Detection of Invasive Ductal Carcinoma from Breast Histopathology Image using Deep Ensemble Neural Networks

Sourodip Ghosh[1], Richik Ghosh[1], Shreya Sahayr[1] and Suprava Patnaik[2]
Dept. of Electronics Engg, KIIT University, Bhubaneshwar, India

[1][sourodip.ghosh02, richikghosh98, sh10findit]@gmail.com)
[2] suprava.patnaikfet@kiit.ac.in

Abstract. This paper presents a Deep Learning Ensemble framework for automatic detection of invasive ductal carcinoma from breast cancer histopathology images for diagnostic assistance. Pathological means of detecting IDC are predominantly arduous and tend to consume significant amount of time since the consulting pathologist must examine large specimens of benign samples before being able to ascertain the malignant ones. In the recent scenario, AI algorithms have meticulously outperformed some of the experts in analyzing medical images. This work presents an algorithm for breast cancer tissue detection, which has been tested on one of the largest datasets containing histopathological images of breast tissues samples. The proposed network has achieved a state of the art performance through a balanced accuracy of 96.16% alongwith Matthews correlation coefficient of 0.923 and kappa score of 0.9231. Furthermore, using extended convolutional layers and the various techniques of feature mapping, the IDC affected regions were methodically isolated from the test slides including the probable regions which might also be vulnerable. The training directory was set to 80%, followed by the test and validation directories, each of size 10%. The ensemble model is proven to be diverse and has significantly less chances of over-fitting without employing methods like cross-validation.

Keywords: Invasive Ductal Carcinoma; Deep NN; Ensemble learning.

1 Introduction

After successful deployment of Deep Neural Network (DNN) for many real word image analysis applications , it is also gaining popularity for medical image processing. Cell structure and texture analysis plays a crucial role in breast cancer pathology. Reliable automotive investigation of histopathological tissues can be a great support for treatment planning.

Invasive Ductal Carcinoma (IDC) is the most prevailing type of breast cancer, encompassing about 80 per cent of its totality. Unlike Ductal Carcinoma In Situ (DCIS), which is confined within ducts, IDC has the tendency to rapidly infiltrate the duct as well as other surrounding areas [1]. The intensity of the disease can be assessed by a procedure called tumour grading which, however, is limited solely to areas particularly affected by invasive cancer [2]. Using a variety of methods, researches have laid the foundation to and inspected the detection of breast cancer in histopathological images. Images were isolated into histologic primitives (such as mitosis, epithelium). These quantifications were then set as the substratum for feature extraction, which are the entities described in various scales for differentiating between benign and malignant areas and further for other types of classification [3]. As discussed in [4], the aforementioned approach also appears to hold certain drawbacks—firstly, such specific approaches demand intensive research. Firstly, in case of nuclei segmentation, it is vital to have a thorough perception of all possible incongruity in texture, morphology and anatomical structures. Secondly, the application of optimal parameters often becomes cumbersome for external observers, because the ability to analyze and adjust such parameters often intrinsic to the developers of the algorithm. Traditional machine learning methods have been extensively used in image analysis procedures, particularly in breast histopathology image analysis, such as mitosis detection, nuclei segmentation and cancer detection.

A large segment of the past methodologies included consolidating an enormous number of complex handcrafted features to create a visual substance as a representation of the Bbreast Cancer (BC) histopathological images. The results of classification of these images depended considerably on types of pre-processing done on the same, constituting of detection and segmentation.In more recent researches, feature learning inculcation and representation that does not assimilate domain proficiency have been integrated in complex learning procedures [5]. Deep Learning (DL) [6] is part of a broader paradigm of feature learning. It is focused on improving upon learned representations with several instances of abstraction and yielding better representations with each iteration. DL approaches have recently gained reputation in the field of cellular feature analysis, pattern recognition and in situations where there is lack of domain understanding for feature introspection. It is also uniquely suited for processing of big data repositories and large datasets.

Recent works on breast cancer detection have employed various CNN architectures since weight sharing and convolution operation has proven to be very efficient for automatic feature extraction. In [7] Han and team proposed the class structure-based deep convolutional neural network (CSDCNN) for multi-classification of different types on the BreakHis dataset. The algorithm leveraged high end hierarchical feature representation. The main advantage of CNN over its predecessors is weight sharing and significantly stronger feature extraction ability.

