the best performance among CNN+SVMs, with 86.40% accuracy at 40X magnification. Kausar et al. [22] proposed a deep CNN model for both binary and multi-class classification of breast cancer. The performance of the developed model was compared using Softmax and SVM as classifiers, and an accuracy of 97.3% was obtained with the SVM classifier. Sharma and Kumar [23] employed a pre-trained Xception model for feature extraction, followed by classification using a SVM with a radial basis function kernel, targeting magnification-specific binary classification tasks. The model demonstrated consistently high performance across varying magnification levels—40X, 100X, 200X, and 400X—achieving classification accuracies of 96.25%, 96.25%, 95.74%, and 94.11%, respectively. Sahu et al. [24], as a result of their experiment using the hybrid ResNet18+SVM model, obtained an accuracy of 92.6% for 200X magnification with the model they proposed. Zerouaoui and Idri [25] employed several pre-trained

CNN models, including DenseNet201, InceptionV3, InceptionResNetV2, MobileNetV2, ResNet50, VGG16, and VGG19, for feature extraction, followed by classification using MLP, SVM, decision trees (DT), and k-nearest neighbors (KNN). The BreakHis and FNAC datasets achieved accuracies of 92.47% and 94.78%, respectively, using InceptionV3+SVM. The results underscored the utility of combining deep learning-driven feature extraction with conventional classifiers in the classification of breast cancer tumors. Clement et al. [26] performed breast cancer histopathological images multi classification using eight pre-trained CNN architectures with an ensemble of four models that performed the best for feature extraction. The ensemble model, which consists of four CNN architectures (ResNet50, ResNet18, DenseNet201, and EfficientNetb0) + SVM, achieved the highest performance, reaching an accuracy of 97.77%. In the study, the experiments were handled in two ways. Firstly, fine-tuned VGG16, VGG19, Xception, InceptionV3, ResNet50 and InceptionResNetV2 models were used for classification and the results were reported. Secondly, these CNN models were used as feature extractors, then SVM was used as classifier and results were reported. As a result of repeated trials and literature reviews, it was determined that SVM used as a classifier together with CNN models could achieve better results, so the study was planned accordingly. The success of all models was evaluated by accuracy, precision, recall, F1 score and AUC. The BreakHis and BACH datasets were chosen to give us an idea of whether a model is generalizable due to their size differences, and because these are frequently used in the literature to diagnose histopathological breast cancer. The aim of the study is to examine the performances of pre-trained convolutional neural networks on breast histopathological images and to compare them with each other.

2. Methods

2.1. Experimental Setup

Google Colaboratory was used to conduct the data analyzes with the Python programming language. Evaluations were performed using the Tensorflow and Keras deep learning frameworks and the scikit-learn machine learning framework in Python. Models were run with GPU support. Both datasets have been divided into by 70%-15%-15% as training-validation-testing. Different data preprocessing methods have been implemented to datasets and set to: Rescale = 1/255.0; Horizontal flip and Vertical flip = True; Shear range, Zoom range, Width shift range and Height shift range = 0.2; Rotation range = 45 degree; Brightness range = [0.5,1.0]; Samplewise center and Samplewise standard normalization = True; Fill mode = Nearest. The hyperparameter adjustments applied in the analyzes are as follows: Epoch=100, Batch size=32, Dropout=[0.3-0.5], Minimum learning rate = 0.0001. The Adam

optimization method was used. Linear kernel method was chosen for SVM. Hyperparameter optimization for the SVM classifier was performed with the GridSearchCV method. 10-fold cross-validation was applied. In this study, the image data was first subjected to pre-processing by introducing it into the algorithm. CNN models were trained. In the first experiment, classification was made on the test dataset with the trained models. In the second experiment, CNN was used as the feature extractor and SVM was used as the classifier. Afterwards, performance measures were obtained and performances were compared. The methodology used in this study is presented as a flowchart in Figure 1.

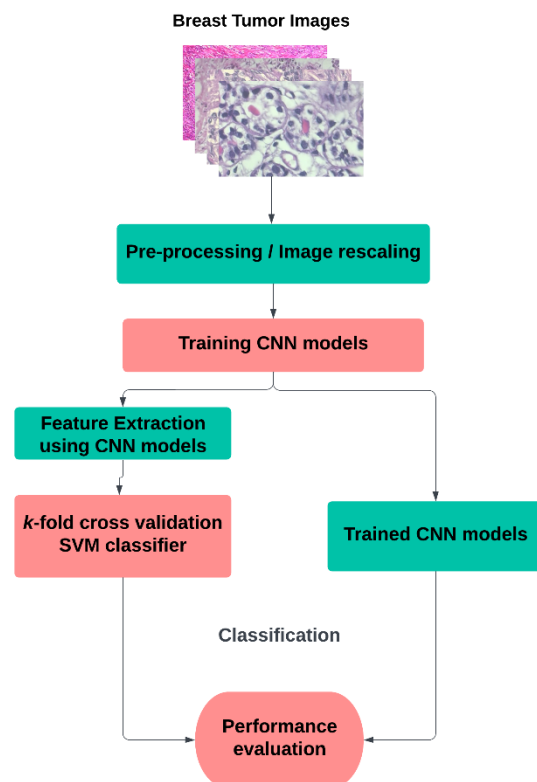


Figure. 1. The Methodology of the Study

2.2. Convolutional Neural Networks

CNN is a deep and feedforward artificial neural network class that is frequently used in image processing and computer vision problems are accepted as a cutting-edge technology for image classification [27]. CNN models extract features from train images by optimizing the loss function in classification studies. CNN models have demonstrated strong performance in image classification tasks across various domains, particularly in medical image analysis and, more specifically, in the interpretation of histopathological images [28].

The training process in CNN is an iterative process of the backpropagation algorithm, which continues until the model minimizes the error to a tolerable extent [29]. CNN learns automatically with the help of filters. These filters help extract accurate and relevant features from the input data [30]. A conventional CNN is organized

into multiple components-such as convolutional, pooling, normalization, activation, fully connected, dropout, and softmax layers-each designed to fulfill a distinct role within the model. The CNN architecture is structured as a series of layers, including convolution, pooling, and normalization layers, arranged sequentially to facilitate hierarchical feature extraction. The overlapping of convolution and pooling operations creates the high-level features to be classified [31].

2.3. Datasets

The open source BreakHis¹ [14] database was used in the study. BreakHis database contains 7909 images (2480 benign, 5429 malignant) of microscopic breast tissue at magnification levels of 40X, 100X, 200X, and 400X. These images were obtained from 82 patients.

The BACH² [13] dataset consists of breast whole-slide and histopathological images. A total of 400 microscopy images (100 benign, 100 normal, 100 invasive, 100 in situ) are available for analysis. The BACH dataset was categorized into two diagnostic groups: benign (encompassing normal and benign lesions) and malignant (including in situ and invasive carcinoma).

