Biomedical Signal Processing and Control 86 (2023) 105152

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Breast cancer diagnosis from histopathology images using deep neural network and XGBoost

Alireza Maleki [a], Mohammad Raahemi [b], Hamid Nasiri [c,*]

[a] Electrical and Computer Engineering Department, Semnan University, Semnan, Iran
[b] University of Ottawa, Intelligent Data Warehouse Lab and Knowledge Discovery and Data Mining Lab, 800 King Edward Avenue, Ottawa, ON K1N6N5, Canada
[c] Department of Computer Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

ARTICLE INFO

Keywords:
BreakHis
Breast cancer
DenseNet201
Transfer learning
XGBoost

ABSTRACT

Background and Objective: Globally, breast cancer is one of the most common diseases among women. As a result of the disadvantages of manual analysis, computer-aided diagnosis (CAD) systems are being used to detect images because of their time-consuming and trustworthy capability. With deep learning techniques based on image analysis and classification, CAD systems can efficiently classify images.
Methods: This paper proposes methodologies for enhancing the speed and precision of histopathological image classification, which is a challenge for therapeutic measures. We assess three different classifiers and six pre-trained networks. A pre-trained model is used to extract features from images and then feed those extracted features into the extreme gradient boosting (XGBoost) method, which is selected as the final classifier. Our methodology is based on transfer learning and uses histopathological images as input. To evaluate the performance of the proposed method, we use the BreakHis dataset, which presents histopathology images in four magnification levels, i.e., 40X, 100X, 200X, and 400X.
Results and Conclusion: The accuracies achieved by the proposed method in 40X, 100X, 200X, and 400X magnifications are 93.6%, 91.3%, 93.8%, and 89.1%, respectively. After analyzing the accuracy achieved in this study, the final method proposed combines the DenseNet201 model as a feature extractor with XGBoost as a classifier.

1. Introduction

Breast cancer causes so many deaths across the globe, and it makes this disease more challenging among women, even more than lung cancer. Due to this problem, researchers focus on diagnosis and increasing the survival rate [1]. According to the reports, studies have shown that identifying and recognizing breast cancer in its early days can boost survival rates by as much as 80% [2]. Even in 2018, statistics show that 2.1 million women were detected with breast cancer, and 25% were identified with malignant breast cancer [3].

Analyzing microscopic images of cancer by an expert could make it classified whether the tumor is benign or malignant [4]. Even this process has some disadvantages, like human mistakes or the requirement of retaking another scanning experiment. Also, the high number of breast cancer cases makes diagnosis much more complicated for radiologists, pathologists, and surgeons [5]. Even if the traditional diagnosis procedure has improved in terms of accuracy, there are still risks for the medical team [6]. Furthermore, other reasons such as humanity's mistakes, time-consuming, and ineffective human efforts are involved. According to these problems, increasing doctors' diagnostic accuracy is vital by applying deep learning computer-aided diagnosis (CAD) systems [7]. To get a better diagnosis, medical image analysis offers systems for diagnosing and treating various diseases, including breast cancer [8]. In order to manually omit cancer detection, we propose a technique to diagnose the type of cancer with high precision and accuracy in just a moment to increase the speed and accuracy of detection. On this occasion, the use of computer-aided diagnosis (CAD) equipment in the automated classification of pathological images is shown off and helps to increase the precision and effectiveness of illness diagnosis along with the understanding of the process of disease progression [3]. Besides, the usage of deep learning (DL) and machine learning (ML) algorithms turn up to simply help us analyze histopathological images. In a nutshell, One of the most excellent methods for identifying tumors and infections brought on by different diseases is to employ deep learning and machine vision [9]. So, the future CAD system consists of these algorithms.

* Corresponding author.
E-mail address: h.nasiri@aut.ac.ir (H. Nasiri).

https://doi.org/10.1016/j.bspc.2023.105152
Received 24 November 2022; Received in revised form 18 March 2023; Accepted 8 June 2023

Our main purpose is to classify images into benign and malignant, while these two classes are the leading labels of the BreakHis dataset. Since DL performed well in classification and many uses of DL methods are published, our study tends to create a model to increase the performance in classifying. This study aims to improve the accuracy of the methods provided compared to previous research. This paper presents a method using deep learning algorithms consisting of pre-trained models for extracting image features and boosting methods to classify them and then measuring the accuracy of the classification. Additionally, this paper aims to answer the following study questions: (1) How can a pre-trained feature extractor be used? (2) How well does boosting perform in classification tasks? (3) How does the provided model perform in detecting breast cancer?

In this paper, we discuss the following topics. In Section 2, we introduce seven pre-trained models and review some other papers in this field. A detailed investigation is conducted around the final proposed method and its architecture in Section 2, followed by fine-tuning parameters in Section 3, titled materials and method. We present the results of the empirical experiments in the fourth Section and illustrate them. Section 4 summarizes the findings and suggests the next steps.

2. Related works

There has been great interest by the research community in recent years regarding breast cancer diagnosis from histopathological images. Spanhol et al. [10] released the largest labeled publicly available dataset, comprising 7909 images of breast cancer including both benign and malignant classes [10]. Since then, many researchers and practitioners have studied this dataset in order to develop automated and reliable approaches to discriminate between these two types of cancers using histopathological images. Our aim in the following section is to conduct evaluative comparisons of recently published studies that investigated the BreakHis dataset for breast cancer classification.

In the early days of breast cancer pathological image classification, mass spectrometry and other analytical methods were primarily used to classify images into cancerous and noncancerous categories. A wide range of computer vision tasks was subsequently solved using DL methods. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are two of the most important DL methods. The use of CNNs has been widely adopted in the classification of pathological images. The paper published by Deniz et al. [11] analyzed the classification methods of deep feature extraction and transfer learning for detecting breast cancer using the BreakHis dataset. Two well-known deep CNN architectures, such as the AlexNet and VGG16 models, were employed for deep feature extraction. Three experimental works were considered. The first one involved extracting and then concatenating the feature vectors from the layer before the last of the AlexNet and VGG16 models. In the second experiment, the authors performed feature extraction using the last layer of both previously mentioned models and then combined the feature vectors. In these two experiments, the Support Vector Machine (SVM) came up to classify images into two classes. The final experiment improved performance since AlexNet was tuned on BreakHis images and brought accuracies as follows: 90.96%, 90.58%, 91.37%, and 91.30% for magnification factors of 40X, 100X, 200X, and 400X, respectively [11].

In other research by Yan et al. [12], the benefits of combining CNNs and RNNs were examined. In the first step, since collected high-resolution images were tested, each image was divided into smaller portions. The CNN and the RNN were used for feature extraction and feature fusion of images, respectively. The results showed the suggested model for four-class classification acquired 90.5% average accuracy [12]. Same authors [13] suggested a novel method consisting of a hybrid convolutional plus RNN to classify breast cancer images back in 2020. Their method integrates the advantages of CNNs and RNNs based on the richer multilevel feature representation of histopathological image patches, keeping the long-term and short-term spatial correlations between patches intact. The suggested method used a combination of Inception-V3, fine-tuning, and richer multilevel features. This structure helped them to reach 91.3% average accuracy for their four-class classification method [13].

