

Article

Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models

Zabit Hameed ^{1,*}, Sofia Zahia ^{1,†}, Begonya Garcia-Zapirain ¹, José Javier Aguirre ^{2,3} and Ana María Vanegas ⁴

¹ eVida Research Group, University of Deusto, 48007 Bilbao, Spain; sofia.zahia95@deusto.es (S.Z.); mbgarciazapi@deusto.es (B.G.-Z.)

² Biokeralty Reseach Institute, 01510 Vitoria, Spain; josej.aguirre@keralty.com

³ Department of Pathological Anatomy, University Hospital of Araba, 01009 Vitoria, Spain

⁴ Clínica Colsanitas, Bogotá 110221, Colombia; anmvanegas@colsanitas.com

* Correspondence: zabithameed@deusto.es

† These authors share first authorship.

Received: 10 June 2020; Accepted: 3 August 2020; Published: 5 August 2020



Abstract: Breast cancer is one of the major public health issues and is considered a leading cause of cancer-related deaths among women worldwide. Its early diagnosis can effectively help in increasing the chances of survival rate. To this end, biopsy is usually followed as a gold standard approach in which tissues are collected for microscopic analysis. However, the histopathological analysis of breast cancer is non-trivial, labor-intensive, and may lead to a high degree of disagreement among pathologists. Therefore, an automatic diagnostic system could assist pathologists to improve the effectiveness of diagnostic processes. This paper presents an ensemble deep learning approach for the definite classification of non-carcinoma and carcinoma breast cancer histopathology images using our collected dataset. We trained four different models based on pre-trained VGG16 and VGG19 architectures. Initially, we followed 5-fold cross-validation operations on all the individual models, namely, fully-trained VGG16, fine-tuned VGG16, fully-trained VGG19, and fine-tuned VGG19 models. Then, we followed an ensemble strategy by taking the average of predicted probabilities and found that the ensemble of fine-tuned VGG16 and fine-tuned VGG19 performed competitive classification performance, especially on the carcinoma class. The ensemble of fine-tuned VGG16 and VGG19 models offered sensitivity of 97.73% for carcinoma class and overall accuracy of 95.29%. Also, it offered an F1 score of 95.29%. These experimental results demonstrated that our proposed deep learning approach is effective for the automatic classification of complex-natured histopathology images of breast cancer, more specifically for carcinoma images.

Keywords: deep learning; histopathology; breast cancer; image classification; ensemble models

1. Introduction

Cancer is one of the critical public health issues around the world. According to the Global Burden of Disease (GBD) study, there have been 24.5 million cancer incidence and 9.6 million cancer deaths worldwide in 2017 [1]. These statistics indicate that cancer incidence expanded by 33% between 2007 and 2017 worldwide [1]. Specifically, breast cancer is the most common malignancy and the leading cause of cancer-related mortalities among women worldwide [1,2]. Thus, premature diagnosis of this pathology is crucial to preclude its progression and reduce its morbidity rates in women.

Breast cancer is a heterogeneous disease, composed of numerous entities with distinctive biological, histological and clinical characteristics [3]. This malignancy erupts from the growth of abnormal breast cells and might invade the adjacent healthy tissues [3]. Its clinical screening is

initially performed by utilizing radiology images, for instance, mammography, ultrasound imaging, and Magnetic Resonance Imaging (MRI) [4,5]. However, these non-invasive imaging approaches may not be capable of determining the cancerous areas efficiently. To this end, the biopsy technique is usually used to analyze the malignancy in breast cancer tissues more comprehensively. The process of biopsy includes the collection of tissue samples, mounting them on microscopic glass slides, and staining these slides for better visualization of nuclei and cytoplasm [6]. Pathologists then carry out the microscopic analysis of these slides in order to finalize the diagnosis of breast cancer [6]. The complete process of biopsy technique is depicted in Figure 1, and is comprehensively described in [7].

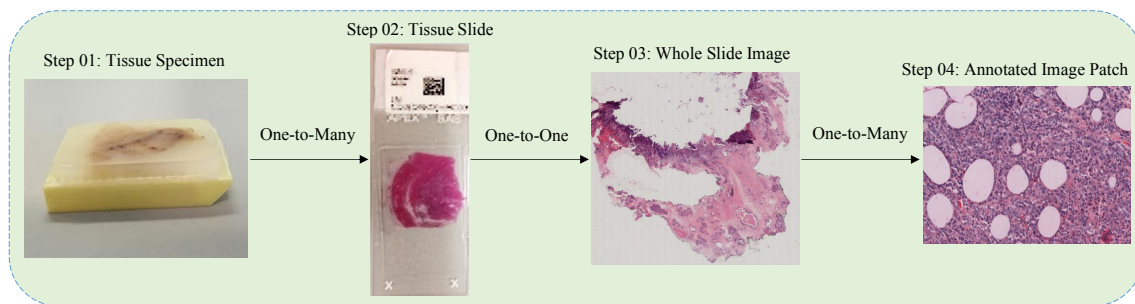


Figure 1. The complete process of biopsy is depicted in Figure. Steps 01 and 02 are taken from [7] whereas steps 03 and 04 are retrieved from our own dataset.

However, the manual analysis of complex-natured histopathological images is fairly a time-consuming and tedious process, and could be prone to errors. Also, the morphological criteria used in the classification of these images are somehow subjective, which leads to the result that an average diagnostic concordance among the pathologists is approximately 75% [8]. Therefore, the computer-assisted diagnosis [4,6,9] plays a significant role to assist pathologists in analyzing the histopathology images. Specifically, it improves the diagnostic accuracy of breast cancer by reducing the inter-pathologist variations in diagnostic decisions [6]. However, the conventional computerized diagnostic approaches, ranging from rule-based systems to machine learning techniques, may not effectively challenge the intra-class variation and inter-class consistency within the histopathology images of breast cancer [10]. Also, these methodologies mainly rely on feature extraction methods like scale-invariant feature transform [11], speed robust features [12] and local binary patterns [13] which all are based on supervised information and can be prone to biased results during the classification of breast cancer histopathology images [10]. Therefore, the need for efficient diagnosis leads to an advanced set of computational models based on multiple layers of nonlinear processing units, called deep learning [14].

Recently, deep learning models [7,14–17] have made remarkable progress in computer vision, specifically in biomedical image processing, due to their abilities to automatically learn complicated and advanced features from images, which inspired various researchers to leverage these models in the classification of breast cancer histopathology images [7]. Especially convolutional neural networks (CNNs) [18] are widely used in image-related tasks due to their abilities to effectively share parameters across various layers within a deep learning model. Numerous CNN-based architectures have been proposed during the past few years; however, AlexNet [19] is considered as of the first deep CNNs to achieve considerable accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) during 2012. Thereafter, VGG architecture [20] introduced the idea of leveraging deeper networks with smaller convolutional filters, and achieved second place at ILSVRC 2014. The intuition of multiple stacked smaller convolutional filters can provide an effective receptive field and is also used in recently proposed pretrained models, including Inception Network [21] and residual neural network (ResNet) [22]. In this work, we employed two different approaches of VGG architecture for an efficient classification of breast cancer histopathology by utilizing our own created dataset. The main contributions of this paper are: First, we created a private dataset of whole slide images

(WSI) from breast cancer patients with the help of experienced pathologists. Then, image patches were extracted from the WSI images, composed of non-carcinoma and carcinoma classes. Next, we selected and trained different combinations of pretrained VGG16 and VGG19 [20] deep learning architectures (discussed in Section 3). Specifically, we evaluated an individual as well as ensemble performances of fully-trained and fine-tuned VGG16 and VGG19 frameworks [20]. Of note, our main objective is the correct classification of the carcinoma class on a priority basis and we found that the ensemble of fine-tuned VGG16 and VGG19 approach [20] provided superior performance in the classification of non-carcinoma and carcinoma histopathology images of breast cancer.