2.4. Transfer Learning

Transfer learning represents an approach within machine learning in which prior knowledge obtained from one problem is exploited to facilitate and accelerate learning in a related problem domain [31]. Transfer learning applies trained parameters to a new classification task using information learned from a pre-trained network [2].

With this method, a model predicted in the source area is reused in a related different target area. [32].

2.5. Support Vector Machines

In the field of machine learning, SVMs are widely recognized as effective methods for addressing problems involving two-class classification. Based on statistical learning, SVMs are used to solve many complex classification problems such as learning linear estimators for high-dimensional feature spaces [33].

The SVM algorithm simply searches for the hyperplane with the largest margin that is linearly separable [34]. The larger the margin, the better the data are separated; therefore, it is easier to find a hyperplane that classifies the dataset correctly [35]. In scenarios where the feature space dimensionality exceeds the number of training samples, SVMs optimize memory usage by relying on support vectors that define the decision boundary. SVMs cannot directly estimate probabilities. Therefore, the probabilities are obtained using the cross-validation method [36].

2.6. Cross Validation

Cross-validation is a repeated sampling procedure that measures the performance of a machine learning algorithm on a dataset [37]. k-fold validation is the most commonly used cross-validation method [38]. In this approach, the dataset is partitioned into k subsets of equal size.

During each iteration, one subset is designated as the validation set, whereas the remaining $k - 1$ subsets are employed for training. The procedure is repeated k times so that every subset is utilized once for validation. Performance is then assessed by aggregating the results across all iterations, with error estimation methods commonly applied to evaluate predictive accuracy [39].

2.7. CNN Layers

Convolution layers are feature extractors that learn feature representations from input samples. Neurons of convolution layers are placed on feature maps. In CNNs, each neuron in a feature map receives inputs from a localized region of the preceding layer, with the strength of these connections determined by the weights of the corresponding filter [40]. Learnable filters that generate feature maps from the output image are combined with the input image in the convolution layer [15].

The reduction of resolution for feature maps, the adjustment of noise- and clutter-resistant features, and the making of learned features resistant to minor changes are all performed in the pooling layer. Essentially, reduction of the feature maps size and the input map downsampling are performed in pooling layer [41]. Neurons tied to activation maps in the previous layer are located in the fully connected layer. After the number of convolutions, pooling, and ReLU layers specified by the researcher, a fully connected layer follows [42]. This layer perform several tasks. In this layer, the features extracted from the previous layer are flattened and assigned to a vector, and the weight is updated to improve the features in the vector. Prior to classification, the final probability for each class label is computed based on the corresponding label feature vectors [43]. The activation layer is the layer where inputs are fed into a scalar function (activation function) and a value is obtained [44]. Each successive layer of the neural network is a linear combination of its inputs. In the case of nonlinearity, entering an activation function ensures that the data can be classified even if they are not linearly separable [45]. Common activation functions used within the activation layer encompass ReLU, sigmoid, hyperbolic tangent (tanh), and similar nonlinear functions. Overfitting occurs in the test dataset owing to the overtraining of a model. In such cases, the model performed poorly. This overfitting

¹This database was obtained from <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>

²This database was obtained from <https://iclar2018-challenge.grand-challenge.org/Dataset/>

problem can be addressed with the dropout process, which is applied by randomly selecting some nodes and removing them from the network. At this stage, it is trained only on reduced network data. The cancelled nodes are then reconnected to the network [46].

2.8. CNN Models

The VGG16 architecture is a convolutional neural network structured with 16 weight-bearing layers, comprising 13 convolution blocks and 3 fully connected layers. In this network, 3×3 filters and stride = padding = 1 are characteristics of the convolution layers. It contains 5 groups of convolution layers, each followed by a 2×2 filtered maximum pooling layer. This layer reduces computational load and prevents overfitting by applying downsampling to images [2].

VGG19 is a CNN comprising 16 convolutional layers followed by 3 fully connected layers. The architecture extensively uses 3×3 convolutional filters, a design choice that significantly reduces the number of parameters in the convolutional layers compared to larger kernels. It involves about 138 million calculation parameters [43]. Xception has VGG-like architecture. Inspired by the InceptionV3's Inception modules. The Xception architecture has residual connections and consists of deeply separable convolutional layers. It has approximately 20 million learnable parameters [47]. The network architecture arranges its 36 convolution layers into 14 distinct modules, where each module—except the initial and final ones—incorporates linear residual connections to facilitate gradient flow [46]. ResNet implements propagation by feeding residual values from layer to layer using residual blocks. ResNet50 is now a 50-layer network type grouped in blocks and has been very successful in training very deep architectures [48]. The InceptionV3 architecture has 44 layers and 21 million learnable parameters [47]. The InceptionV3 model was trained using a 1000-class dataset from the ImageNet database [49]. The InceptionResNetV2 architecture is based on the inception architecture as a combination of connections. The InceptionResNetV2 architecture uses residual links with high efficiency [48]. The architecture addresses the gradient degradation problem by adding units to the inception modules. This degradation can also be solved by selecting different filter sizes and adding skip layers [50].

2.9. Model Evaluation

The performance of the architectures was evaluated according to specific criteria for the accuracy and efficiency of the classification results. Commonly used performance metrics include the accuracy, sensitivity, precision, recall, AUC, and F1 score [50]. The performance metrics used were calculated using a confusion matrix.

True positive (TP) refers to the count of positive instances that are correctly identified. False positive (FP) denotes the number of instances that are falsely

classified as positive, despite being negative. True negative (TN) indicates the number of negative instances that are accurately classified, while false negative (FN) represents the number of positive instances that are incorrectly classified as negative.

Accuracy (Acc). Classification accuracy is the ratio of correctly classified samples [51] as shown in Formula 1.

$$Acc = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Precision (Pr). The precision refers to the number of predicted positives are correct [52] as shown in Formula 2.

$$Pr = TP / (TP + FP) \quad (2)$$

F1 score. The F1 score represents the harmonic mean of precision and recall, combining both metrics into a single performance indicator, as illustrated in Formula 3 [53].

$$F1 = (2 \times TP) / (2 \times TP + FP + FN) \quad (3)$$

The F1 score ranges from 0 to 1, where 0 denotes the worst performance and 1 represents the best performance. The F1 score was employed as the primary metric to assess the model's overall performance.

Recall. The recall indicates the number of positive cases correctly predicted by the classifier. This is also known as the Sensitivity, True Positive Rate (TPR) and Probability of Detection [54] as shown in Formula 4.

$$Recall = TP / (TP + FN) \quad (4)$$

AUC. The Receiver Operating Characteristic (ROC) curve depicts the trade-off between the TPR and the FPR, offering a visual means of evaluating a model's diagnostic capability. AUC quantifies this performance by measuring the entire area beneath the ROC curve. An increased AUC score reflects improved model capability in differentiating between positive and negative classes.