2 Literature Survey

A novel diagnosis architecture of breast cancer using whole slide images was introduced in [8]. It utilizes both histopathological image classification and content based image classification image retrieval (CBHIR). The malignant patches in the processed images were detected through a probability map. He et al. [9] used a stacked sparse auto-encoder for rank-level fusion of local and integrated features for enhanced breast cancer diagnosis using histopathological images. Even though the classical machine learning architectures have managed to perform well and have achieved significantly good scores, the diagnostic performance still remains unreliable and to some extent, devoid of potential scope.To further enhance the performance capabilities, Deep Learning has been widely incorporated in many spheres of image and pattern analysis researches. Its techniques have the power to carry out automated image extraction and information retrieval from the data and thereby learn improved and abridged representations of the same. Hence, many researches have been conducted that constituted the use of Deep Learning for breast cancer classification and detection [10] - [11] [12] [13] [14] [15] [16] [17]. In [15] Celik et al. proposed transfer learning through DenseNet-161 and ResNet-50 models to distinguish between IDC negative and IDC positive histopathology images to achieve best balanced accuracy of 91.57% through DenseNet- 161 and best F-score performance of 94.11% through ResNet- 50 while also laying out insightful comparison with results of other previous works.

Han et al. [7] laid out a proposition of the Class Structured Deep Convolutional Neural Network (CSDCNN), a new comprehensive recognition framework for multi-class classification of subordinate classes of breast cancer ductal carcinoma, fibrodenoma, lobular carcinoma and so on, reaching an average accuracy of 93.2%. Kassani et al. [18] used an ensemble network consisting of VGG19, MobileNetV2 and DenseNet201 on four different open-source breast histopathology image dataset to showcase the exceeding performance of the ensemble network with respect to that of the three aforementioned CNN architectures. The resultant accuracy scores were then compared against other state-of-art classifiers and different machine learning models. Performance, as previously achieved by classical DCNNs such as DenseNet, ResNet, InceptionV3 and VGG-16 individually, is seen to be boosted further in [19] where Wang et al. devise multi-network feature extraction model involving multi-level Inception V3 and multi-level VGG-16, which are thereby further combined with DenseNet-121 and ResNet-50. They pointed out that existing models fail to acknowledge contemporary relation among the features of two different DCNNs. In order to tackle this they proposed the DOLL (dual-network orthogonal low-rank learning) method which utilizes various feature relations to remove redundant features. They also replaced the softmax layer in DCNN with ensemble-SVM classifiers.

This experimental setup was then shown to improve performance and thereby proved its superiority over other contemporary methods. In [20] authors have proposed the MuDeRN framework that used residual learning for classification of sub-classes of benign and malignant breast cancer. Deep Learning has evidently proven to be much more efficient and reliable than traditional machine learning models, however many a times the former is noticed to be prone to overfitting due to clueless addition of layers for hierarchical feature abstraction. Which subsequently leads in huge number of parameters and restricted training samples. Visual analysis and automatic detection of invasive ductal carcinoma (IDC) tissue segments of breast cancer using deep learning techniques was carried out by Cruz-Roa et al. in their paper [21], where CNN yielded the highest balanced accuracy (BAC) of 84.23% and F-measure of 71.80%. In [22], Arau´jo et al. designed an architecture for information retrieval at different scales, including nuclei and whole tissue structure. The features thus extracted were used to train a SVM network, generating optimum results over that of contemporary methods.

3 Proposed Ensemble DNN model

This research paper presents a Deeply Convoluted Ensemble Network comprising of two pre-trained architectures namely VGG16 and DenseNet, a lightweight CNN architecture called ShuffleNetV2 and a self designed Custom CNN architecture. The proposed model is a framework of Deep Convoluted Ensemble Network. Model training process involves two major steps. As shown in Fig. 1, the initial classification employs the pre-trained architectures available at public disposal, namely VGG16, DenseNet and ShuffleNetV2. The motivation for selecting these particular architectures is because of their exemplary performance in terms of standard error metrics with respect to other architectures when tested on this particular dataset. The second step involves, considering the performance of individual architectures and preparing the final model,by using Deep Ensemble Neural Net- work architecture.