Sudharshan et al. [14] studied the applicability of Multiple Instance Learning (MIL) for CAD of breast cancer patients based on the analysis of histopathology images and offered a weakly supervised learning framework. Their MIL included grouping instances as images into bags as the patient, with no requirement to label every instance. The research was done using the publicly available BreakHis dataset. The authors inspected some of the cutting-edge models, such as APR, Diverse Density, MI-SVM, and citation-KNN, and then more modern works like a non-parametric method and deep learning-based approach (MIL-CNN). Eventually, the results demonstrated that the non-parametric MIL and MILCNN, which were recently presented, are very effective for the tasks of patient and image classification. The accuracy values achieved were 92.1%, 89.1%, 87.2%, and 82.7% for magnification factors of 40X, 100X, 200X, and 400X, respectively [14]. Li et al. [3] employed an analysis of the 'BreakHis' dataset. Based on deep features, their study suggested a modern strategy for classifying benign and malignant breast cancer on three levels. At the data pre-processing level, the authors designed Sliding + Random and Sliding + Class Balance Random window slicing methods. Two of these procedures increased model generalization and classification performance. After that, the AlexNet model was used for the feature extraction level. Eventually, Different ML models were given from different levels of features to classify data, and the optimal combination was selected. The integration of intermediate- and high-level features with SVM produced the best classification results when characteristics of various levels were mixed with an ML classifier during the stage of deep feature classification. Sliding + Class Balance Random window slicing was the appropriate data pre-processing method according to model performance. The best classification result was achieved when intermediate and high-level features were combined with SVM. At various magnifications, the classification accuracy varied from 85.30% to 88.76% [3].

In another study by Zerouaoui & Idri [8], the results of an empirical comparison of 28 hybrid architectures came out using SVM, Multilayer perceptron, K-Nearest Neighbors (KNN), and Decision Tree all as classifiers, and then DL methods came to help as feature extractors, such as DenseNet201, MobileNet V2, ResNet 50, Inception V3, ReseNetV2, VGG16, and VGG19, to classify breast cancer images. Over two datasets, BreakHis, which consists of four magnification factors, and FNAC, the experimental evaluations (accuracy, precision, recall, and SK test) were calculated. Using the KNN classifier and DenseNet201 called KDEN for the BreakHis dataset, the third-ranked hybrid architectures had accuracy values of 83.35%, 84.82%, 83.27%, and 80.56% for the magnification factors of 40X, 100X, 200X, and 400X, respectively. A significant effect was observed in the accuracy outcomes of hybrid architectures derived from DL techniques for feature extraction and classifier development. The results showed that DenseNet201 was the best performer for hybrid architectures with magnification factors of 40X, 100X, 200X, and 400X, with accuracy scores of 92.61%, 92%, 93.93%, and 91.73%, respectively [8].

Recent research has not only focused on DL models and modern classifiers. According to Sharma & Mehra [15], magnification factors have an impact on the selection of appropriate layers for fine-tuning depending on the depth of the pre-trained network for fine-tuning. Based on their study, fine-tuning impacts image magnification differently. The most effective results were obtained with moderate fine-tuning both for binary and multi-classification at a 40X magnification factor. As a practical method to attain the best performance in the classification of histopathological images as well as for other computer vision-related applications, layer-wise fine-tuning might be recommended, according to the results. Compared to the shallow and deep tuning of the pre-trained network, which depends on the size and distribution of a dataset, a moderate fine-tuning level is ideal for classification images of histology at different magnification levels. By combining a pre-trained "AlexNet" model with appropriate fine-tuning methods, the BreakHis dataset was classified. Furthermore, they calculated the effective depth of fine-tuning for four distinct magnification levels: 40X, 100X, 200X, and 400X, with accuracy scores of 89.31%, 85.75%, 83.95%, and 84.33%, respectively [15].

3. Materials and methodology

3.1. Dataset

Images of benign and malignant breast cancers taken during microscopic biopsies can be found in the BreakHis dataset [10]. Images were gathered from January to December 2014 as part of a clinical investigation. Through this process, the patients were invited to participate in this volunteer experiment and recourse to the P&D lab in Brazil. The lab provides a dataset that contains 7909 breast cancer histopathology images gathered from 82 patients. The dataset is divided into two main type of cancers, benign and malignant, and there are subclasses such as Adenosis, Fibroadenoma, Phyllodes Tumor, and Tubular Adenoma for benign tumors and some subtypes for malignant like Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma. The images in the dataset are gathered in four different magnification levels: 40X, 100X, 200X, and 400X. The benign label consists of 2489 images; on the other side, 5429 samples are provided for malignant tumors. The original images' dimensions are 700 460 pixels. Spanhol et al. [10] published this dataset in 2015, which is available to the public [10]. Table 1 provides a better view of this dataset.

3.2. Methodology

This study identifies benign and malignant breast cells that indicate the absence and presence of tumors in the given image. Our proposed method consists of several distinct phases, as illustrated in Fig. 1. To begin with, using all images in the BreakHis dataset, all images with different magnifying rates are extracted and grouped into training (i.e., 70% of all images) and testing (i.e., 30% of all images). Following pre-processing of training images, the images are fed into six pre-trained feature extraction models, including VGG16, VGG19, ResNet50, DenseNet201, DenseNet169, and DenseNet121, which are implemented by the Keras open-source library. Images in the test set were subjected to the same pre-processing phases. In the end, the extracted features were passed to gradient boosting methods (i.e., XGBoost, LightGBM, and CatBoost), and all their parameters were tuned and optimized accordingly. In order to fine-tune transfer learning models, the top layer is changed to classify photos based on binary classifications. Using the final constructed models, the test images are classified. Classification results are represented in binary classes labeled "Benign" or "Malignant".

3.3. Data pre-processing

The following subsection describes how the chosen datasets were pre-processed in this study.

Using different pre-processing techniques is the first and most important step of the data preparation process [16]. The experimental dataset was divided into training and testing sets using binary labels. Classes 0 and 1 correspond to benign and malignant classes, respectively. Initially, the images were 700 pixels by 460 pixels in size. The images were resized to 224 pixels by 224 pixels to be compatible with deep neural networks. Two lists are eventually created to store images and their labels. In order to improve speed, lists are transformed into NumPy arrays.

Table 1
Distributions of images in BreakHis dataset.