3. Results and Discussions

3.1 Results

Deep feature extraction was performed with the fine-tuned VGG16, VGG19, Xception, InceptionV3, ResNet50, and InceptionResNetV2 models pre-trained with the ImageNet database. Deep features extracted using CNN models were transferred to the linear SVM machine learning model by transfer fine-tuning learning, the model was trained, and the performance evaluation was performed for each CNN model. Model efficacy was evaluated using established performance metrics—accuracy, precision, recall, F1 score, and AUC. The performance metrics of the CNN models for the BreakHis and BACH datasets are listed in Tables 1 and 2, respectively. For the BreakHis dataset, the InceptionResNetV2+SVM model performance at 40X magnification, with an accuracy of 99.3%, precision of 99.0%, recall of 100.0%, F1 score of 99.5%, and AUC

of 98.9%. InceptionResNetV2 achieved a precision of 99.5%. At 40X magnification, the VGG16+SVM model achieved an accuracy of 93.3%, precision of 97.4%, recall of 92.7%, F1 score of 95.0%, and AUC of 93.7%. Using the Xception+SVM model at 100X magnification, an accuracy of 99.0%, precision of 99.1%, recall of 99.5%, F1 score of 99.3%, and AUC of 98.7% were obtained. At 100X magnification, the InceptionV3+SVM model showed performance with accuracy of 97.4%, precision of 97.7%, recall of 98.6%, F1 score of 98.2% and AUC of 96.7%. At 100X magnification, the ResNet50+SVM model achieved an accuracy of 97.1%, a precision of 97.3%, a recall of 98.6%, an F1 score of 97.9%, and an AUC of 96.2%.

For the BACH dataset, using the InceptionResNetV2 model, an accuracy and AUC of 96.7%, precision of 93.8%, recall of 100.0%, and F1 score of 96.8% were obtained. The InceptionResNetV2+SVM model showed an accuracy and AUC of 91.7%, a precision of 87.9%, a recall of 96.7%, and an F1 score of 92.1%. The Xception+SVM model achieved an accuracy and AUC of 90.0%, a precision of 87.5%, a recall of 93.3%, and an F1 score of 90.3%. For the InceptionV3+SVM model, the accuracy and AUC were 95.0%, precision of 93.5%, recall of 96.7%, F1 score of 95.1% were obtained. The ResNet50+SVM model exhibited an accuracy and AUC of 93.3%, precision of 90.6%, recall of 96.7%, and F1 score of 93.5%.

3.2 Discussions

A pivotal moment in image recognition came with the 2009 launch of ImageNet—an extensive, annotated image database—that subsequently formed the foundation for the annual ImageNet Large Scale Visual Recognition Challenge, aimed at driving down classification error rates in object recognition tasks. The emergence of deep learning as a dominant approach has been largely driven by the advancements and scale of the ImageNet dataset [55].

The rapid advancements in deep learning—characterized by the availability of large-scale annotated datasets and powerful GPU acceleration—have markedly elevated the prominence of CNNs in image classification tasks. Compared with handcrafted feature-based methods, high-level image features automatically learned by state-of-the-art CNN architectures are more representative [56]. After extracting the features learned by CNN algorithms, a machine learning method is typically used as a classifier. The high dimensionality of the feature space results in problems with the sample complexity and computational complexity. To tackle these challenges, SVM algorithms employ large-margin decision boundaries, effectively managing complexity in high-dimensional feature spaces. Even if certain training samples lie within the margin or are misclassified, a soft-margin SVM can still maintain a wide separation boundary by allowing slack in the

optimization process [33]. Figure 2 presents the learning curves of the models that achieved the highest performance scores on the BreakHis test dataset. Figure 3 illustrates the learning curves of the models that achieved the highest performance metrics on the BACH test dataset. The configuration of learning curves reveals insights into a model's architecture and highlights potential areas for refinement. The parallel learning curves show that the model training is successful and the tendency to overlearning is optimized. When the learning curves are examined, it is seen that model training with VGG16 at 40X magnification, with Xception at 100X magnification, with InceptionV3 at 200X magnification, with InceptionResNetV2 at 40X magnification for the BreakHis dataset passed well and overlearning was optimized. With the BACH dataset, it was seen that the most optimal training phase was realized with the InceptionResNetV2 model. Figures 4 and 5 illustrate the confusion matrices of the top-performing models on the test sets of the BreakHis and BACH datasets, respectively. Confusion matrices display both correct and incorrect predictions made during the classification process using the test dataset. Accordingly, the best classification performance for the BreakHis dataset was obtained with the InceptionResNetV2+SVM model at 40X magnification, while the best classification performance for the BACH dataset was obtained with the InceptionResNetV2 model. The use of transfer learning methods is extremely important for generalizability. Even if the dataset is small in size, it performs well on a specific medical imaging task because it uses general features learned on a vast dataset like ImageNet. The fine-tuned transfer learning models used in this study, which achieve high performance values and have parallel learning curves, are generalizable, given the learning curves, the size difference of the datasets used, and the classification results. However, they still need to be different fine-tuned for each task, each type of data set and for different models. The most successful model among all the models and magnifications for the BreakHis dataset was InceptionResNetV2+SVM at 40X magnification, with an accuracy of 99.3%, a precision of 99.0%, a recall of 100.0%, an F1 score of 99.5%, and an AUC of 98.9%. InceptionResNetV2 achieved the highest precision (99.5% at 200X magnification). Xception+SVM was the most successful model, with an accuracy of 99.0%, a precision of 99.1%, a recall of 99.5%, an F1 score of 99.3%, and an AUC of 98.7% at 100X magnification. For magnification 200X, the InceptionResNetV2+SVM model had the best performance, with an accuracy of 98.7%, precision of 99.0%, recall of 99.0%, F1 score of 99.0%, and AUC of 98.5%. The InceptionResNetV2+SVM model was determined to have the best performance at 400X magnification, with an accuracy of 95.6%, precision of 94.8%, recall of 98.9%, F1 score of 96.8%, and AUC of 93.8%.