The remaining sections of this paper are provided as follows. Section 2 presents related work. Section 3 demonstrates the materials and methods used to conduct this research. Section 4 shows experimental setup and Section 5 illustrates the results along with discussion. Finally, Section 6 highlights the conclusion and future direction of this study.

2. Related Work

With the evolution of machine learning in biomedical engineering, numerous studies leveraged handcrafted features-based approaches for the classification of histopathology images related to breast cancer. For instance, Kowal et al. [23] focused on the nuclei segmentation and extracted forty-two morphological, topological and texture features from the segmented nuclei of 500 fine-needle biopsy images of breast cancer. Then, these features were utilized to train three different classifiers in order to classify these images into benign and malignant classes. Similarly, Filipczuk et al. [24] also showed interest in the segmentation of nuclei and extracted twenty-five shape-based and texture-based features from the segmented nuclei of 737 cytology images of breast cancer. Based on these features, four different machine learning classifiers, namely, KNN (K-nearest neighbor), NB (Naive Bayes), DT (decision tree), and SVM (support vector machine), were trained for the classification of these cytological images into benign and malignant cases. Apart from nuclei segmentation [23,24], other studies focused on the extraction of global features from the whole images. For instance, Zhang et al. [25] combined local binary patterns, statistics from the gray level co-occurrence matrix and the curvelet transform, and designed a cascade random space ensemble scheme (with rejection options) for an efficient classification of the microscopic biopsy images of breast cancer. Although the traditional machine learning approaches have made satisfactory performances in analyzing the histological images of breast cancer, their performances mainly rely on the selection of features on which they are trained. Furthermore, they might not be capable of effectively extracting and organizing the discriminative information from data [26].

In contrast to the traditional machine learning approaches based on hand-crafted features, deep learning models have the ability to yield complicated and high-level features from images automatically [26]. Consequently, numerous recent studies employed deep learning approaches, with and without leveraging the pre-trained models, for the classification of breast cancer histopathology images. Of note, most of these studies employed BreakHis dataset [27] for the classification task. For instance, Spanhol et al. [28] employed CNN for the classification of breast cancer histopathology images and achieved 4 to 6 percentage points higher accuracy on BreakHis dataset [27] when using a variation of AlexNet [19]. Similarly, Bayramoglu et al. [29] utilized CNN in order to classify the histopathology images breast cancer irrespectively of their resolution using BreakHis dataset [27]. Specifically, the authors proposed single-task and multi-task CNN architectures; whereas the former was capable of predicting malignancy only and the latter was able to predict malignancy and magnification intensity of images simultaneously. These studies leveraging BreakHis dataset provided various state-of-the-art performances; however, they are relying on the same dataset. In this study, we followed the recent approaches of Araújo et al. [30] and Yan et al. [31] and presented a dataset for the classification of breast cancer histology images using deep learning models. However, our dataset contains only non-carcinoma and carcinoma classes, unlike [30,31] which have four classes

in their classification problem. The explanation of our dataset and proposed methodologies are comprehensively discussed in the next Section 3.

3. Materials and Methods

In this section, we introduced our dataset, followed by its preprocessing methodology and training, validation, and testing criteria along with the augmentation process. Then, we discussed the layout of the VGG model and finally, we described the architecture of our proposed ensemble architecture.

3.1. Data Collection

We collected overall 544 whole slides images (WSI) from 80 patients suffering from breast cancer in the pathology department of Colsanitas Colombia University, Bogotá, Colombia. The tumor tissue fragments were fixed in formalin and embedded in paraffin. Subsequently, 4 mm cuts were made that were stained with hematoxylin and eosin (H & E). For the Immunohistochemistry studies, the paraffin-embedded tissue sections were treated with xylene to render them diaphanous (the paraffin being removed later by passing it through decreasing alcohol concentrations until 100% water was reached). Rehydrated sections were rinsed in phosphate buffered saline (PBS) containing 1% Tween-20. For the detection of proteins, sections were heated in a high pH Envision FLEX target retrieval solution at 65 °C for 20 min and then incubated for 20 min at room temperature in the same solution. Endogenous peroxidase activity (3% H₂O₂) and non-specific binding (33% foetal calf serum) were blocked and the sections were incubated overnight at 4 °C with the following primary antibodies: anti-ER (estrogen receptor), anti-PR (progesterone receptor), anti-HER-2 (human epidermal growth factor receptor-2), anti-myosin, anti-Ki-67 (proliferation-associated biomarker). Next, an Ultra View universal DAB kit was used following the manufacturer's recommendations in conjunction with an automated staining procedure.

The tissue sections were then scanned at high resolution (400×) using a Roche iScan HT scanner (<https://diagnostics.roche.com/global/en/products/instruments/ventana-iscan-ht.html>). These WSI images representing multiple cases from every patient were analyzed using H & E, hormone receptors, including ER, PR, HER2, myosin, and Ki-67. Next, two pathologists examined the digital whole slides of tissue stained with H & E and extracted 845 areas from WSI, among which 408 are non-carcinoma and 437 are carcinoma images. The carcinoma class has images of malignant tumors whereas the non-carcinoma class contains images of normal tissues as well as benign images of non-tumor glandular tissues. These areas were photographed at 200× (50 micrometers of resolution) and exported to png format using Qupath 0.1.2 software [32]. The dimensions of these images were noted as 1278 × 760 pixels. This dataset is considered to be balanced and its statistics are represented in Table 1. The main objective related to this dataset is the automatic classification of breast cancer histopathology images, most importantly the carcinoma images.

Table 1. Characteristics of our proposed dataset.

Images	Quantity	Color Model	Staining
carcinoma	437	RGB	H & E
non-carcinoma	408	RGB	H & E
Total	845	RGB	H & E

3.2. Preprocessing

The dataset used in this paper contains histopathology images of breast cancer stained with H & E, which is widely used to assist pathologists during the microscopic assessment of tissue slides. However, it is difficult to maintain the same staining concentration through all the slides, which results in color differences among the acquired images. These contrast differences may adversely affect the training process of the CNN model and thus the color normalization is usually applied. In this

paper, we followed the recent studies [30,31] and employed the approach proposed by [33] for colour normalization. In this method, images are first converted into optical density (OD) by using a logarithmic transformation. Next, singular value decomposition (SVD) is applied to OD tuples to obtain two-dimensional projections with higher variance. Then, the resulting color space transform is applied to the original images. Finally, the histogram of images is stretched in order to cover the lower 90% of data. However, the classification performance of our proposed model deteriorated upon using the normalized images, which is also comprehensively explained in [34]. Eventually, we omitted the stain normalization process and thus used the original images in this paper. The example of original and normalized carcinoma images are shown in Figure 2.