Magnification	Benign	Malignant	Total
40X	625	1370	1995
100X	644	1437	2081
200X	623	1390	2013
400X	588	1232	1820
Total	2480	5429	7909
Number of Patients	24	58	82

3.4. Feature extractors

Extracting features is a process of transforming the initial data into usable information. The main purpose of feature extraction is to reduce the amount of data that must be processed and adequately characterize the initial data simultaneously. In image processing, feature extractors can be useful to find properties, including shapes, edges, and movements. Feature extraction is a beneficial technique when fewer resources are required for processing without losing the key or essential data [17]. The model or a subset of the model may pre-process the input to produce an output (i.e., a vector) for each input image, which may then be used as input for training a new model [18].

We applied six types of deep learning pre-trained models for data extraction, including VGG16, VGG19, ResNet50, DenseNet201, DenseNet169, and DenseNet121. Two essential parameters need to be initialized: type of pooling and including or excluding the last layer. In the following part, we aim to introduce these pre-trained models used in this paper briefly.

VGG16, VGG19: Take 224 × 224 RGB images as input. VGG16 (Visual Geometry Group) is expected to have a better performance concerning to image classification and visualization [19]. These VGGs are considered powerful CNNs. These models use small receptive fields (3 × 3 with a stride of (1), padding, and 2 × 2 max-pooling filters with the same stride. The output is provided by three fully connected layers (FC) and a softmax layer. A VGG16 model consists of 16 layers, while a VGG19 model consists of 19 layers. All hidden layers contain rectified linear units (ReLUs) non-linearity.

Resnet 50: ResNet50 has 48 Convolution layers along with one max pooling and one average pooling layer. The default input size that ResNet50 takes is 224 × 224. ResNet's architecture is inspired by VGG, which has 3 × 3 filters that adhere to two simple design rules: (1) the number of filters should be the same across the layers for a similar output feature map size, and (2) to maintain the time complexity per layer the number of filters should be doubled.

DenseNet201, Densenet121, and Densenet169: Densnet201 is similar to ResNets, although, in DenseNet201 architecture, each layer receives inputs from all previous layers and outputs feature maps [20]. A dense block consists of batch normalization, ReLU activation, and 3 × 3 convolution. A transition layer comprises of batch normalization, a 1 × 1 convolution, and average pooling. The main difference between DenseNet121 and DenseNet169 is the number of layers each model has. Each of these pre-trained models includes four dense blocks with different numbers of layers as follows: DenseNet121 has 6-12-24-16 layers in four dense blocks, while Densenet169 consists of 6-12-32-32 layers.

The first step in the process is choosing the type of pooling: To smooth out images in the BreakHis dataset, average pooling was used since the histopathological images contain a lot of edges and sharp features. For example, consider the situation where cancer is only visible in a portion of the image. In this situation, the parts of the pooling region that match the background pixels will tend to dominate the pooled representation, so average pooling might not be the most effective method. However, the average pooling method may be more appropriate in other cases, such as classifying abnormal images from

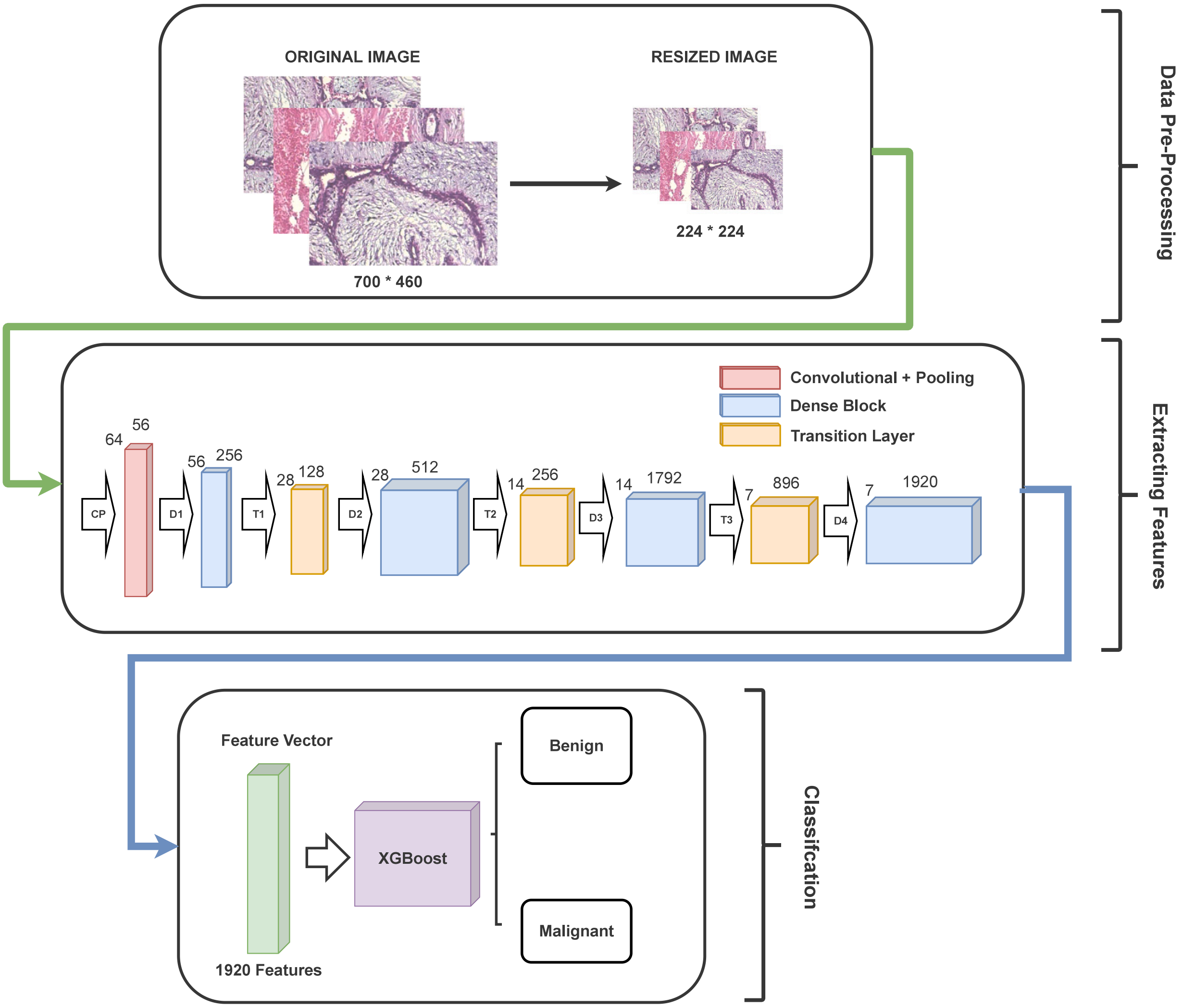


Fig. 1. Proposed method architecture.
Source: The original and resized images extracted from the BreakHis dataset [10].

normal ones when the abnormalities are distributed throughout [21]. In the next step, we exclude the last layer, which is known as the prediction layer. In a model without the prediction layer, the final convolutional or pooling layer's activations are outputted directly [18]. This is followed by specifying the input of the shape, which takes 224 × 224x3 images. In the end, an array of numbers is fed into the classifiers as input. Without sufficient training examples, CNN will become overfit and lose its ability to generalize [22].