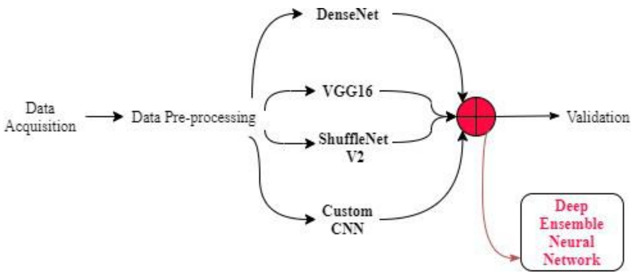

Fig. 1: Schematic of proposed Ensemble Network

3.1 Histopathologi Database

The Breast Histopathology Images dataset has been used for evaluation of the proposed model. The original publication can be found in [4]. This data contains 162 whole mount slide images of breast cancer, scanned in 40x magnification. This has been processed to extract patches of size 50 x 50, resulting in a total range of 277,524. The number of patches we used in our model is 160,000. A ground truth control setup was designated as the 0 labels, indicating non-IDC class. The data cohort consisted of two sub-folders of each patient labeled as '0' and '1'– '0' for non-IDC and '1' for IDC. The experiments were conducted and the data was analysed using TPU as a hardware assist. It is a custom-developed ASIC used to accelerate workload processing power. It facilitates advanced performance of linear computation and is thus used extensively in deep learning model training. Compared to GPU, it increases train computation time by 10x. The Breast Histopathology images data is significantly large, thus it requires more potent processing power for fast and efficient computation. As the number of patches provided is 160,000, batches of size 10 each were arranged which subsequently proved to be helpful in running the model with the assistance of TPU. These run faster than GPUs if the appropriate amount of batched data is passed to it. This helped the authors to reinforce the scope of experimentation, as well as favoured the incorporation of high end feature models.

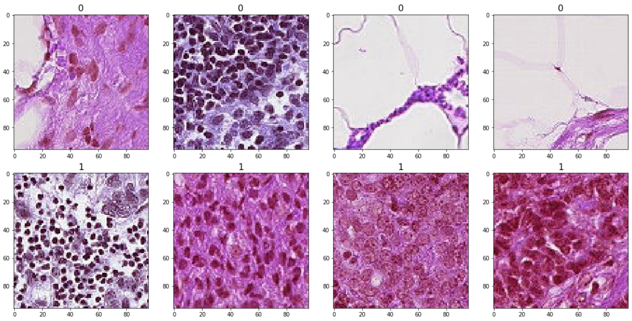

Fig. 2: Representation of labels in original dataset (0: Non- IDC, 1: IDC)

3.2 Data pre-processing

The data initially consisted of 277,524 patches, with an individual class count of 198,738 Non-IDC and 78,786 IDC patches. The dataset is very refined with high feature- containing patches that were extracted to make best-suited fea- tures for individual model training. No features were changed or color channels were altered to prevent damaging of feature maps. The dataset was reduced to 160,000 images and was dissected into training, test, and validation directories, each of size 80%, 10%, and 10% of the original respectively. The layout was carefully organized to implement hyper-parameter tuning. The final dataset highlights after model pre-processing is shown in Table I. The images after data pre-processing with respect to each directory are shown in Fig. 3. The images depict the void of any interference with the original data. Some patches had disrupted features which were not suitable for model building; they were subsequently removed manually from the dataset.

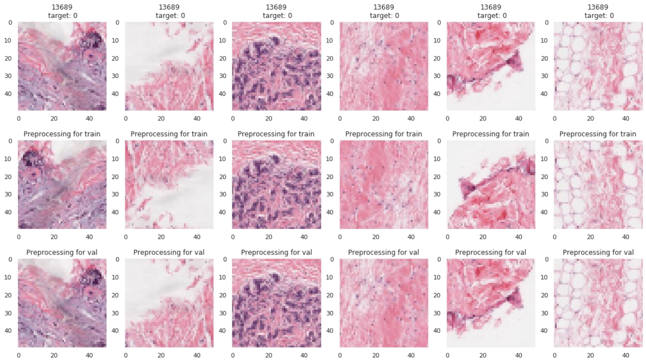

Fig. 3: Representation of slices in the respective directories after Data Pre-processing

TABLE I: Number of Images after Data Pre-processing

	Training Set	Testing Set	Validation Set
No. of images	1,44,000	16,000	16,000

3.3 Detail Architecture

The proposed model is an ensemble network, built using model architectures, namely

1) Pre-trained Architectures: DenseNet, VGG16, ShuffleNet V2

2)Custom CNN Architecture

Pre-trained architectures: The individual pre-trained DCNN models were validated through 10 epochs, keeping the batch size fixed at 10. The batch size was kept small so that the hardware accelerator, TPU, performs to its optimal ability. Also, TPU demands lesser batch size to create tensor buckets for processing information and accessing features from images manually. The size of the images were kept at 94 x 94. No layers were frozen during the computation process.