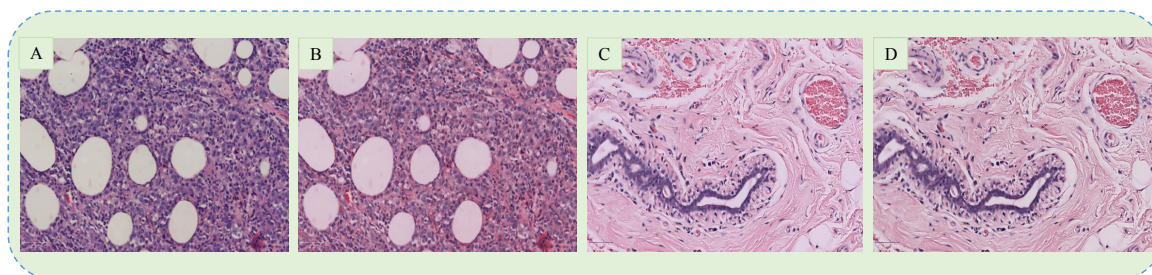


Figure 2. The examples of original (A,C) and normalized (B,D) images of carcinoma and non-carcinoma cases.

3.3. Training Criteria

For the individual and ensemble models, we selected 80% of images for training and the remaining 20% for testing purposes with the same percentage of carcinoma and non-carcinoma images. In this way, 675 images were used for training whereas the remaining 170 images were kept for testing the model. Following [35], we used 5-fold cross-validation on training images which means that 540 images were used for training and 135 images for validation purpose. Again, we have an equal percentage of non-carcinoma and carcinoma images in training and validation. These statistics about training, validation, and testing the models are depicted in Table 2.

Table 2. Criteria for the selection of training, validation, and test images.

	No. of Images	Percentage
Training	540	64%
Validation	135	16%
Test	170	20%
Total	845	100%

3.4. Data Augmentation

Image data augmentation is a technique used to expand the dataset by generating modified images during the training process. By employing the *ImageDataGenerator* provided by Keras deep learning library [36], we generate batches of tensor image data with real-time data augmentation. With this type of data augmentation, we want to ensure that our network, when trained, sees new variations of our data at each and every epoch. Firstly, an input batch of images is presented to the *ImageDataGenerator*, which then transforms each image in the batch by a series of random translations, rotations, etc. The rotation which we specified “rotation range = 40” corresponds to a random rotation angle between $[-40, 40]$ degrees. We also set the “width and height shift range = 0.2” which specifies the upper bound of the fraction of the total width by which the image is to be randomly shifted, either towards the left or right for width or up or down for height. Of note, the rotation operation may rotate some pixels out of the image frame and leave behind empty pixels within the frame which must be filled. We used the “reflect mode” in order to fill these empty pixels. Finally, the randomly transformed

batch is then returned to the calling function. All these parameters along with their values are shown in Table 3.

Table 3. Parameters of data augmentation.

Parameters of Image Augmentation	Values
Zoom range	0.2
Rotation range	40
Width shift range	0.2
Height shift range	0.2
Horizontal flip	True
Fill mode	Reflect

3.5. VGG Architecture

Pretrained models usually help in a better initialization and convergence when the dataset is comparably small as compared to natural image datasets, and this result has been extensively used in other areas of medical imaging too [31]. To this end, we employed deep CNN-based pretrained model proposed by Visual Geometry Group (VGG) of Oxford University [20]. The VGG model is in fact one of the most influential contributions since it reinforced the notion that CNNs have to have a deep network of layers in order for this hierarchical representation of visual data to work. Although numerous follow-up works made improvements in VGG architecture; however, we used its early layouts in this paper, called VGG16 and VGG19 architectures. These names are given because of the fact that VGG16 contains sixteen weight layers whereas VGG19 carries nineteen weight layers in their basic structures [20].

The complete framework of the VGG16 model is portrayed in Figure 3. It is composed of five convolutional blocks and every block has multiple convolution layers (with relu activation), together with a max-pooling layer. It strictly uses 3×3 filters with stride and pad of 1, along with 2×2 maxpooling layers with stride 2. The basic architecture of VGG19 is the same as that of VGG16, except three extra convolutional layers. We tried four different approaches by using these two pretrained architectures. For fully-trained VGG16, we employed all the five blocks and replaced the last three layers by a single dense layer with 256 nodes, as shown in Figure 3. The final output layer is composed of binary cross-entropy loss function which is mathematically shown in Equation (1). Also, for fine-tuned VGG16, we froze the first block (with two convolutional layers and one max-pooling layer) and used the remaining four blocks for the training purpose. Again, we used one dense layer of 256 nodes along with the same loss function of binary cross-entropy. Similarly, for fully-trained VGG19, we trained all the blocks along with one dense layer of 128 nodes. Also, we froze the first block and trained the remaining blocks in the fine-tuned VGG19 model along with a single dense layer of 128 nodes. The final layer in case of VGG19 is also composed of binary cross-entropy loss function, as shown in Equation (1).

$$\text{Binary cross entropy} = -\frac{1}{m} \sum_i^m (y_i * \log(p(y_i)) + (1 - y_i) * \log(1 - p(y_i))) \quad (1)$$

3.6. Proposed Ensemble Approach

The architecture of our proposed ensemble approach is illustrated in Figure 4. It is composed of an ensemble of fine-tuned VGG16 and fine-tuned VGG19 models. First, for both models, training images (80%) are arranged in 5-folds, out of which four are used for training and one is used for model validation or evaluation. Of note, these folds are mutually exclusive and have equal percentages of non-carcinoma and carcinoma images. Also, we used image augmentation during the training process,

as described in Table 3. In every fold, we trained each model for 200 epochs; however, we saved weights of the best model only, based on a minimum value of loss function. In this way, we saved the weight for 5 folds for both models. Then, the test images (20%) are utilized in order to make the final prediction in the form of probabilities. The average probability for every class (non-carcinoma and carcinoma) is derived by taking the mean of ten probability values, obtained from 5-fold VGG16 and 5-fold VGG19 models (10 folds in total). In this way, we considered the average probability of both the models in order to classify images into non-carcinoma or carcinoma classes. The final results of our proposed ensemble deep learning approach are discussed in Section 5.

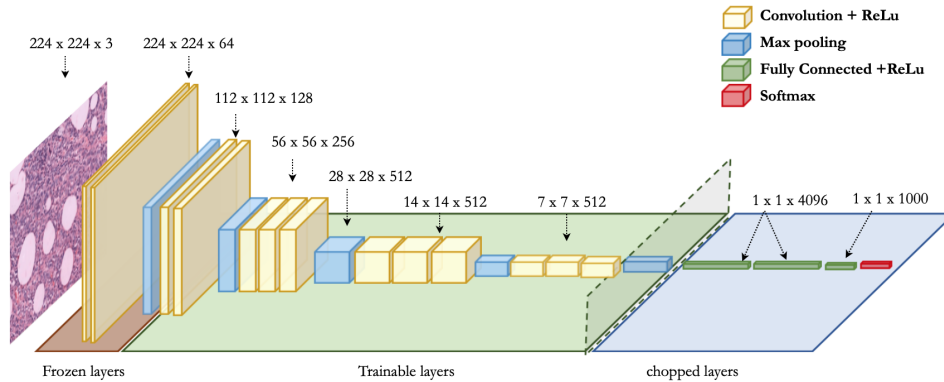


Figure 3. Representation of fine-tuned VGG16 architecture [20]. In fine-tuned VGG16 and VGG19 models, the first block (comprising two convolutional layers and one max-pooling layer) is frozen whereas the rest of layers are trainable. However, in fully-trained VGG16 and VGG19 models, all the five blocks are trainable.