3.5. *Classifiers*

Despite artificial neural networks recently regaining popularity, boosting methods became more useful for medium datasets since training times are fast and tuning parameters are not time-consuming. Boosting methods combine various weak classifiers to create a more accurate classifier. This is done by dividing the training data. Then each part trains different models or one model with a different setting. In the end, the results are combined [23]. This paper benefits from three gradient boosting algorithms developed on decision trees, which have become increasingly popular.

3.5.1. *XGBoost*

In 2016, Chen et al. [24] provided a system with high scalability known as XGBoost. This algorithm is an implementation of Gradient Boosted Decision Tree (GBDT) [25], which is provided by Zhang & Haghani [26]. Compared to other boosting decision trees, the most distinguishing feature of XGBoost is its scalability. Other published methods are ten times slower than this one. Despite its fast-learning capabilities, the classifier can overfit the data. XGBoost's regularization technique prevents overfitting, differentiating from other gradient-boosting algorithms. Consequently, model tuning becomes faster and more robust [26,27]. XGBoost adds a regularization term to the objective function as follows:

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{1}$$

where $L(\theta)$ is a loss function, and $\Omega(\theta)$ is a regularization function that avoids overfitting by controlling the complication of the model [28]. The regularization function is calculated as follows:

$$\Omega(\theta) = \gamma N + \frac{1}{2}\lambda \|w\|^2 \tag{2}$$

Table 2
XGBoost hyperparameters setting.

Parameters	Value
Base learner	Gradient boosted tree
Learning rate (η)	0.5
Number of trees	250
Minimum loss reduction	0

Table 3
LightGBM hyperparameters setting.

Parameters	Value
Number of iterations of the algorithm	200
Learning rate (η)	0.5
max number of bins	100

Table 4
Binary classification accuracy score in the BreakHis dataset.

Feature extractor	Method	Result (Accuracy%)				
		40X	100X	200X	400X	Average
VGG16 [19]	XGBoost	88.1	89.4	87.7	86.8	88
	LightGBM	89.1	89.1	86.7	86.2	87.7
	CatBoost	86.6	88	84.9	84	85.8
VGG19 [19]	XGBoost	88.4	88.3	88.5	84.2	87.3
	LightGBM	90.1	88.3	88.2	85.8	88.1
	CatBoost	87.8	88.1	84.9	84.4	86.3
ResNet50 [36]	XGBoost	91.3	89.6	91.8	89.5	90.5
	LightGBM	91.8	90.2	92.5	**89.7**	91
	CatBoost	88.4	85.9	89.2	86.4	87.4
DenseNet201 [20]	XGBoost	**93.6**	91.3	**93.8**	89	**91.93**
	LightGBM	93	**92.4**	93	89	91.85
	CatBoost	92.3	89.2	91.5	86.4	89.85
DenseNet169 [20]	XGBoost	91.3	89.9	92.7	89.5	90.85
	LightGBM	92.3	91.2	93.5	89	91.5
	CatBoost	88.9	89.1	89.4	86.4	88.4
DenseNet121 [20]	XGBoost	90.3	89.9	90.5	87.5	89.55
	LightGBM	91.8	90.2	90.8	88.8	90.4
	CatBoost	89.3	90.2	88.9	86.4	88.7

Where N denotes the number of leaf nodes in the decision tree and w represents the node's weights [29]. Regularization parameters are γ and λ which specify the penalty limit related to the decision tree. One important step is setting XGBoost hyperparameters value, a crucial key feature over other machine learning techniques that enables better tuning [30].

In order to obtain satisfactory results, the grid search technique was utilized to determine optimal hyperparameter values for the XGBoost algorithm. Three crucial hyperparameters were selected for initialization and evaluation based on their effectiveness. Each hyperparameter was assigned a range of values to search for the optimal value. The first hyperparameter, namely the "learning rate", which controls the weights of newly added trees in the model and helps prevent overfitting, was fine-tuned with a grid search approach. Specifically, a range of values from 0.1 to 0.9 with a step size of 0.1 was tested, and the grid search algorithm selected 0.5 as the best value. The second hyperparameter, "the number of gradient boosted trees", was initialized to 250 and was fine-tuned by grid search over a range of values from 150 to 350 with a step size of 50. Lastly, the "minimum loss reduction" hyperparameter was fine-tuned to obtain better algorithm control. The grid search algorithm selected a value of 0, within the range of integers from 0 to 10. The selected hyperparameter values for XGBoost are presented in Table 2.

3.5.2. LightGBM

In April 2017, Microsoft developed the light gradient boosting machine (LightGBM) to reduce implementation time. Unlike other decision trees, LightGBM grows decision trees leaf by leaf, instead of checking each new leaf against all others [31,32]. Since LightGBM generates more complex trees compared to other boosting methods, it is known as a precise and reliable decision tree boosting algorithm [33]. According to experiments done on some available datasets, it turns out that LightGBM can accelerate the training process of gradient boosting decision trees up to 20 times while maintaining similar accuracy. Followed by using LightGBM, tuning the hyperparameters is required to get a better trade-off between speed and accuracy. Gradient-based One-Side Sampling and Exclusive Feature Bundling are two unique techniques in LightGBM that are used to handle vast data instances and features, respectively [31]. Once again, we used grid search to set the best values for LightGBM parameters. Table 3 summarizes the LightGBM hyperparameters setting.

3.5.3. Catboost

CatBoost (for "categorical boosting") uses permutation techniques, one hot max size (OHMS), and target-based statistics to boost categorical columns. The greedy method is used by CatBoost to solve the exponential growth of feature combinations at every split of the tree [34]. The CatBoost process uses these steps for features with more categories than OHMS (an input parameter): Dividing, converting, and transforming [35].

4. Results and discussion

4.1. Evaluation criteria

We applied a cross-validation approach to evaluate the performance validity and verify the model's outcome. A variety of metrics, including accuracy, precision, recall, specificity, and F_1-Score, are used in our proposed approach to evaluate classification efficiency. The F_1-Score is a measure of a model's accuracy that takes into account both precision and recall. It is commonly used in machine learning and information retrieval to evaluate the performance of a classification model. The F_1-Score is calculated as the harmonic mean of precision and recall, giving equal weight to both measures. A high F_1-Score indicates that the model has high precision and recall, which means it can correctly identify the most relevant instances while minimizing false positives. Accordingly, the mentioned metrics' formulas are as follows:

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F_1 - Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{7}$$

where *TP* represents the number of correct malignant predictions, *TN* stands for the number of correct benign predictions, *FP* indicates the number of incorrect malignant predictions, and *FN* denotes the number of incorrect benign predictions.