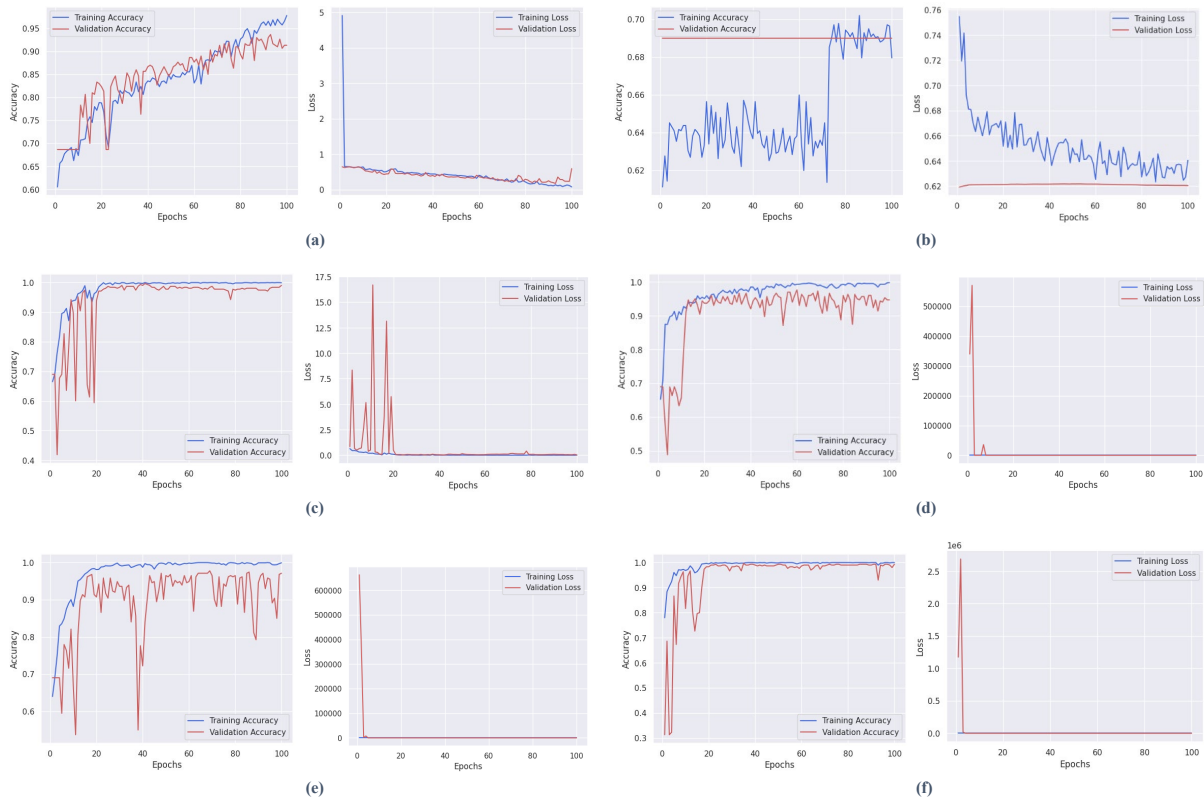


Figure 2. Learning Curves for BreakHis Dataset which has the Best Accuracy in Test Dataset, (a) Learning Curves of VGG16 Model for 40X Images, (b) Learning Curves of VGG19 Model for 100X Images, (c) Learning Curves of Xception Model for 100X Images, (d) Learning Curves of InceptionV3 Model for 200X Images, (e) Learning Curves of ResNet50 Model for 100X Images, (f) Learning Curves of InceptionResNetV2 Model for 40X Images

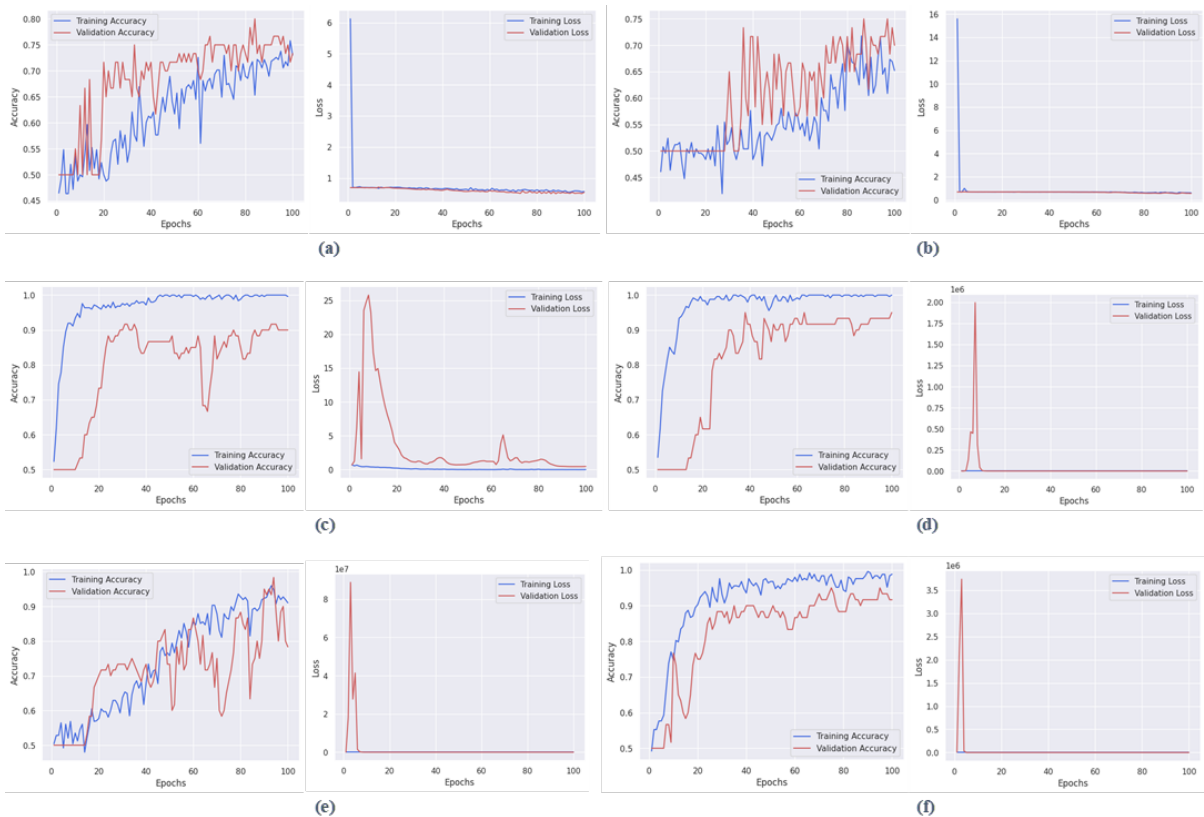


Figure 3. Learning Curves for BACH Dataset, (a) Learning Curves for VGG16 Model, (b) Learning Curves for VGG19 Model, (c) Learning Curves for Xception Model, (d) Learning Curves for InceptionV3 Model, (e) Learning Curves for ResNet50 Model, (f) Learning Curves for InceptionResNetV2 Model

Table 1. Performance Metrics for Binary Classification Models for BreakHis Dataset