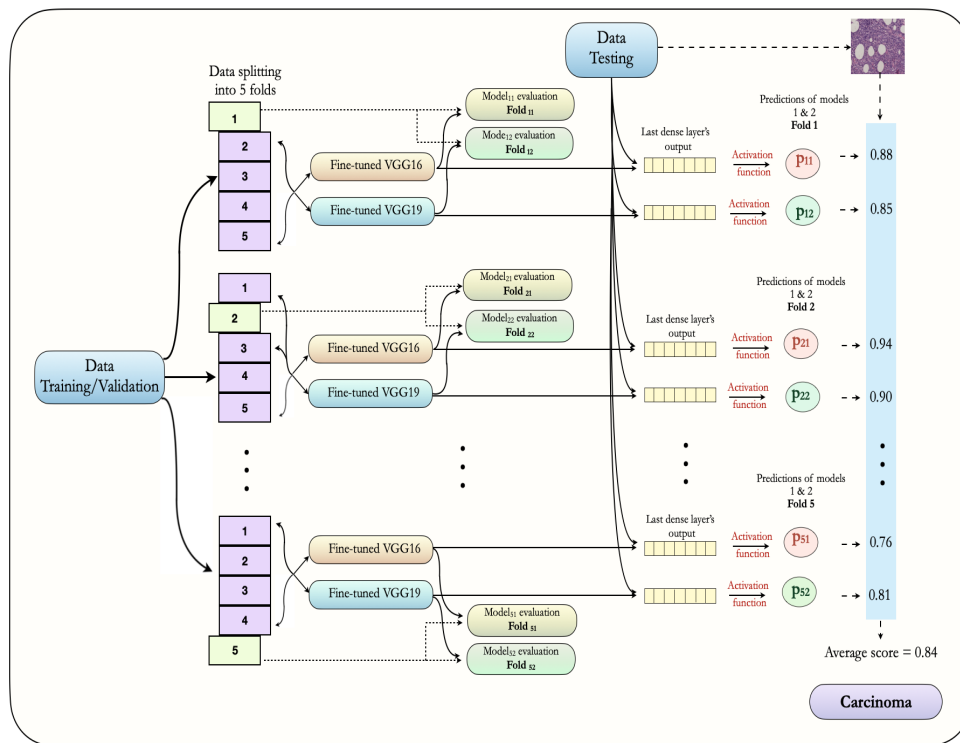


Figure 4. The proposed ensemble architecture using the fine-tuned VGG16 and VGG19 models along with 5-fold cross-validation approach.

4. Experimental Setup

In this section, we explained the experimental environment, followed by the interpretation of evaluation metrics in our proposed model, and finally, we elucidated the tuning of hyperparameters.

4.1. Implementation

We implemented all the experiments related to this article by using *Python 3.7.6* along with *TensorFlow 2.1.0* and *Keras 2.2.4* installed on a standard PC with dual Nvidia GeForce GTX 2070 graphical processing unit (GPU) support. Moreover, this PC has a RAM capacity of 32.0 GB and holds a 3.60 GHz Intel® Core™ i9-9900K processor with 16 logical threads as well as 16 MB of cache memory.

4.2. Evaluation Metrics

The overall performance of our proposed model relies on elements of confusion matrix, also called error matrix or contingency table. This evaluation matrix contains four terms, namely, True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). In our problem, TP refers to those images that were correctly classified as carcinoma and the FP represents the non-carcinoma images mistakenly classified as carcinoma. Whereas, the FN represents the images belonging to carcinoma class that were classified as non-carcinoma, and the TN refers to the non-carcinoma images correctly classified. The classification performance of our proposed model was evaluated on the testing set using four performance measures based on confusion matrix, namely, precision, sensitivity (recall), overall accuracy, and F1-score, using python scikit-learn module. These performance measures can be calculated as follow:

- **Precision:** It quantifies exactness of a model, and represents the ratio of carcinoma images accurately classified out of the union of predicted same-class images.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- **Sensitivity:** Sensitivity, also called “recall” computes completeness of a model. It represents the ratio of images accurately classified as carcinoma out of the total number of carcinoma images.

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

- **Accuracy:** It evaluates correctness of a model, and is the ratio of the number of images accurately classified out of the total number of testing images.

$$OverallAccuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

- **F1-score:** It represents the harmonic average of precision and recall, and is usually used for the optimization of a model towards either precision or recall.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

4.3. Hyperparameter Tuning

Neural networks have a powerful property of learning sophisticated connections between their inputs and outputs automatically [37]. However, some of these connections might be the result of sampling noise, so they can prevail during the training process but could not exist within the real test dataset. This issue may lead to overfitting problems and thus may degrade the prediction performance of a deep learning model [37]. For this very reason, we followed the tuning process of hyperparameters in order to get the generalized performance of our proposed model. The methodology used for the selection of optimal hyperparameters is as follows: First, we selected binary cross-entropy as

a loss function for our binary classification problem. Then, Adam (adaptive moment estimation) algorithm [38,39] was used during the training process in order to perform optimization through 200 epochs. At this stage, we tried three different learning rates (0.001, 0.0001, and 0.00001) and three different batch sizes (16, 32, and 64) while keeping in mind the values used in the recently published study [39,40]. During the model training, our primary aim was to minimize the generalization gap between training loss and validation loss, and found that the batch size of 32 worked well together with the learning rate of 0.0001. Also, we used a dropout of 0.3 in order to prevent the model from overfitting during the training process [41]. Next, we saved the weights of five best models based on their minimal validation loss by using a 5-fold cross validation approach. Finally, we employed these weights for the class prediction on the test dataset. Of note, we used the convolutional filters, pooling filters, strides, and padding with their default values mentioned in the original VGG16 and VGG19 architectures [20]. All the optimal values of hyperparameters used in this study are provided in Table 4.

Table 4. Hyperparameters used in the individual and an ensemble models.

Hyperparameters	VGG16 with Data Augmentation	VGG19 with Data Augmentation
Train approach	5-fold cross-validation	5-fold cross-validation
Optimizer	Adam	Adam
Loss function	Binary cross-entropy	Binary cross-entropy
Learning rate	0.0001	0.0001
Batch size	32	32
Convolution	3 × 3 with stride 1	3 × 3 with stride 1
Padding	Same	Same
Pooling	2 × 2 max-pooling with stride 2	2 × 2 max-pooling with stride 2
Epochs	200	200
Drop out	0.3	0.3
Regularizer	N/A	N/A
Architecture	Fully-trained and Fine-tuned	Fully-trained and Fine-tuned

5. Results and Discussion

In this section, we evaluated the performances of our proposed deep learning models by taking into consideration the average predicted probabilities. First, we highlighted the performance metrics of individual models and then we discussed the competitiveness of our proposed models with recently published studies, especially in terms of carcinoma classification.

5.1. Results of VGG16 Architecture

The performance metrics of fully-trained VGG16 architecture on our dataset are shown in Table 5. It can be noticed that these metrics vary across different folds although using the same test samples. Interestingly, the average recall value (sensitivity) of carcinoma class is noted as 94.55% (± 2.59). Also, the highest accuracy and F1 score are noted during Fold 1, in contrast to their lowest values during Fold 2. The overall accuracy of the fully-trained VGG16 model is 91.41 (± 3.40) along with the average F1 score of 91.38 (± 3.42). The accuracy curves of this model are depicted in Figure 5, whereas its loss curves are displayed in Figure 6.