4.2. Classification results

The present study examined six pre-trained models using the Keras package to extract features from histopathological images. Then these

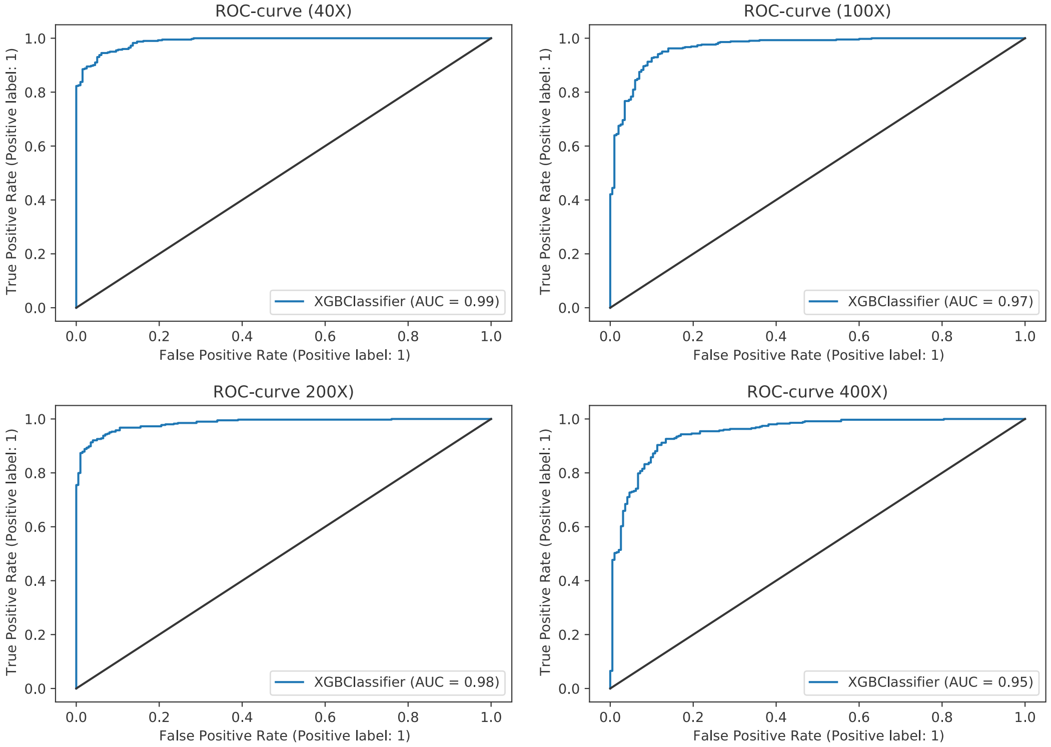


Fig. 2. ROC curves of the proposed method for different magnification factors.

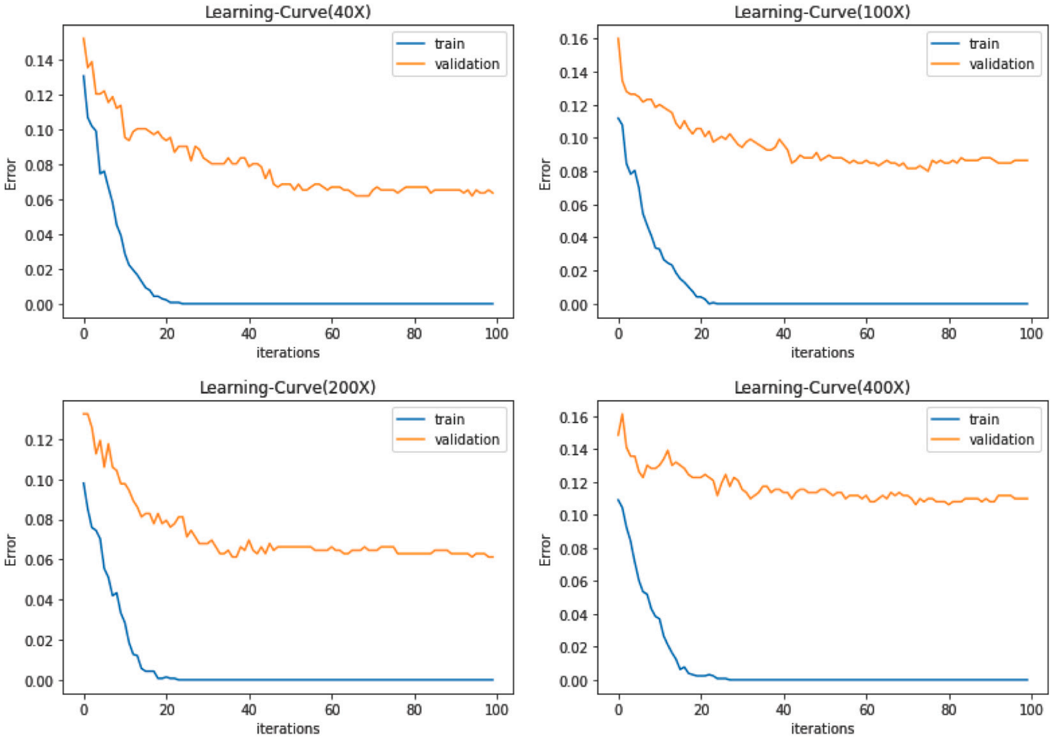


Fig. 3. Learning curves of the proposed method for 40X, 100X, 200X, and 400X magnification factors.

features were passed to three classifiers, i.e., XGboost, LightGBM, and CatBoost, to predict appropriate labels. Table 4 shows the results based on the accuracy score.

As shown in Table 4, the model, which consists of DenseNet201 as a feature extractor and XGBoost as a classifier, achieved 93.8% accuracy. A total of 18 experiments were conducted for this research. Obtained

Table 5
Evaluation criteria for the proposed method.

Magnification	Accuracy%	Precision	Recall	Specificity	F_1-Score
40X	93.6	0.921	0.99	0.828	0.954
100X	91.3	0.907	0.971	0.79	0.938
200X	93.8	0.942	0.967	0.88	0.954
400X	89	0.888	0.948	0.783	0.917
Average	91.9	0.915	0.969	0.82	0.941

results based on accuracy led us to calculate additional evaluation metrics in addition to the accuracy Table 5. Fig. 2 shows ROC curves of the proposed method for different magnification factors. Learning curves of the proposed method for 40X, 100X, 200X, and 400X magnification factors are shown in Fig. 3. Moreover, the outputs of the Grad-CAM algorithm are presented in Table 6.