Model	Magnification	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
VGG16	40X	87.7	99.4	82.5	90.2	90.7
Avg* Acc=86.9	100X	87.5	88.6	94.0	91.2	83.6
Avg Pr=91.8	200X	87.8	93.4	88.5	90.9	87.3
Avg Recall=89.5	400X	84.7	85.6	93.0	89.1	80.2
Avg F1 Score=90.4						
Avg AUC=85.5						
VGG16 + SVM	40X	93.3	97.4	92.7	95.0	93.7
Avg Acc=89.1	100X	88.2	89.4	94.0	91.6	84.6
Avg Pr=91.0	200X	88.4	91.0	92.3	91.7	86.1
Avg Recall=93.5	400X	86.5	86.3	95.1	90.5	81.8
Avg F1 score=92.2						
Avg AUC=86.6						
VGG19	40X	68.7	68.7	100.0	81.4	50.0
Avg Acc=68.6	100X	69.0	69.0	100.0	81.7	50.0
Avg Pr=68.6	200X	69.0	69.0	100.0	81.6	50.0
Avg Recall=100.0	400X	67.5	67.5	100.0	80.6	50.0
Avg F1 score=81.3						
Avg AUC=50.0						
VGG19 + SVM	40X	68.7	68.7	100.0	81.4	50.0
Avg Acc=68.6	100X	69.0	69.0	100.0	81.7	50.0
Avg Pr=68.6	200X	69.0	69.0	100.0	81.6	50.0
Avg Recall=100.0	400X	67.5	67.5	100.0	80.6	50.0
Avg F1 score=81.3						
Avg AUC=50.0						
Xception	40X	97.0	99.0	96.6	97.8	97.2
Avg Acc=95.2	100X	97.1	97.7	98.1	97.9	96.5
Avg Pr=96.2	200X	94.4	93.6	98.6	96.0	91.8
Avg Recall=96.9	400X	92.3	94.6	94.1	94.3	91.4
Avg F1 score=96.5						
Avg AUC=94.2						
Xception + SVM	40X	98.7	99.0	99.0	99.0	98.5
Avg Acc=97.0	100X	99.0	99.1	99.5	99.3	98.7
Avg Pr=97.1	200X	95.7	95.0	99.0	97.0	93.7
Avg Recall=98.6	400X	94.5	95.2	96.8	96.0	93.3
Avg F1 score=97.8						
Avg AUC=96.1						
InceptionV3	40X	96.3	95.3	99.5	97.4	94.4
Avg Acc=96.3	100X	97.4	97.3	99.1	98.2	96.4
Avg Pr=96.1	200X	98.0	98.6	98.6	98.6	97.7
Avg Recall=98.6	400X	93.4	93.3	97.3	95.2	91.3
Avg F1 score=97.4						
Avg AUC=95.0						
InceptionV3 + SVM	40X	95.0	94.0	99.0	96.5	92.6
Avg Acc=95.1	100X	97.4	97.7	98.6	98.2	96.7
Avg Pr=95.1	200X	94.7	96.6	95.7	96.2	94.1
Avg Recall=97.9	400X	93.1	91.9	98.4	95.0	90.2
Avg F1 score=96.5						
Avg AUC=93.4						
ResNet50	40X	77.0	99.3	67.0	80.0	83.0
Avg Acc=88.7	100X	92.7	92.1	97.7	94.8	89.6
Avg Pr=94.6	200X	95.4	94.5	99.0	96.7	93.1
Avg Recall=89.0	400X	89.8	92.4	92.4	92.4	88.4
Avg F1 score=91.0						
Avg AUC=88.5						
ResNet50 + SVM	40X	91.0	99.4	87.4	93.0	93.2
Avg Acc=94.1	100X	97.1	97.3	98.6	97.9	96.2
Avg Pr=96.5	200X	96.7	95.9	99.5	97.7	95.0
Avg Recall=94.9	400X	91.6	93.5	94.1	93.8	90.3
Avg F1 score=95.6						
Avg AUC=93.7						
InceptionResNetV2	40X	98.3	98.6	99.0	98.8	97.9
Avg Acc=97.5	100X	96.8	96.8	98.6	97.7	95.7
Avg Pr=98.1	200X	98.3	99.5	98.1	98.8	98.5
Avg Recall=98.3	400X	96.4	97.3	97.3	97.3	95.8
Avg F1 score=98.2						
Avg AUC=97.0						
InceptionResNetV2 + SVM	40X	99.3	99.0	100.0	99.5	98.9
Avg Acc=98.1	100X	98.7	99.1	99.1	99.1	98.5
Avg Pr=98.0	200X	98.7	99.0	99.0	99.0	98.5
Avg Recall=99.3	400X	95.6	94.8	98.9	96.8	93.8
Avg F1 score=98.6						
Avg AUC=97.4						

*: Average

Table 2. Performance Metrics for Binary Classification Models for BACH Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
VGG16	66.7	63.9	76.7	69.7	66.7
VGG16 + SVM	75.0	70.3	86.7	77.6	75.0
VGG19	65.0	80.0	93.3	53.3	65.0
VGG19 + SVM	71.7	69.7	76.7	73.0	71.7
Xception	83.3	81.3	86.7	83.9	83.3
Xception + SVM	90.0	87.5	93.3	90.3	90.0
InceptionV3	81.7	78.8	86.7	82.5	81.7
InceptionV3 + SVM	95.0	93.5	96.7	95.1	95.0
ResNet50	85.0	76.9	100.0	87.0	85.0
ResNet50 + SVM	93.3	90.6	96.7	93.5	93.3
InceptionResNetV2	96.7	93.8	100.0	96.8	96.7
InceptionResNetV2 + SVM	91.7	87.9	96.7	92.1	91.7