Similar to fully-trained VGG16 architecture, the performance metrics of fine-tuned VGG16 framework are also presented in Table 5. Again, we used the same test set across all the folds. In this case, the average recall value of carcinoma class can be noticed as 94.09% (± 3.35). Moreover, the highest accuracy and F1 score are found during Fold 5, whereas their respective lowest values

can be seen during Fold 1. Overall, the fine-tuned VGG16 models provided an average accuracy of 91.67% (± 3.69) as well as an average F1 score of 91.63% (± 3.69). The accuracy curves of this model are also illustrated in Figure 5, whereas its loss curves are presented in Figure 6. Lastly, the training and prediction times of fully-trained and fine-tuned VGG16 models are provided in Table 6.

Table 5. Performance metrics of VGG16 architecture on our dataset.

Architecture	Folds	Confusion Matrices			Performance Evaluation (%)				Average (%)	
		Predict → Actual ↓	NC	C	Precision	Recall	F1	Test	Acc.	F1
Fully-Trained VGG16	Fold 1	non-carcinoma carcinoma	75 2	7 86	97.40 92.47	91.46 97.73	94.34 95.03	82 88	94.71	94.70
	Fold 2	non-carcinoma carcinoma	65 7	17 81	90.28 82.65	79.27 92.05	84.42 87.10	82 88	85.88	85.80
	Fold 3	non-carcinoma carcinoma	73 5	9 83	93.59 90.22	89.02 94.32	91.25 92.22	82 88	91.76	91.75
	Fold 4	non-carcinoma carcinoma	70 3	12 85	95.89 87.63	85.37 96.59	90.32 91.89	82 88	91.18	91.13
	Fold 5	non-carcinoma carcinoma	78 7	4 81	91.76 95.29	95.12 92.05	93.41 93.64	82 88	93.53	93.53
	Avg.	non-carcinoma carcinoma	– –	– –	93.78 89.65	88.05 94.55	90.75 91.98	82 88	91.41	91.38
	Fold 1	non-carcinoma carcinoma	67 10	15 78	87.01 83.87	81.71 88.64	84.28 86.19	82 88	85.29	85.27
	Fold 2	non-carcinoma carcinoma	74 6	8 82	92.50 91.11	90.24 93.18	91.36 92.13	82 88	91.76	91.76
	Fold 3	non-carcinoma carcinoma	76 4	6 84	95.00 93.33	92.68 95.45	93.83 94.38	82 88	94.12	94.11
	Fold 4	non-carcinoma carcinoma	73 3	9 85	96.05 90.43	89.02 96.59	92.41 93.41	82 88	92.94	92.92
Fold 5	non-carcinoma carcinoma	75 3	7 85	96.15 92.39	91.46 96.59	93.75 94.44	82 88	94.12	94.11	
Avg.	non-carcinoma carcinoma	– –	– –	93.34 90.23	89.02 94.09	91.13 92.11	82 88	91.67	91.63	

Table 6. The training and prediction times of fully-trained and fine-tuned models.

Model	Single Training Time	5-Fold Training Time	Prediction Time
Fully-trained VGG16	17 min 50 s	89 min	30 s
Fine-tuned VGG16	17 min 25 s	87 min	31 s
Fully-trained VGG19	20 min 40 s	103 min	35 s
Fine-tuned VGG19	19 min 55 s	99 min	36 s

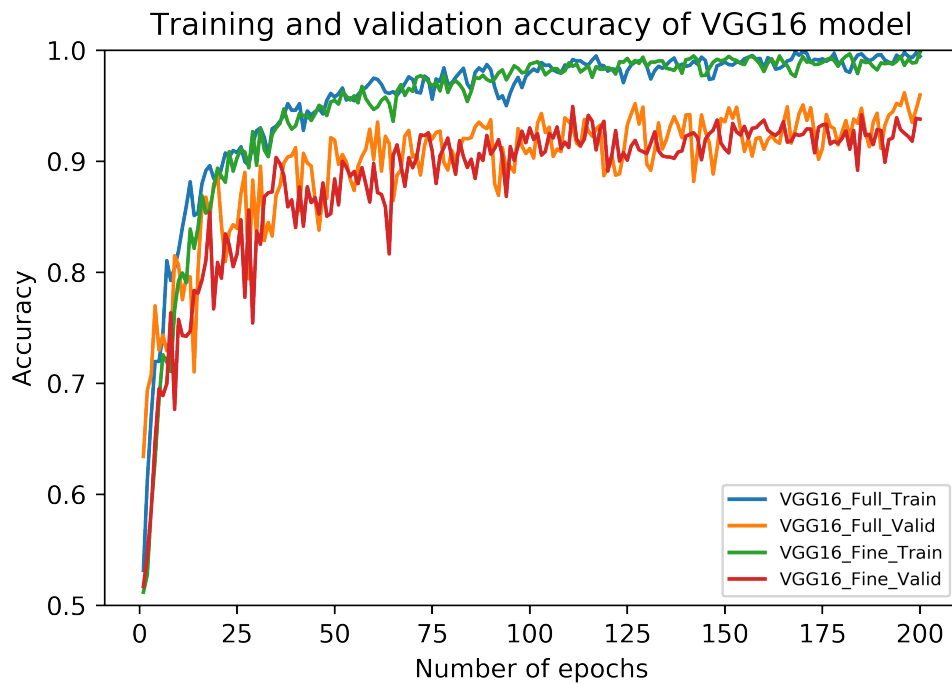


Figure 5. The training and validation accuracy curves of fully-trained and fine-tuned VGG16 models.

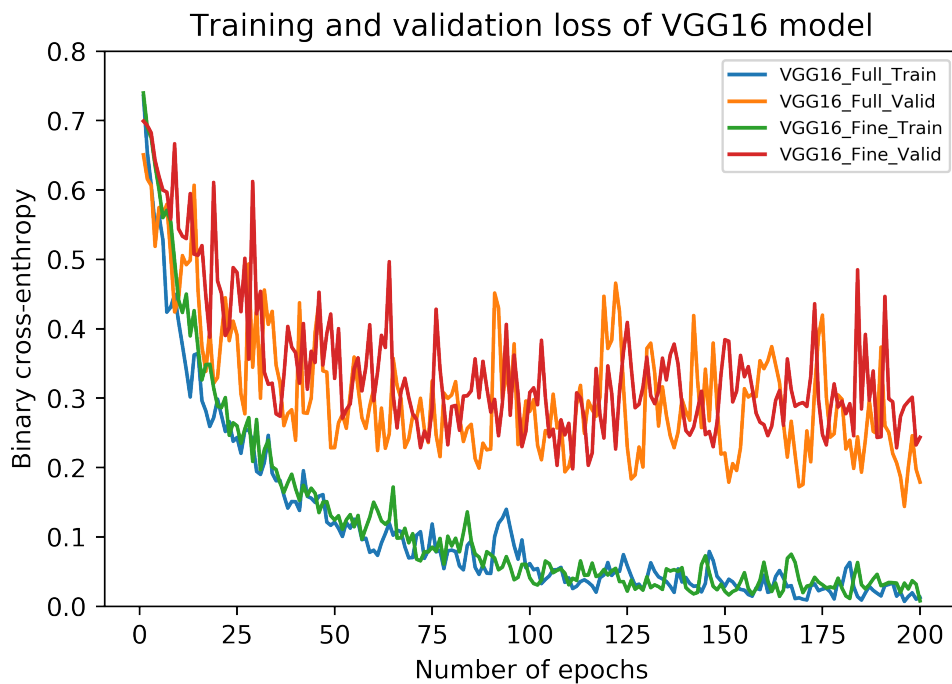


Figure 6. The training and validation loss curves of fully-trained and fine-tuned VGG16 models.