DenseNet: The DenseNet [23] model contains 3 blocks convoluted together. Each independent block contains 12 layers, supported with BN-ReLU conv layers. The architecture uses feature-maps to transfer information, layer by layer. The initial weights were updated with Adam optimizer. This optimizer is suitable for situations where parameters are high in number. It is invariant to diagonal rescale of gradients.

Also, the hyper-parameters typically require very less tuning. Adam [24] was used to optimise the model parameters. This specifically converges the loss functions and updates weights. It combines the optimised performances of AdaGrad and RMSProp algorithms to regulate sparse gradients in channel noise. The number of epochs was set keeping in mind the fluctuations in validation and training accuracy and error, therefore avoiding chances of model overfitting.

VGG16: VGG16 [24] is a huge network with approximately 138 million parameters. The 16 in VGG16 betokens that it has 16 layers that have weights. This follows a convolution neural network architecture with convolution layers of 3x3 filter having stride of 1. The padding method used here is same, along with which maxpool layers of 2x2 filter with a stride of 2 are being employed. The ending of this architecture consists of two fully connected layers that terminate with a block comprising softmax activation function for multiclass classification. Adam optimizer was used to evaluate loss function by utilizing updated weights.

ShuffleNet V2: The ShuffleNet V2 architecture [25] was custom designed with added conv layers and FC layers in addition to dropouts. ShuffleNet is a very efficient CNN architecture for mobile devices. Here, depthwise separable convolutions are used, which deal with the spatial patterns alongwith the depth dimensions. This splits the kernel in two segments for depth-wise and grouped convolutions on 1x1 conv layers which are point-wise grouped convolutions, to retrieve the channel dimensions. Once these operations are over the channel is shuffled to make efficient computations. This architecture is guided by indirect metrics of computation complexity which are referred to as FLOPs. The direct metric, speed, is mainly dependable on factors like memory access expense and platform components. The model is optimized by the Adam optimizer which converges the mean square error with the increasing number of epochs.

***Custom CNN Architecture*:** The input dimensions of the patches that are passed into the custom CNN model are 94 x 94 x 3. Custom CNN model consists of three groups of three convolutional layers, each in- cluding ReLU activation function. Each of those groups further has one 3X3 max-pool layer and one dropout layer to avoid overfitting. This is succeeded by an FC layer of 256 output classes from which input is passed via a dropout layer to another FC layer consisting of two classes as output with a softmax activation func- tion. The learning rate was intentionally kept less to make the overall model more sensitive. This model is trained by 10 epochs with batch size set to 10 in accordance to the accuracy and loss graph which was found to flatten after the mentioned number of epochs. NADAM rule is used for weight updating having a learning rate set to 0.0001 and using binary cross-entropy as a loss function.

3.4 Ensemble Learning

At this stage, the organization of the considerable number of results accomplished from the individual architectures was done and the evaluation highlights of the models

on 16,000 test pictures were developed. Non-linear weights were allocated based on the performance of the selected models. Negative Correlation (NC) Learning addresses mis-class diversity in neural network ensembles. As the ensemble size manifolds, the upper and lower bounds converge, thereby minimising the error rate.
The FC and Dense Layers of the Ensemble model were kept intact, and the remaining layers were at first frozen. The deep ensemble strategy guarantees the advancement of the model in terms of diversity, boosting accuracy, and alleviating the issue of over-fitting. Along these lines, this implies the exclusivity of the proposed technique from every other study recently conducted on this dataset. The workflow on the ensemble model is showcased in Fig. 4.

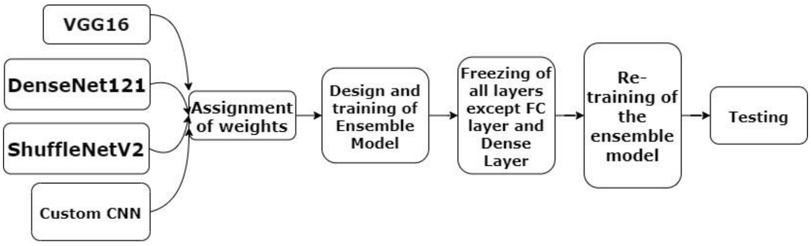
VGG16
DenseNet121
ShuffleNetV2
Custom CNN
Assignment of weights
Design and training of Ensemble Model
Freezing of all layers except FC layer and Dense Layer
Re-training of the ensemble model
Testing

Fig. 4: Proposed Ensemble model

4 Results and Conclusion

4.1 Model Evaluation

The assigned weights from the individual architectures are passed to the Ensemble model during design implementation. All layers, apart from FC layer and dense layer were frozen and the model was re-trained through 10 epochs. This ensured better optimization capability of the final model with respect to individual parameters (and thus the ensemble model was ready). Table II shows performance of individual architectures after first training, in terms of Accuracy, Precision, Recall and F1-Score.