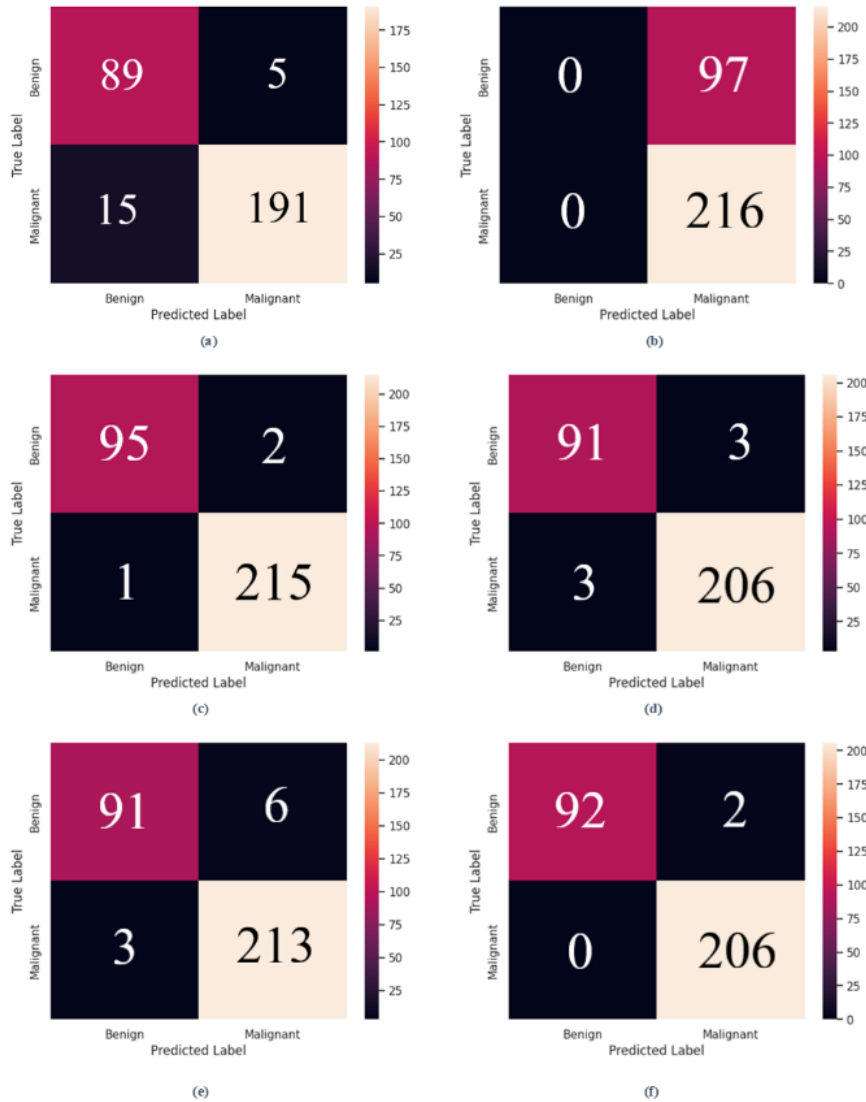


Figure 4. Confusion Matrices of Models which has the Best Accuracy in Test Dataset, (a) Confusion Matrix of the VGG16+SVM Model using 40X Images, (b) Confusion Matrix of the VGG19+SVM Model using 100X Images, (c) Confusion Matrix of the Xception+SVM Model using 100X Images, (d) Confusion Matrix of the InceptionV3 Model using 200X Images, (e) Confusion Matrix of the ResNet50+SVM Model using 100X Images, (f) Confusion Matrix of the InceptionResNetV2+SVM Model using 40X Images

The VGG19 and VGG19+SVM models exhibited the worst performance for each magnification. Considering the average performance metrics for each model, it was determined that InceptionResNetV2+SVM was the most successful model, with an accuracy of 98.1%, precision of 98.0%, recall of 99.3%, F1 score of 98.6%, and AUC of 97.4%. It was found that adding SVM as a

classifier to all other CNNs, except for the Xception and VGG19 models, increased the performance. For the BACH dataset, the most successful model was InceptionResNetV2, with an accuracy of 96.7%, a precision of 93.8%, a recall of 100.0%, an F1 score of 96.8%, and an AUC of 96.7%. Using the Xception+SVM, InceptionV3+SVM, ResNet50+SVM,

and InceptionResNetV2+SVM models, performance measures of 90% and above were achieved. The precision measures of the Xception+SVM and InceptionResNetV2+SVM models were 87.5% and 87.9%, respectively. The VGG16, VGG19,

VGG19+SVM, and VGG16+SVM models exhibited the worst performance for the BACH dataset. The performance measures for these models remained below 80% (except for recall measures).

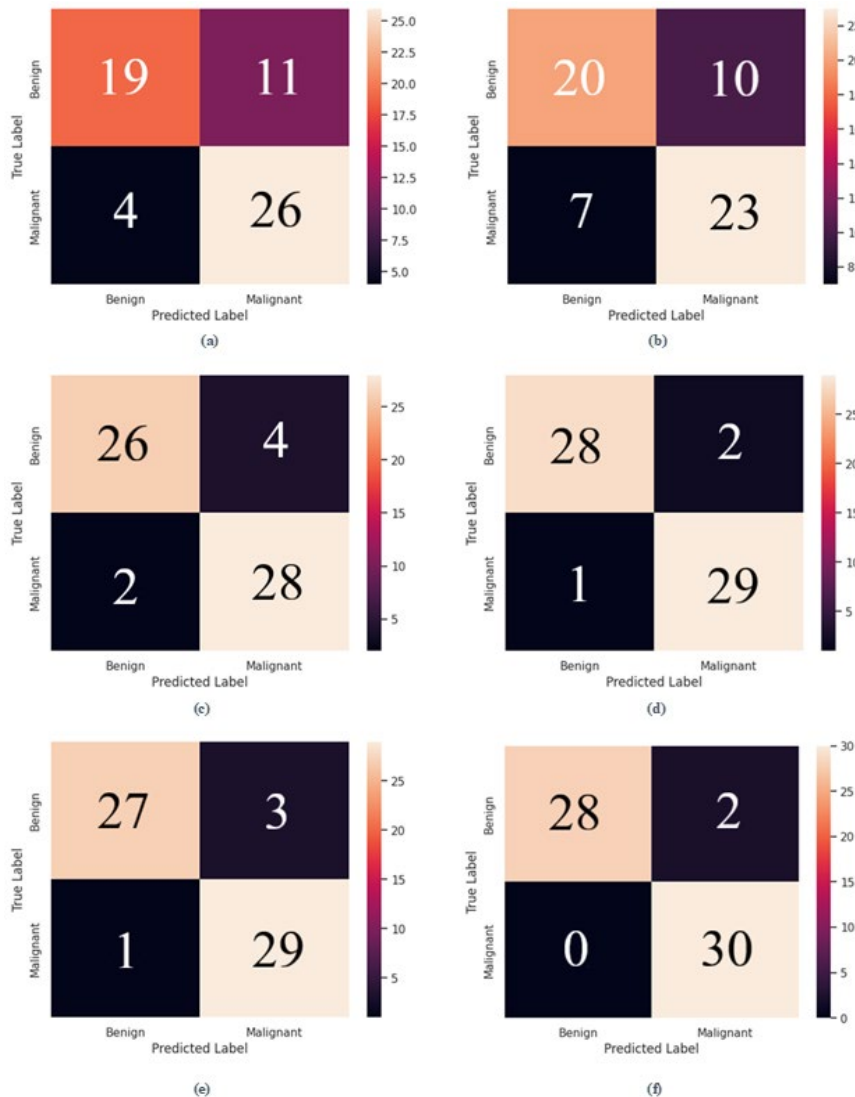


Figure 5. Confusion Matrices of Models which has the Best Accuracy in BACH Dataset (a) Confusion Matrix of the VGG16+SVM Model, (b) Confusion Matrix of the VGG19+SVM Model, (c) Confusion Matrix of the Xception+SVM Model, (d) Confusion Matrix of the InceptionV3+SVM Model, (e) Confusion Matrix of the ResNet50+SVM Model, (f) Confusion Matrix of the InceptionResNetV2 Model

4. Conclusions

Considering all the results, it was concluded that all models yielded successful results, except for VGG19 and VGG19+SVM. Although the VGG16 and VGG16+SVM models obtained successful results for BreakHis, successful results were not obtained for the BACH dataset. According to the learning curves, InceptionResNetV2 was found to be the model with the most successful training process for both datasets.

The learning curves and classification results also showed that it has a high generalization capability. All other models examined in this study are very successful for image classification owing to their deep architectures. It can be said that the results obtained may

differ depending on the hyperparameters determined and adjusted by the researcher's experience according to the dataset and the nature of the problem, algorithm differences, and the random and independent separation of the training and testing sets for each study.

The results from this study showed that more-layered models are costly in terms of time and memory, but perform better in classification and feature extraction. In addition, it was observed that the use of the SVM classifier together with the CNN models increased the performance.