5.2. Results of VGG19 Architecture

The performance metrics of fully-trained VGG19 architecture on our dataset are presented in Table 7. In this case, the average recall value (sensitivity) for carcinoma images is 95.45% (± 3.41) which is 0.9 percentage points higher than that of the fully-trained VGG16 model. Also, the maximum values of the accuracy and an F1 scores occurred during Fold 3, whereas their minimum values found during Fold 4. Finally, the overall accuracy of the fully-trained VGG19 model is 90.35% (± 1.35) in

together with the average F1 score of 90.31% (± 1.35). The accuracy curves of this model are illustrated in Figure 7, whereas its loss curves are portrayed in Figure 8.

Similar to the fully-trained VGG19 model, the performance metrics of fine-tuned VGG19 architecture are portrayed in Table 7. The average recall value for carcinoma cases is 95.68% (± 3.15) which reflects 1.59 percentage points higher than that of the fine-tuned VGG16 model. In this case, the highest values of accuracy and an F1 score are noted for Fold 3 and 4, whereas their low values occurred during Fold 1. The average accuracy and F1 score in this case are 91.67% (± 2.99) and 91.63% (± 3.03), respectively. The accuracy curves of this model are also presented in Figure 7, whereas its loss curves are shown in Figure 8. Finally, like the VGG16 models, the training and prediction times of fully-trained and fine-tuned VGG19 frameworks are also given in Table 6.

Table 7. Performance metrics of VGG19 architecture on our dataset.

Architecture	Folds	Confusion Matrices			Performance Evaluation (%)				Average (%)	
		Predict \rightarrow Actual \downarrow	NC	C	Precision	Recall	F1	Test	Acc.	F1
Fully-Trained VGG19	Fold 1	non-carcinoma carcinoma	66 1	16 87	98.51 84.47	80.49 98.86	88.59 91.10	82 88	90.00	89.89
	Fold 2	non-carcinoma carcinoma	71 4	11 84	94.67 88.42	86.59 95.45	90.45 91.80	82 88	91.18	91.15
	Fold 3	non-carcinoma carcinoma	69 1	13 87	98.57 87.00	84.15 98.86	90.79 92.55	82 88	91.76	91.70
	Fold 4	non-carcinoma carcinoma	69 7	13 81	90.79 86.17	84.15 92.05	87.34 89.01	82 88	88.24	88.21
	Fold 5	non-carcinoma carcinoma	73 7	9 81	91.25 90.00	89.02 92.05	90.12 91.01	82 88	90.59	90.58
	Avg.	non-carcinoma carcinoma	– –	– –	94.76 87.21	84.88 95.45	89.46 91.09	82 88	90.35	90.31
Fine-Tuned VGG19	Fold 1	non-carcinoma carcinoma	64 1	18 87	98.46 82.86	78.05 98.86	87.07 90.16	82 88	88.82	88.67
	Fold 2	non-carcinoma carcinoma	75 5	7 83	93.75 92.22	91.46 94.32	92.59 93.26	82 88	92.94	92.94
	Fold 3	non-carcinoma carcinoma	75 2	7 86	97.40 92.47	91.46 97.73	94.34 95.03	82 88	94.71	94.70
	Fold 4	non-carcinoma carcinoma	76 3	6 85	96.20 93.41	92.68 96.59	94.41 94.97	82 88	94.71	94.70
	Fold 5	non-carcinoma carcinoma	71 8	11 80	89.87 87.91	86.59 90.91	88.20 89.39	82 88	88.82	88.81
	Avg.	non-carcinoma carcinoma	– –	– –	95.14 89.77	88.05 95.68	91.32 92.56	82 88	92.00	91.96

5.3. Results of Ensemble VGG16 and VGG19

The performance metrics of the ensemble VGG16 and VGG19 framework are shown in Table 8. In this approach, we ensemble the fully-trained VGG16 and VGG19 architectures and the fine-tuned VGG16 and VGG19 frameworks by taking the average of output probabilities among all the folds in the aforementioned architectures. Interestingly, the recall value for the carcinoma class is noted as the same (97.73%) in both fully-trained and fine-tuned ensemble approaches. However, the fine-tuned approach offered high accuracy and F1 score (overall) compared to the fully-trained approach, as shown in Table 8.

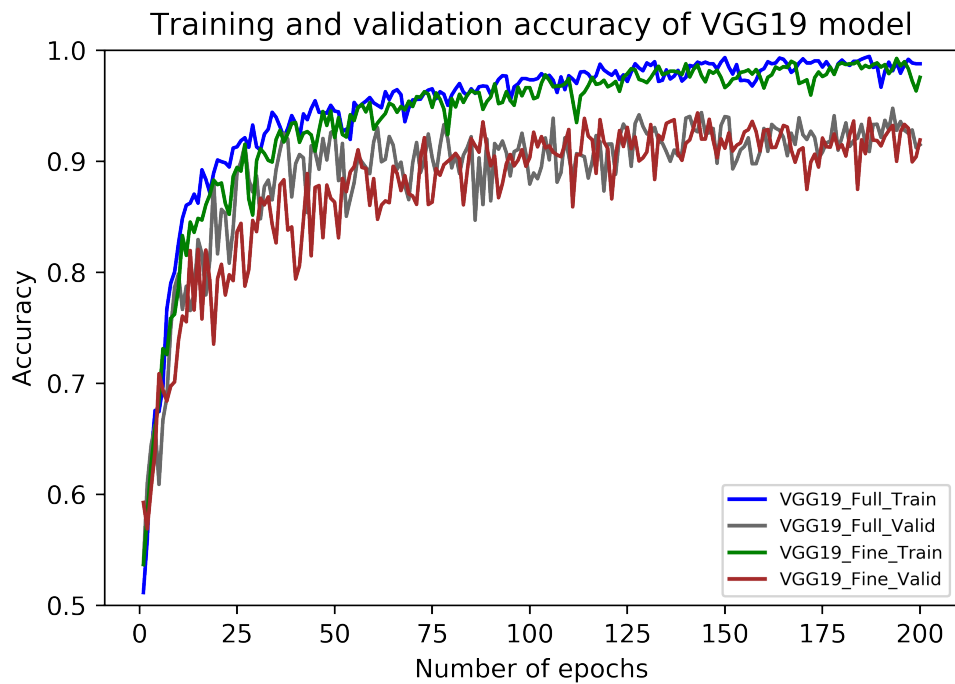


Figure 7. The training and validation accuracy curves of fully-trained and fine-tuned VGG19 models.