TABLE II: Performance Evaluation of individual architectures

Metrics	DenseNet	Custom CNN	VGG16	ShuffleNetV2
Accuracy	0.93	0.943	0.95	0.95
Precision	0.94	0.95	0.95	0.95
Recall	0.92	0.94	0.94	0.96
F1-Score	0.93	0.94	0.95	0.95

4.2 Result Analysis

The confusion matrix deduced from the ensemble frame- work is depicted in Fig 6. The balanced test accuracy (BAC) settled at 0.962 with an error rate of 0.038. The class specific scores in terms of Sensitivity, Precision and F1-score are listed in Table III. The overall sensitivity and precision was valued to be 0.973 and 0.949 respectively. Using values from sensitivity and precision, F1-score was calculated to be 0.961. The ensem- ble model correctly classifies 15,385 patches out of 16,000 test patches. This results to the final Balanced Accuracy (BAC) of 0.9616. Furthermore, Matthews Correlation Coefficient (MCC) was calculated to be 0.923. Cohen's kappa score stood at 0.923125, which indicates almost perfect agreement between the true labels and predicted labels during meta-analysis. The performance on training and validation directories per epoch wise are shown in Fig 5. The graph shows that the model has very low chances of over-fitting and thus guarantees an unbiased model. Correct classification functioning of the model is monitored using Matthews Correlation Coefficient (MCC) parameter.

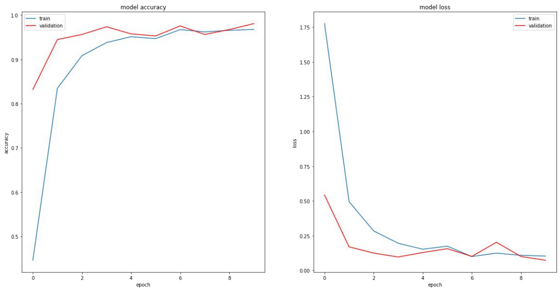

Fig. 5: Accuracy and loss graph vs epochs (Ensemble model)

615 patches out of 16,000 test patches are mis-classified. Firstly, this is due to the evident fact that the model was unable to classify all the images with a perfect scope of precision. Secondly, there might be a possibility of error from the domain expert itself, in labelling all 277,524 patches with perfect precision, since the task was tedious and time-consuming. The model might have predicted the images correctly, while the medical expert had some ground truths, false. Fig. 7 shows the breast histopathology tissue image of Patient ID: 10295, holding a ground truth of IDC positive. The tissue slice is visualized and the possible zone containing cancer tissue is highlighted in red. The third part shows the probability of cancer affected region in the tissue in colour gradients of red (low-high : light red-dark red). These visualizations could further help doctors and researchers to correctly identify IDC affected tissues, as well as the cancer affected region in that tissue.

TABLE III: Performance analysis using evaluation metric

Precision	0.949	False Discovery Rate	0.051
Sensitivity	0.973	False Negative Rate	0.027
Specificity	0.950	Matthews Corre Coefficient	0.923
Negative Predictive Value	0.974	Balanced Accuracy (BAC)	0.962
False Positive Rate	0.049	Cohen's kappa [27]	0.923
F1-Score [8]	0.961		