The results of the study showed that using CNN models for feature extraction and feeding the generated feature maps to a machine learning classifier method

significantly affected the model performances. A critical determinant of deep learning algorithms' success is the availability of a substantial volume of labeled training data. However, high performance resulting from training can be an indicator of excessive learning. For this reason, an appropriate regularization method is preferred to reduce overfitting. A large data size allows deep architectures to yield successful results.

An architecture that is multi-layered and complex may not always yield the best results. The results depend on data preprocessing, models appropriate to the problem, the train process, the size of the data used in train, the split ratio, hyperparameter tuning methods, and regularization methods. In terms of classification performance, it can be said that using machine learning methods as classifiers together with CNN models is a powerful and applicable procedure. Many methods have been developed and used for classification in the literature. The depth of the model does not imply that it will performed the best for the problem under consideration. This was because the model had to fit the data. The more compatible the model is with the data, the higher its performance. In this study, it was found that InceptionResNetV2 and InceptionResNetV2+SVM models have good generalization capability and can respond to clinical problems.

In future studies, we plan to evaluate the current histopathological breast tissue images with higher dimensions and more diversity, with different CNN models, binary classification, or different split ratios according to benign-malignant subgroups, regularization methods, hyperparameter selections and adjustments, and different performance metrics. At the same time, it is recommended to try different models by using other machine learning techniques as classifiers, and thus provide an approach for disease diagnosis by using medical images from other fields.

References

- [1] A. S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, no. 5, 2017. [Online]. Available: <https://doi.org/10.1093/bib/bbw068>.
- [2] X. Zhen, J. Chen, Z. Zhong, B. Hrycushko, L. Zhou, S. Jiang, K. Albuquerque, and X. Gu, "Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: A feasibility study," *Physics in Medicine and Biology*, vol. 62, no. 21, 2017. [Online]. Available: <https://doi.org/10.1088/1361-6560/aa8d09>.
- [3] Y. Bengio and Y. Lecun, "Scaling learning algorithms towards AI," in *Large-Scale Kernel Machines*, New York, vol. 1, 2007.
- [4] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, 2015. [Online]. Available: <https://doi.org/10.1186/s40537-014-0007-7>.
- [5] B. Krishnakumar, K. Kousalya, R. S. Mohana, K. Dinesh, and S. Santhiya, "Classification of breast cancer using deep learning architecture," *International Journal of Recent Technology and Engineering (IJR)*, 2019.
- [6] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," *arXiv preprint*, arXiv:1710.09282, 2017.
- [7] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," *Future Generation Computer Systems*, vol. 102, 2020. [Online]. Available: <https://doi.org/10.1016/j.future.2019.09.015>.
- [8] H. Mechria, M. S. Gouider, and K. Hassine, "Breast cancer detection using deep convolutional neural network," in *Proc. 11th Int. Conf. on Agents and Artificial Intelligence (ICAART)*, 2019, vol. 2. [Online]. Available: <https://doi.org/10.5220/0007386206550660>.
- [9] A. Golatkar, D. Anand, and A. Sethi, "Classification of breast cancer histology using deep learning," *Lecture Notes in Computer Science*, vol. 10882, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-93000-8_95.
- [10] G. Aresta et al., "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, 2019. [Online]. Available: <https://doi.org/10.1016/j.media.2019.05.010>.
- [11] Y. Sahu, A. Tripathi, R. K. Gupta, P. Gautam, R. K. Pateriya, A. Gupta, "A CNN-SVM based computer aided diagnosis of breast Cancer using histogram K-means segmentation technique," *Multimed Tools Appl*, 82(9), 14055-14075, 2023.
- [12] T. Kausar, Y. Lu, A. Kausar, "Breast cancer diagnosis using lightweight deep convolution neural network model", *IEEE Access*, 11, 124869-124886, 2023.
- [13] S. K. V.N. and M. Vijay, "Diagnosing Cancer Cells Using Histopathological Images with Deep Learning" *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, pp. 148-152, 2021. doi: 10.1109/WiSPNET51692.2021.9419468.
- [14] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, 2016. [Online]. Available: <https://doi.org/10.1109/TBME.2015.2496264>.
- [15] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2016. [Online]. Available: <https://doi.org/10.1109/IJCNN.2016.7727519>.
- [16] J. Chang, J. Yu, T. Han, H. J. Chang, and E. Park, "A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer," in *Proc. 19th IEEE Int. Conf. E-Health Networking, Applications and Services (Healthcom)*, 2017. [Online]. Available: <https://doi.org/10.1109/HealthCom.2017.8210843>.
- [17] R. Mehra, "Breast cancer histology images classification: Training from scratch or transfer learning?," *ICT Express*, vol. 4, no. 4, pp. 247-254, 2018. [Online]. Available: <https://doi.org/10.1016/j.ict.2018.10.007>.
- [18] D. S. Nathan, R. Saravanan, J. Anbazhagan, and P. Koduganty, "Comparison of deep feature classification and fine tuning for breast cancer histopathology image classification," in *Recent Trends in Image Processing and Pattern Recognition*, vol. 2, pp. 58-68, Springer, 2019.
- [19] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Classification of histopathological biopsy images using ensemble of deep learning networks," in *Proc. CASCON 2019*, 2020.
- [20] J. Zhang, X. Wei, C. Che, and Q. Zhang, "Breast cancer histopathological image classification based on convolutional neural networks," *J. Med. Imaging Health Informatics*, vol. 9, no. 4, 2019. [Online]. Available: <https://doi.org/10.1166/jmhi.2019.2648>.
- [21] K. Gupta and N. Chawla, "Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained CNN," *Procedia Computer Science*, vol. 167, 2020. [Online]. Available: <https://doi.org/10.1016/j.procs.2020.03.427>.
- [22] T. Kausar, M. A. Ashraf, A. Kausar, and I. Riaz, "Convolution neural network-based approach for breast cancer type classification," in *Proc. 18th Int. Bhurban Conf. Appl. Sci.*