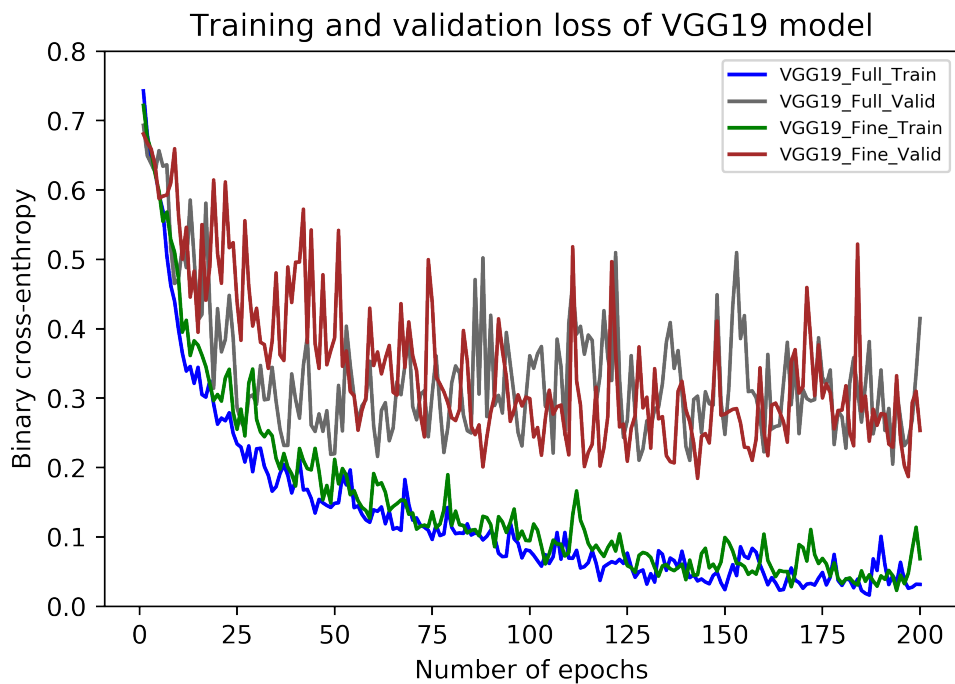


Figure 8. The training and validation loss curves of fully-trained and fine-tuned VGG19 models.

5.4. Discussion

The effectiveness of our proposed ensembling approach can be compared with various state-of-the-art studies used for the classification of breast cancer histopathology images. Most of these novel deep learning approaches are based on BreakHis dataset [27]. For instance, Spanhol et al. [28] employed a variant of AlexNet [19] for the classification of benign and malignant images of BreakHis dataset [27]. The authors used sum, product and maximum fusions rules along with different patch sizes and reported an image level accuracy of 84.0% (± 3.2) for $200\times$ image magnification. In the

following year, Bayramoglu et al. [29] proposed a magnification independent approach for BreakHis dataset. Specifically, the authors presented “single task CNN” and “multi-task CNN” frameworks, where the former predicts malignancy and the latter predicts malignancy as well as the magnification level in the benign and malignant images. For $200\times$ magnification, the authors reported an accuracy of 84.63% (± 2.72) and 82.56% (± 3.49) for single task CNN and multi-task CNN, respectively. Both of these studies [28,29] reported better classification performance than the traditional hand-crafted machine learning approaches. In comparison with Spanhol et al. [28] and Bayramoglu et al. [29], our approach shows better classification performance despite using a comparatively small dataset. Recently, Han et al. [42] proposed a structured deep learning model called class structure-based deep CNN (CSDCNN) for the classification of benign and malignant histopathology images of breast cancer, and reported an accuracy of 96.7% (± 2.0) on BreakHis dataset for $200\times$ magnification factor. Similarly, Nahid et al. [43] first used clustering algorithm in order to retrieve the statistical and geometrical clusters hidden in the histopathology images. The authors then evaluated the effect of deep CNN in together with short-term memory (LSTM) network for the efficient classification of benign and malignant images, and thus achieved an accuracy of 91.0% on BreakHis dataset for $200\times$ magnification. Lastly, Daniz et al. [44] employed fine-tuned AlexNet [19] and VGG16 [20] models for the classification of breast cancer histopathology images. The authors followed 5-fold cross-validation approach and reported a maximum accuracy of 91.37% (± 1.72) when using fine-tuned AlexNet [19] on BreakHis dataset for $200\times$ magnification. These state-of-the-art studies [28,29,42–44] along with other novel frameworks are comprehensively reviewed in [45]. Although having a small dataset, our results are still competitive with the novel deep learning frameworks [28,29,42–45]. In summary, the results demonstrated that our proposed ensemble deep learning model can retrieve various multi-level and multi-scale features from histopathology images of breast cancer. It also became clear from the comparison process that the results of our proposed architecture is competitive with numerous state-of-the-art studies using comparably bigger datasets.

Table 8. Performance metrics of ensemble VGG16 and VGG19 architectures.

Ensemble Method	Confusion Matrices			Performance Evaluation (%)				Average (%)	
	Predict \rightarrow Actual \downarrow	NC	C	Precision	Recall	F1	Test	Accuracy	F1
Full-Trained	non-carcinoma	73	9	97.33	89.02	92.99	82	93.53	93.51
VGG16+VGG19	carcinoma	2	86	90.53	97.73	93.99	88		
Fine-Tuned	non-carcinoma	76	6	97.44	92.68	95.00	82	95.29	95.29
VGG16+VGG19	carcinoma	2	86	93.48	97.73	95.56	88		

6. Conclusions

In this paper, we presented an ensemble deep learning approach for the classification of breast cancer histopathology images using our collected dataset. The main objective of this work was to effectively classify carcinoma images. We found that it could be better to use the average predicted probabilities of two individual models. To this end, we employed an ensemble of fine-tuned VGG16 and VGG19 models and achieved a relatively more robust model. The proposed ensemble approach provides competitive performance on the classification of complex-natured histopathology images of breast cancer. However, our collected dataset is comparatively small in contrast to the datasets used in numerous state-of-the-art studies. Also, our dataset contains merely two-class images. The future indications of this study include the extension of our dataset and the inclusion of images for multi-class classification problems. Also, other pretrained models need to be included in the future work. Finally, it will be interesting to apply similar ensemble criteria to histopathology images of different cancers, such as lung cancer.

Author Contributions: Conceptualization, Z.H., S.Z., B.G.-Z., J.J.A., and A.M.V.; methodology, Z.H., S.Z., and B.G.-Z.; software, Z.H. and S.Z.; validation, Z.H., S.Z. and B.G.-Z.; formal analysis, B.G.-Z.; investigation, Z.H., S.Z., B.G.-Z., J.J.A., and A.M.V.; resources, Z.H., S.Z., and B.G.-Z.; data curation, Z.H. and S.Z.; writing—original draft preparation, Z.H.; writing—review and editing, B.G.-Z., S.Z., J.J.A., and A.M.V.; visualization, Z.H. and S.Z.; supervision, B.G.-Z., J.J.A., and A.M.V.; project administration, B.G.-Z., J.J.A., and A.M.V.; funding acquisition, B.G.-Z. and J.J.A. All authors have read and agreed to the published version of the manuscript.

Funding: Acknowledgment to the Basque Country project MIFLUDAN that partially provided funds for this work in collaboration with eVida Research Group IT 905-16, University of Deusto, Bilbao, Spain.