4.3. Discussion

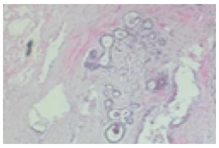
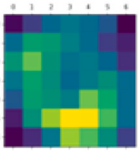
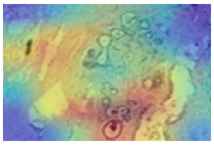
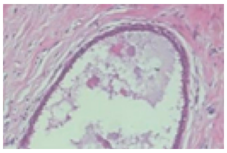
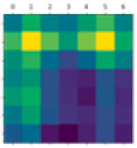
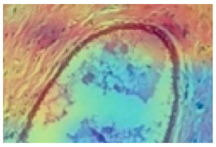
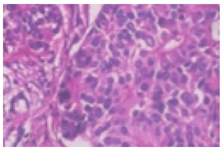
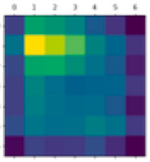
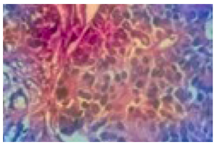
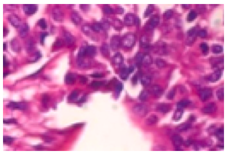
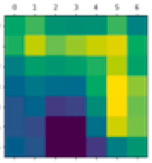
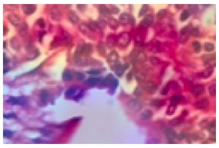
Increasingly, histopathological images are considered highly important in the real world. Our priority is to build systems that minimize human errors and time-consuming processes. Using our experiments, we recommended a model that combines a pre-trained DenseNet201 model and XGBoost as a classifier. We believe that this combination can be considered a better computer-aided design system (CAD). CAD systems are developed and compete to eliminate manual analysis and reduce the problems of other systems. The dataset architecture underwent no specific changes, including augmentation or image pre-processing. Considering Table 4, our proposed model obtained accuracies in the range of 89.0% to 93.8% with all magnification levels showing it has a low dependency on the magnification factor. In Table 6, we illustrated the heatmap of every magnification level with both benign and malignant images using the Grad-CAM algorithm that shows which parts of the images are sensitized by the models and which part was less noticed. Table 5 shows the other metrics calculated for the final proposed method, consisting of precision, sensitivity, specificity, and F1-score. Besides its high accuracy, the chosen approach has 94.2% precision on a 200X magnifier. Also, our proposed method achieved 99% sensitivity (Recall) on a 40X magnifier, while high recall can be seen at other magnification levels. The high F1-score achieved by our proposed model shows the harmonic between recall and precision regarding unbalanced data in the dataset. As observed from Table 4 and Table 5, the outcomes obtained from 400X magnification are comparatively inferior to those of the other resolutions. This can be attributed to the excessive magnification at the 400X level, resulting in the possible loss of certain discriminative features that aid in the identification of cancer classes. Consequently, the accuracy of the resolution at this level is decreased.

For the performance analysis of various classifiers, we have calculated learning curves in Fig. 3 and ROC curves in Fig. 2. We can understand well how our model truly predicts images. As shown, area under the curve (AUC) values are 0.99, 0.98, 0.97, and 0.95 for 40X, 100X, 200X, and 400X magnification factors, respectively. A receiver operating characteristic curve (ROC) is used to evaluate the classifier's ability to predict a class. Higher values AUC, indicate better performance of the classifier. Fig. 2 illustrates the ROC curve for XGBoost classifier performance in predicting the true positive (malignant) class. The results represent how the chosen classifier performs accurately in predicting. The ROC-Curve visualizes the AUC value. As shown, the curve tends to the True positive rate (TPR) axis, a vertical axis in ROC-curve diagrams. From Table 4, it is clearly visible that the performance of DenseNet201-XGBoost is quite significant, whereas VGG16-CatBoost has shown poor performance on the same dataset.

Fig. 3 represents how well a model learns per iteration based on errors. A learning curve shows how a model's learning performance evolves with time. During training, the model can be evaluated on the training and holdout validation dataset, and learning curve plots can be created using the measured performance. Fig. 3 shows the learning curve for 100 epochs. Eventually, the training curve stabilized after 25 epochs which is known as the generalization gap.

Fig. 4 shows a two-dimensional t-SNE projection of image patches, where each patch represents an image belonging to one of the classes. Two sets of diagrams for all magnifications are shown in the figure, left ones showing the patches before and the right ones showing them after

Table 6
Grad-CAM outputs consist of Heat-map for benign and malignant cancer in 40X, 100X, 200X, and 400X magnification factors. Source of original images: [10].

	MF	Original image	Heat-map	Grad-CAM output
Benign	**40X**			
	100X			
Malignant	**200X**			
	400X			