- Technol. (IBCAST), 2021. [Online]. Available: <https://doi.org/10.1109/IBCAST51254.2021.9393249>.
- [23] S. Sharma and S. Kumar, "The Xception model: A potential feature extractor in breast cancer histology images classification," *ICT Express*, vol. 8, no. 1, 2022. [Online]. Available: <https://doi.org/10.1016/j.icte.2021.11.010>.
- [24] Y. Sahu, A. Tripathi, R. K. Gupta, P. Gautam, R. K. Pateriya, and A. Gupta, "A CNN-SVM based computer-aided diagnosis of breast cancer using histogram K-means segmentation technique," *Multimedia Tools and Applications*, vol. 82, no. 9, 2023. [Online]. Available: <https://doi.org/10.1007/s11042-022-13807-x>.
- [25] H. Zerouaoui and A. Idri, "Deep hybrid architectures for binary classification of medical breast cancer images," *Biomed. Signal Process. Control*, vol. 71, 2022. [Online]. Available: <https://doi.org/10.1016/j.bspc.2021.103226>.
- [26] D. Clement, E. Agu, M. A. Suleiman, J. Obayemi, S. Adeshina, and W. Soboyejo, "Multi-class breast cancer histopathological image classification using multi-scale pooled image feature representation (MPIFR) and one-versus-one support vector machines," *Appl. Sci.*, vol. 13, no. 1, 2023. [Online]. Available: <https://doi.org/10.3390/app13010156>.
- [27] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, 2017. [Online]. Available: <https://doi.org/10.1016/j.neucom.2016.12.038>.
- [28] T. Araujo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polonia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PLoS ONE*, vol. 12, no. 6, 2017. [Online]. Available: <https://doi.org/10.1371/journal.pone.0177544>.
- [29] S. A. Adeshina, A. P. Adedigba, A. A. Adeniyi, and A. M. Aibinu, "Breast cancer histopathology image classification with deep convolutional neural networks," in *Proc. 14th Int. Conf. on Electronics Computer and Computation (ICECCO)*, 2019. [Online]. Available: <https://doi.org/10.1109/ICECCO.2018.8634690>.
- [30] A. Pai, "CNN vs. RNN vs. ANN—analyzing 3 types of neural networks in deep learning," *Analytics Vidhya*, 2020. Accessed: Feb. 17, 2020. [Online]. Available: <https://www.analyticsvidhya.com>.
- [31] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, "Transfer learning based histopathologic image classification for breast cancer detection," *Health Inf. Sci. Syst.*, vol. 6, no. 1, 2018. [Online]. Available: <https://doi.org/10.1007/s13755-018-0057-x>.
- [32] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, "A technical review of convolutional neural network-based mammographic breast cancer diagnosis," *Computational and Mathematical Methods in Medicine*, 2019. [Online]. Available: <https://doi.org/10.1155/2019/6509357>.
- [33] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, vol. 9781107057135, 2013. [Online]. Available: <https://doi.org/10.1017/CBO9781107298019>.
- [34] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999. [Online]. Available: <https://doi.org/10.1023/A:1018628609742>.
- [35] A. Smola and S. Vishwanathan, *Introduction to Machine Learning*, Cambridge University, UK, vol. 32, no. 34, 2008.
- [36] Scikit-learn developers, "Support Vector Machine," *Scikit-learn Documentation*, 2019. Accessed: Feb. 17, 2020. [Online]. Available: <https://scikit-learn.org/stable/modules/svm.html>.
- [37] J. Brownlee, "A gentle introduction to k-fold cross-validation," *Statistics*, 2018. [Online]. Available: <https://machinelearningmastery.com>.
- [38] O. Theobald, *Machine Learning for Absolute Beginners*, News.Ge, vol. 4, no. 1, 2017.
- [39] S. Saud, B. Jamil, Y. Upadhyay, and K. Irshad, "Performance improvement of empirical models for estimation of global solar radiation in India: A k-fold cross-validation approach," *Sustainable Energy Technol. Assess.*, vol. 40, 2020. [Online]. Available: <https://doi.org/10.1016/j.seta.2020.100768>.
- [40] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, 2017. [Online]. Available: https://doi.org/10.1162/NECO_a_00990.
- [41] I. Namatëvs, "Deep convolutional neural networks: Structure, feature extraction and training," *Inf. Technol. Manage. Sci.*, vol. 20, no. 1, pp. 39–47, 2018. [Online]. Available: <https://doi.org/10.1515/itms-2017-0007>.
- [42] T. Beysolow II, *Introduction to Deep Learning Using R: A Step-by-Step Guide to Learning and Implementing Deep Learning Models Using R*, 2017.
- [43] M. Toğaçar, B. Ergen, and Z. Cömert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Medical Hypotheses*, vol. 135, 2020. [Online]. Available: <https://doi.org/10.1016/j.mehy.2019.109503>.
- [44] B. Karlik and A. V. Olgac, "Performance analysis of various activation functions in generalized MLP architectures of neural networks," *Int. J. Artif. Intell. Expert Syst.*, vol. 1, no. 4, pp. 111–122, 2011.
- [45] P. Karkare, "Convolutional neural networks-simplified," *AI Graduate*, vol. 15, 2019.
- [46] M. A. Wani, F. A. Bhat, S. Afzal, and A. I. Khan, "Basics of supervised deep learning," in *Studies in Big Data*, vol. 57, 2020. [Online]. Available: https://doi.org/10.1007/978-981-13-6794-6_2.
- [47] Z. A. Varaich and S. Khalid, "Recognizing actions of distracted drivers using Inception v3 and Xception convolutional neural networks," in *Proc. 2nd Int. Conf. Advancements in Computational Sciences (ICACS)*, 2019. [Online]. Available: <https://doi.org/10.23919/ICACS.2019.8689131>.
- [48] R. M. Kamble, M. Kokare, G. C. Y. Chan, O. Perdomo, F. A. González, H. Müller, and F. Mériaudeau, "Automated diabetic macular edema (DME) analysis using fine-tuning with inception-resnet-v2 on OCT images," in *Proc. 2018 IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES)*, 2019. [Online]. Available: <https://doi.org/10.1109/IECBES.2018.8626616>.
- [49] A. Milton-Barker, "Inception V3 deep convolutional architecture for classifying acute myeloid/lymphoblastic leukemia," 2019. Accessed: Feb. 17, 2020. [Online]. Available: <https://www.researchgate.net>.
- [50] J. Xie, R. Liu, J. Luttrell, and C. Zhang, "Deep learning based analysis of histopathological images of breast cancer," *Front. Genet.*, vol. 10, no. FEB, 2019. [Online]. Available: <https://doi.org/10.3389/fgene.2019.00080>.
- [51] R. Mirzaei, R. Nopour, Z. Asghari Varzaneh, M. Shafiee, M. Shanbehzadeh, and H. Kazemi-Arpanahi, "Which are best for successful aging prediction? Bagging, boosting, or simple machine learning algorithms?," *Biomedical Engineering Online*, vol. 22, no. 1, p. 85, 2023.
- [52] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.
- [53] J. Mohajon, "Confusion matrix for your multi-class machine learning model," *Towardsdatascience.Com*, 2020. Accessed: Feb. 17, 2020. [Online]. Available: <https://towardsdatascience.com>.
- [54] K. Riehl, M. Neunteufel, and M. Hemberg, "Hierarchical confusion matrix for classification performance evaluation," *arXiv preprint*, arXiv:2306.09461, 2023. [Online]. Available: <https://arxiv.org/abs/2306.09461>.
- [55] M. Salvaris, D. Dean, and W. H. Tok, *Deep Learning with Azure*, 2018. [Online]. Available: <https://doi.org/10.1007/978-1-4842-3679-6>.
- [56] Y. Xiao, Z. Tian, J. Yu, Y. Zhang, S. Liu, S. Du, and X. Lan, "A review of object detection based on deep learning," *Multimedia Tools and Applications*, vol. 79, no. 33–34, 2020. [Online]. Available: <https://doi.org/10.1007/s11042-020-08976-6>.