Acknowledgments: Acknowledgment to the team and partners of MIFLUDAN project and to the Colson Hospital for their support to this research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GBD	Global Burden of Disease
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
KNN	K-Nearest Neighbor
NB	Naive Bayes
DT	Decision Tree
SVM	Support Vector Machine
H&E	Hematoxylin and Eosin
SVD	Singular Value Decomposition
OD	Optical Density
WSI	Whole Slides Image
ER	Estrogen Receptor
PR	Progesterone Receptor
HER2	Human Epidermal growth factor Receptor-2

References

1. Fitzmaurice, C. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: A systematic analysis for the global burden of disease study. *JAMA Oncol.* **2019**, *5*, 1749–1768.
2. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [[CrossRef](#)] [[PubMed](#)]
3. Weigelt, B.; Geyer, F.C.; Reis-Filho, J.S. Histological types of breast cancer: How special are they? *Mol. Oncol.* **2010**, *4*, 192–208. [[CrossRef](#)] [[PubMed](#)]
4. Dromain, C.; Boyer, B.; Ferré, R.; Canale, S.; Delalogue, S.; Balleyguier, C. Computed-aided diagnosis (CAD) in the detection of breast cancer. *Eur. J. Radiol.* **2013**, *82*, 417–423. [[CrossRef](#)]
5. Wang, L. Early Diagnosis of Breast Cancer. *Sensors* **2017**, *17*, 1572. [[CrossRef](#)]
6. Veta, M.; Pluim, J.P.W.; Diest, P.J.V.; Viergever, M.A. Breast Cancer Histopathology Image Analysis: A Review. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 1400–1411. [[CrossRef](#)]
7. Dimitriou, N.; Arandjelović, O.; Caie, P.D. Deep Learning for Whole Slide Image Analysis: An Overview. *Front. Med.* **2019**, *6*. [[CrossRef](#)]
8. Elmore, J.G.; Longton, G.M.; Carney, P.A.; Geller, B.M.; Onega, T.; Tosteson, A.N.A.; Nelson, H.D.; Pepe, M.S.; Allison, K.H.; Schnitt, S.J.; et al. Diagnostic Concordance among Pathologists Interpreting Breast Biopsy Specimens. *JAMA* **2015**, *313*, 1122. [[CrossRef](#)]
9. Aswathy, M.; Jagannath, M. Detection of breast cancer on digital histopathology images: Present status and future possibilities. *Inform. Med. Unlocked* **2017**, *8*, 74–79. [[CrossRef](#)]
10. Robertson, S.; Azizpour, H.; Smith, K.; Hartman, J. Digital image analysis in breast pathology—From image processing techniques to artificial intelligence. *Transl. Res.* **2018**, *194*, 19–35. [[CrossRef](#)]

11. Lowe, D. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–25 September 1999.
12. Bay, H.; Tuytelaars, T.; Gool, L.V. SURF: Speeded Up Robust Features. In Proceedings of the Computer Vision—ECCV 2006 Lecture Notes in Computer Science, Graz, Austria, 7–13 May 2006; pp. 404–417.
13. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [[CrossRef](#)]
14. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
15. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Laak, J.A.V.D.; Ginneken, B.V.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [[CrossRef](#)]
16. Zahia, S.; Zapirain, M.B.G.; Sevillano, X.; González, A.; Kim, P.J.; Elmaghraby, A. Pressure injury image analysis with machine learning techniques: A systematic review on previous and possible future methods. *Artif. Intell. Med.* **2020**, *102*, 101742. [[CrossRef](#)]
17. Bour, A.; Castillo-Olea, C.; Garcia-Zapirain, B.; Zahia, S. Automatic colon polyp classification using Convolutional Neural Network: A Case Study at Basque Country. In Proceedings of the 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, 10–12 December 2019.
18. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
19. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012.
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
23. Kowal, M.; Filipczuk, P.; Obuchowicz, A.; Korbicz, J.; Monczak, R. Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Comput. Biol. Med.* **2013**, *43*, 1563–1572. [[CrossRef](#)]
24. Filipczuk, P.; Fevens, T.; Krzyzak, A.; Monczak, R. Computer-Aided Breast Cancer Diagnosis Based on the Analysis of Cytological Images of Fine Needle Biopsies. *IEEE Trans. Med. Imaging* **2013**, *32*, 2169–2178. [[CrossRef](#)]
25. Zhang, Y.; Zhang, B.; Coenen, F.; Lu, W. Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles. *Mach. Vis. Appl.* **2012**, *24*, 1405–1420. [[CrossRef](#)]
26. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828 [[CrossRef](#)]
27. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1455–1462. [[CrossRef](#)]
28. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. Breast cancer histopathological image classification using Convolutional Neural Networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016.
29. Bayramoglu, N.; Kannala, J.; Heikkila, J. Deep learning for magnification independent breast cancer histopathology image classification. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancún, Mexico, 4–8 December 2016.
30. Araújo, T.; Aresta, G.; Castro, E.; Rouco, J.; Aguiar, P.; Eloy, C.; Polónia, A.; Campilho, A. Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS ONE* **2017**, *12*, e0177544. [[CrossRef](#)] [[PubMed](#)]
31. Yan, R.; Ren, F.; Wang, Z.; Wang, L.; Zhang, T.; Liu, Y.; Rao, X.; Zheng, C.; Zhang, F. Breast cancer histopathological image classification using a hybrid deep neural network. *Methods* **2020**, *173*, 52–60. [[CrossRef](#)] [[PubMed](#)]

32. Bankhead, P.; Loughrey, M.B.; Fernández, J.A.; Dombrowski, Y.; Mcart, D.G.; Dunne, P.D.; Mcquaid, S.; Gray, R.T.; Murray, L.J.; Coleman, H.G.; et al. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **2017**, *7*, 1–7. [[CrossRef](#)] [[PubMed](#)]
33. Macenko, M.; Niethammer, M.; Marron, J.S.; Borland, D.; Woosley, J.T.; Guan, X.; Schmitt, C.; Thomas, N.E. A method for normalizing histology slides for quantitative analysis. In Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June–1 July 2009.
34. Bianconi, F.; Kather, J.N.; Reyes-Aldasoro, C.C. Evaluation of Colour Pre-processing on Patch-Based Classification of H&E-Stained Images. In *European Congress on Digital Pathology*; Springer: Cham, Switzerland, 2019; pp. 56–64.
35. Yao, H.; Zhang, X.; Zhou, X.; Liu, S. Parallel Structure Deep Neural Network Using CNN and RNN with an Attention Mechanism for Breast Cancer Histology Image Classification. *Cancers* **2019**, *11*, 1901. [[CrossRef](#)] [[PubMed](#)]
36. Chollet, F. Keras: Deep Learning Library. 2015. Available online: <https://keras.io/api/preprocessing/image/> (accessed on 24 January 2020).
37. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2017.
38. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
39. Xie, J.; Liu, R.; Luttrell, J.; Zhang, C. Deep Learning Based Analysis of Histopathological Images of Breast Cancer. *Front. Genet.* **2019**, *10*, 80. [[CrossRef](#)] [[PubMed](#)]
40. Bardou, D.; Zhang, K.; Ahmad, S.M. Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks. *IEEE Access* **2018**, *6*, 24680–24693. [[CrossRef](#)]
41. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
42. Han, Z.; Wei, B.; Zheng, Y.; Yin, Y.; Li, K.; Li, S. Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Sci. Rep.* **2017**, *7*, 1–10. [[CrossRef](#)] [[PubMed](#)]
43. Nahid, A.-A.; Mehrabi, M.A.; Kong, Y. Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering. *BioMed Res. Int.* **2018**, *2018*, 2362108. [[CrossRef](#)] [[PubMed](#)]
44. Deniz, E.; Şengür, A.; Kadiroğlu, Z.; Guo, Y.; Bajaj, V.; Budak, Ü. Transfer learning based histopathologic image classification for breast cancer detection. *Health Inf. Sci. Syst.* **2018**, *6*, 18. [[CrossRef](#)] [[PubMed](#)]
45. Benhammou, Y.; Achchab, B.; Herrera, F.; Tabik, S. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing* **2020**, *375*, 9–24. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).