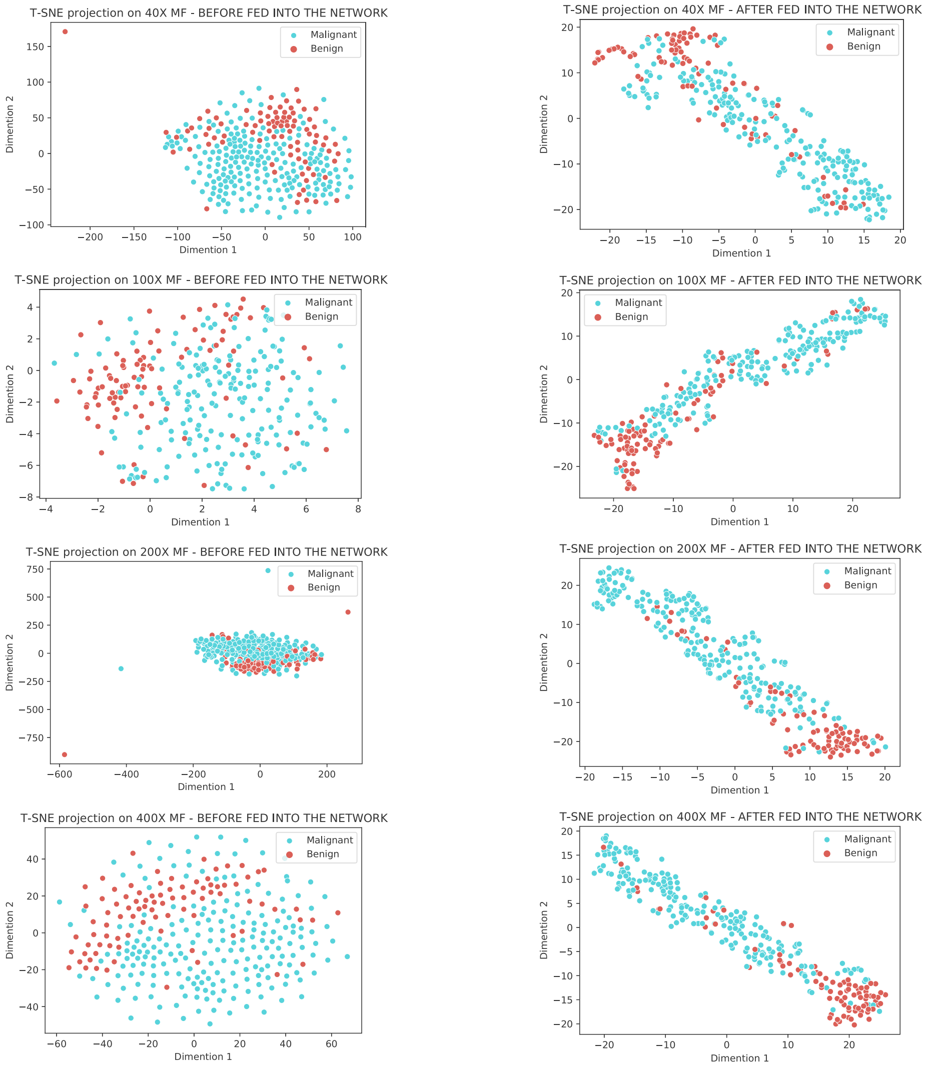


Fig. 4. t-SNE projection on 300 images in BreakHis dataset before and after extracting features by the DenseNet201 model.

DenseNet201 is applied to extract the features. The right diagrams show how well patches are clustered within the same class after extracting features and training a model on the dataset. Through the t-SNE technique, the dimensions of the samples are reduced and embedded. Due to better visualization of clustering on both classes, the experiment is conducted on 300 samples from the dataset and the Euclidean metric is chosen in order to measure the similarity between the samples.

Eventually, XGBoost has the best performance among the classifiers with the highest average accuracy (i.e., 91.93%), precision (i.e., 91.5%), recall (i.e., 96.9%), and F1-score (94.1%). LightGBM is the second-best classifier with 91.85% accuracy. The worst average accuracy achieved through three classifiers is scored by CatBoost and through pre-trained models used is VGG16 with 85.8%. As can be seen from Table 7, our proposed methodology outperformed other state-of-the-art methods. Fig. 5 illustrates how well our proposed method performs in classification using confusion matrices.

Table 7
Performance comparison with state-of-the-art counterparts.

Method	Accuracy %				
	40X	100X	200X	400X	Average
Proposed Method	**93.6**	**91.3**	**93.8**	89.1	**91.9**
(Deniz et al. 2018) [11]	90.9	90.5	91.3	**91.3**	91
(Yan et al.,2018) [12]	–	–	–	–	90.5
(Yan et al.,2020) [13]	–	–	–	–	91.3
(Sudharshan et al. 2019) [14]	92.1	89.1	87.2	82.7	87.7
(Li et al. 2021) [3]	87.8	86.6	87.7	85.3	86.8
(Sharma & Mehra, 2020) [15]	89.3	85.7	83.9	84.3	85.5
(Joseph et al. 2022) [37]	90.9	89.6	91.6	88.7	90.1

5. Conclusion

This study proposes a method where features are extracted by a pre-trained feature extractor, and then the extracted features are concatenated to a boosting method to classify images into benign and malignant tumors. In total, 18 various architectures were evaluated using six pre-trained models and three classifiers. The experiments were conducted over the BreakHis dataset, which contains histopathology images at magnifications of 40X, 100X, 200X, and 400X. By using grid search algorithm, we could fine-tune the classifiers' parameters. Using a pre-trained DenseNet201 model and XGBoost classifier, the proposed model achieved 93.6%, 91.3%, 93.8%, and 89.0% for magnifications of 40X, 100X, 200X, and 400X, respectively. In addition to the accuracy

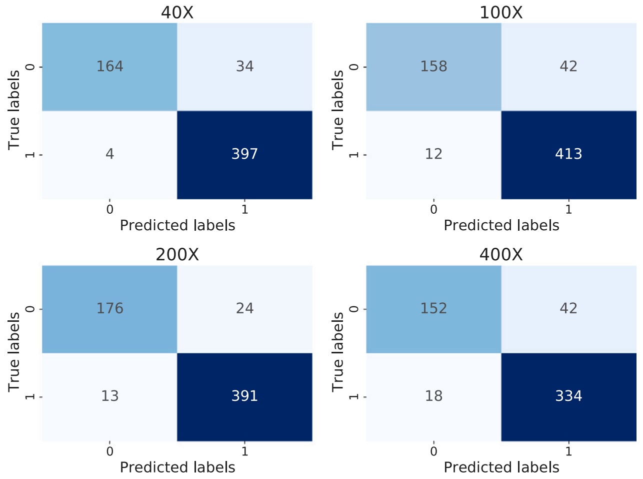


Fig. 5. Confusion matrix of the proposed method for 40X, 100X, 200X, and 400X magnification levels.

score, four metrics are considered for empirical evaluations: Precision, recall (sensitivity), specificity, and F1-score. The proposed model not only achieves a satisfactory classification for minority class (benign) instances but also demonstrates a more promising prediction for majority class instances (malignant). Based on these results, the proposed model is viable for providing definitive opinions on benign and malignant cases. However, this study has some limitations, including the absence of stain normalization. To maximize the efficiency of binary classification in the future, we will attempt to use augmentation techniques. Since the proposed model is implemented for binary classification, in the future, we will test our suggested methodology on the subclasses released recently via BreakHis dataset such as ductal carcinoma, etc.

Data and code availability

A publicly available dataset, i.e., BreakHis, was used in this study, which is available at BreaKHis.[1] In addition, The source code of the proposed method required to reproduce the predictions and results is available at Github.[2]

CRediT authorship contribution statement

Alireza Maleki: Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Mohammad Raahemi:** Software, Validation, Writing – review & editing. **Hamid Nasiri:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data and code at "Data and Code Availability" section of the manuscript

[1] http://www.inf.ufpr.br/vri/databases/BreaKHis_v1.tar.gz

[2] https://github.com/allirezamaleki/BreastCancerDiagnosis

References

[1] Taye Girma Debelee, Friedhelm Schwenker, Achim Ibenthal, Dereje Yohannes, Survey of deep learning in breast cancer image analysis, Evol. Syst. 11 (1) (2020) 143–163.

[2] Thomas H.F. Tsang, K.H. Wong, Kate Allen, Karen K.L. Chan, Miranda C.M. Chan, David V.K. Chao, A.N. Cheung, Cecilia Y.M. Fan, Edwin P. Hui, Dennis K.M. Ip, et al., Update on the recommendations on breast cancer screening by the cancer expert working group on cancer prevention and screening, Hong Kong Med. J. 28 (2022) 161–168.

[3] Xin Li, HongBo Li, WenSheng Cui, ZhaoHui Cai, MeiJuan Jia, Classification on digital pathological images of breast cancer based on deep features of different levels, Math. Probl. Eng. 2021 (2021).

[4] William Al Noumah, Assef Jafar, Kadan Al Joumaa, Using parallel pre-trained types of DCNN model to predict breast cancer with color normalization, BMC Res. Notes 15 (1) (2022) 1–6.