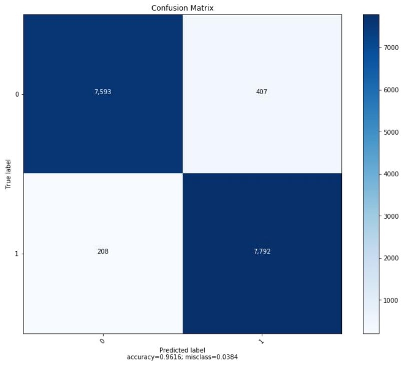

Fig. 6: Confusion matrix of Ensemble Model

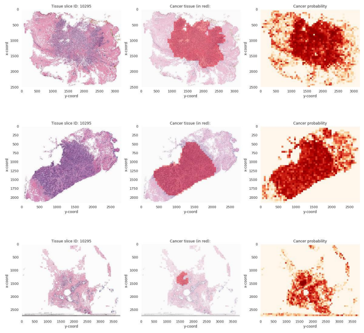

Fig. 7: Sample Predicted region

4.3 Conclusion and scope for future work

This paper introduces an ensemble framework generated using four distinct architectures which were selected for their exceptional performance in the binary classification of IDC from normal patches across 277,524 patches. The overall performance was projected on the basis of various distinct evaluation metrics such as precision, sensitivity, Matthews Correlation Coefficient (MCC), Cohen's kappa score and so on, as shown in Table III. The authors further encourage research in the domain of Deep Ensemble Networks to predict more ground truth to address this vulnerable disease.

References

1. C. DeSantis, R. Siegel, P. Bandi, and A. Jemal, "Breast cancer statistics, " *CA: a cancer journal for clinicians*, vol. 61, no. 6, pp. 408–418, 2011.
2. C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991.
3. S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in 2008 5th IEEE International Symposium on Biomedical Imaging: from Nano to Macro, pp. 284–287, 2008.
4. A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of pathology informatics*, vol. 7, 2016).
5. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
6. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
7. Z. Han, B. Wei, Y. Zheng, Y. Yin, & S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," Scientific reports, vol. 7, no. 1, pp. 1–10, 2017.
8. Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, and Y. Zhao, "Histopathological whole slide image analysis using context-based cbir," *IEEE Trans on Medical Imaging*, vol. 37, pp. 1641–1652, 2018

9. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
10. A. Pimkin, G. Makarchuk, V. Kondratenko, M. Pisov, E. Krivov, and M. Belyaev, “Ensembling neural networks for digital pathology images classification and segmentation,” in Int. Conf Image Analysis and Recognition, pp. 877–886, Springer, 2018.
11. S. Vesal, N. Ravikumar, A. Davari, S. Ellmann, and A. Maier, “Clas- sification of breast cancer histology images using transfer learning,” in *International conference image analysis and recognition*, pp. 812–819, Springer, 2018.
12. N. Brancati, M. Frucci, and D. Riccio, “Multi-classification of breast cancer histology images by using a fine-tuning strategy,” in*International conference image analysis and recognition*, pp. 771–778, Springer, 2018.
13. A. Nahid and Y. Kong, “Histopathological breast-image classification using local and frequency domains by convolutional neural network,” *Information*, vol. 9, no. 1, p. 19, 2018.
14. F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, “Deep features for breast cancer histopathological image classification,” in *2017 IEEE International Conference on Systems, Man, and Cyber- netics (SMC)*, pp. 1868–1873, IEEE, 2017.
15. Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, “Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images,” *Pattern Recognition Letters*, 2020.
16. G. J. S. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. D. Nagtegaal, I. Kovacs, C. H. van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak, “Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis,” *Scientific Reports*, vol. 6, 2016.
17. D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, “Deep learning for identifying metastatic breast cancer,” *ArXiv*, vol. abs/1606.05718, 2016
18. S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, “Classification of histopathological biopsy images using ensemble of deep learning networks,” *arXiv preprint arXiv:1909.11870*, 2019.
19. Y. Wang, B. Lei, A. Elazab, E.-L. Tan, W. Wang, F. Huang, X. Gong, and T. Wang, “Breast cancer image classification via multi-network features and dual-network orthogonal low-rank learning,” IEEE Access, vol. 8, pp. 27779–27792, 2020.
20. Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, “Mudern: Multi- category classification of breast histopathological image using deep residual networks,” *Artificial intelligence in medicine*, vol. 88, pp. 14– 24, 2018.
21. A. Cruz-Roa, A. Basavanhally, F. Gonza´lez, H. Gilmore, M. Feldman,S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, “Automatic detection of invasive ductal carcinoma in whole slide images with convo- lutional neural networks,” in Medical Imaging 2014: Digital Pathology, vol. 9041, p. 904103, International Society for Optics and Photonics, 2014
22. T. Arau´jo, G. Aresta, E. V. de Castro, J. F. Rouco, P. Aguiar, C. Eloy, A. Polo´nia, and A. Campilho, “Classification of breast cancer histology images using convolutional neural networks,” PLoS ONE, vol. 12, 2017.
23. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
24. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
25. N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: Practical guidelines for efficient cnn architecture design,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018