[5] Gensheng Zhang, Wei Wang, Jucheol Moon, Jeong K. Pack, Soon Ik Jeon, A review of breast tissue classification in mammograms, in: Proceedings of the 2011 ACM Symposium on Research in Applied Computation, 2011, pp. 232–237.

[6] Abhijit Bhattacharyya, Divyanshu Bhaik, Sunil Kumar, Prayas Thakur, Rahul Sharma, Ram Bilas Pachori, A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images, Biomed. Signal Process. Control 71 (2022) 103182.

[7] Hua Li, Shasha Zhuang, Deng-ao Li, Jumin Zhao, Yanyun Ma, Benign and malignant classification of mammogram images based on deep learning, Biomed. Signal Process. Control 51 (2019) 347–354.

[8] Hasnae Zerouaoui, Ali Idri, Deep hybrid architectures for binary classification of medical breast cancer images, Biomed. Signal Process. Control 71 (2022) 103226.

[9] Mohammad Rahimzadeh, Abolfazl Attar, Seyed Mohammad Sakhaei, A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset, Biomed. Signal Process. Control 68 (2021) 102588.

[10] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, Laurent Heutte, A dataset for breast cancer histopathological image classification, IEEE Trans. Biomed. Eng. 63 (7) (2015) 1455–1462.

[11] Erkan Deniz, Abdulkadir Şengür, Zehra Kadiroğlu, Yanhui Guo, Varun Bajaj, Ümit Budak, Transfer learning based histopathologic image classification for breast cancer detection, Health Inf. Sci. Syst. 6 (1) (2018) 1–7.

[12] Rui Yan, Fei Ren, Zihao Wang, Lihua Wang, Yubo Ren, Yudong Liu, Xiaosong Rao, Chunhou Zheng, Fa Zhang, A hybrid convolutional and recurrent deep neural network for breast cancer pathological image classification, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2018, pp. 957–962.

[13] Rui Yan, Fei Ren, Zihao Wang, Lihua Wang, Tong Zhang, Yudong Liu, Xiaosong Rao, Chunhou Zheng, Fa Zhang, Breast cancer histopathological image classification using a hybrid deep neural network, Methods 173 (2020) 52–60.

[14] P.J. Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, Paul Honeine, Multiple instance learning for histopathological breast cancer image classification, Expert Syst. Appl. 117 (2019) 103–111.

[15] Shallu Sharma, Rajesh Mehra, Effect of layer-wise fine-tuning in magnification-dependent classification of breast cancer histopathological image, Vis. Comput. 36 (9) (2020) 1755–1769.

[16] Muhammad Imran Razzak, Saeeda Naz, Ahmad Zaib, Deep learning for medical image processing: Overview, challenges and the future, Classif. BioApps (2018) 323–350.

[17] Samina Khalid, Tehmina Khalil, Shamila Nasreen, A survey of feature selection and feature extraction techniques in machine learning, in: 2014 Science and Information Conference, IEEE, 2014, pp. 372–378.

[18] Jason Brownlee, Transfer learning in keras with computer vision models, Mach. Learn. Mastery (2020).

[19] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[21] Rajendran Nirthika, Siyamalan Manivannan, Amirthalingam Ramanan, Ruixuan Wang, Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study, Neural Comput. Appl. (2022) 1–27.

[22] Pin Wang, Jiaxin Wang, Yongming Li, Pufei Li, Linyu Li, Mingfeng Jiang, Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing, Biomed. Signal Process. Control 65 (2021) 102341.

[23] Essam Al Daoud, Comparison between xgboost, lightgbm and CatBoost using a home credit dataset, Int. J. Comput. Inf. Eng. 13 (1) (2019) 6–10.

[24] Tianqi Chen, Carlos Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[25] Hamid Nasiri, Ghazal Kheyroddin, Morteza Dorrigiv, Mona Esmaeili, Amir Raeisi Nafchi, Mohsen Haji Ghorbani, Payman Zarkesh-Ha, Classification of COVID-19 in chest X-ray images using fusion of deep features and LightGBM, in: 2022 IEEE World AI IoT Congress (AIIoT), IEEE, 2022, pp. 201–206.

[26] Yanru Zhang, Ali Haghani, A gradient boosting method to improve travel time prediction, Transp. Res. C 58 (2015) 308–324.

[27] Saeed Chehreh Chelgani, Hamid Nasiri, Arash Tohry, H.R. Heidari, Modeling industrial hydrocyclone operational variables by SHAP-CatBoost - A "conscious lab" approach, Powder Technol. 420 (2023) 118416, http://dx.doi.org/10.1016/j.powtec.2023.118416.

[28] Rasoul Fatahi, Hamid Nasiri, Arman Homafar, Rasoul Khosravi, Hossein Siavoshi, Saeed Chehreh Chelgani, Modeling operational cement rotary kiln variables with explainable artificial intelligence methods – a "conscious lab" development, Particul. Sci. Technol. (2022) 1–10, http://dx.doi.org/10.1080/02726351.2022.2135470.

[29] Hamid Nasiri, Seyed Ali Alavi, A novel framework based on deep learning and ANOVA feature selection method for diagnosis of COVID-19 cases from chest X-ray images, Comput. Intell. Neurosci. 2022 (2022) 4694567, http://dx.doi.org/10.1155/2022/4694567.

[30] Rasoul Fatahi, Hamid Nasiri, Ehsan Dadfar, Saeed Chehreh Chelgani, Modeling of energy consumption factors for an industrial cement vertical roller mill by SHAP-XGBoost: a "conscious lab" approach, Sci. Rep. 12 (1) (2022) 1–13.

[31] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).

[32] Mobina Ezzoddin, Hamid Nasiri, Morteza Dorrigiv, Diagnosis of COVID-19 cases from chest X-ray images using deep neural network and lightgbm, in: 2022 International Conference on Machine Vision and Image Processing, MVIP, IEEE, 2022, pp. 1–7.

[33] Hamid Nasiri, Sharif Hasani, Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost, Radiography 28 (3) (2022) 732–738.

[34] Mohammad Reza Abbasniya, Sayed Ali Sheikholeslamzadeh, Hamid Nasiri, Samaneh Emami, Classification of breast tumors based on histopathology images using deep features and ensemble of gradient boosting methods, Comput. Electr. Eng. 103 (2022) 108382.

[35] Anna Veronika Dorogush, Vasily Ershov, Andrey Gulin, CatBoost: gradient boosting with categorical features support, 2018, arXiv preprint arXiv:1810.11363.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[37] Agaba Ameh Joseph, Mohammed Abdullahi, Sahalu Balarabe Junaidu, Hayatu Hassan Ibrahim, Haruna Chiroma, Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer), Intell. Syst. Appl. 14 (2022) 200066.