International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

TRANSFER LEARNING BASED LIGHTWEIGHT ESSEMBLED MODEL FOR IMBALANCED BREST CANCER

[1] *Jashwanth*, [2] *Asahi*
Department of Master of Computer Application
Graduate School of Management, Kyoto University

Received: 09-06-2022 Accepted: 14-7-2022 Published: 21-7-2022

ABSTRACT

Accurate breast cancer detection using automated algorithms remains a problem within the literature. Although a plethora of work has tried to address this issue, an exact solution is yet to be found. This problem is further exacerbated by the fact that most of the existing datasets are imbalanced, i.e. the number of instances of a particular class far exceeds that of the others. In this paper, we propose a framework based on the notion of transfer learning to address this issue and focus our efforts on histopathological and imbalanced image classification. We use the popular VGG-19 as the base model and complement it with several state-of-theart techniques to improve the overall performance of the system. With the ImageNet dataset taken as the source domain, we apply the learned knowledge in the target domain consisting of histopathological images. With experimentation performed on a large-scale dataset consisting of 277,524 images, we show that the framework proposed in this paper gives superior performance than those available in the existing literature. Through numerical simulations conducted on a supercomputer, we also present guidelines for work in transfer learning and imbalanced image classification. **Index Terms—**Transfer Learning, Imbalanced Class, Classification, Deep Learning.

I. INTRODUCTION:

Breast **Cancer** is one of most prevalent form of cancer among women worldwide. Despite the progress made in past decade in the field of cancer theranostics, this disease ranks second after lung cancer as world's leading causes of death according to the cancer mortality survey [1]. Accounting for about 80 per cent of all breast cancers, Invasive Ductal Carcinoma (IDC) is the most predominant subtype of breast cancer [2], [3]. In the IDC, the cells invade through the basement membrane into the surrounding stroma and are no longer confined to the affected duct [4]. With time, prognosis of IDC may worsen as the infiltrating cells metastasize to the lymph nodes and to other parts of the body. The regions with invasive cancer determine the aggressiveness of the disease. Early diagnosis significantly improves the chances of patient survival however, the process is very tedious, time-consuming and requires trained pathologists.

Diagnosing IDC generally involves procedures such as physical examination, mammography, magnetic resonance imaging (MRI), ultrasound, fine needle aspiration cytology (FNAC), histopathological examination of tissue sections. [5], [6]. Various non-invasive approaches such as SPR biosensor assays and multiplexed immunoassays have also been explored for serological detection of cancers, but none of these techniques is presently in clinical use [7], [8], [9]. If a patient presents a suspected region, histopathological examination of the biopsy is the only diagnostic procedure to confirm breast cancer with confidence. Histopathological examination of Hematoxylin and Eosin (H&E) stained tissue sections is usually done by trained pathologists and suffers from both intra-observer as well as inter-observer variability with an

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

average 68.39% interclass agreement between pathologists [10], [11]. In order to identify the presence of IDC, the tissue regions containing invasive regions needs to be differentiated from the non-invasive tissues. Recent advancements in the field of oncology and computer engineering has also supported to the development of computer-aided diagnosis (CAD) systems for helping the pathologists in the diagnosis procedure and reducing their workload. Accurately identifying and categorizing invasive and non-invasive tissues is an important clinical task, and automated methods can greatly speed up diagnosis, reduce errors and lead to better healthcare and treatment. Therefore, researchers realized the potential that automated systems could have in this regard. As a consequence, there is a significant research going on medical imaging, e.g. [12], [13].

This paper, therefore, attempts to classify IDC based on histopathological images using state-of-the-art machine-driven algorithms. If an effective system for classifying IDC and non-IDC images could be devised, the scientific and societal impact would be huge. This however is easier said than done. The issue is further exacerbated by the fact that existing datasets, e.g. [14], [15], [16], are highly imbalanced. Although the problem of imbalanced datasets and imbalanced class classifications is not new, and significant efforts have been made to overcome this issue, e.g. [17], [18], [19], [20], an efficient machine-driven solution has yet to be devised. Therefore, the first challenge in histopathological image classification is: How to handle the issue of imbalanced classes?

Recently, image classification based on convolutional neural networks (CNNs) has made several strides forward. The field of medical image processing was challenged by the invention of CNN, and work in this direction used deep models that pushed the system's performance to never before seen results. There are multiple documented instances where CNNs have given excellent results [12], [21], [22]. However, they are with some limitations. CNNs require a large set of annotated images [23], which is especially problematic considering the fact that training a large CNN takes a lot of time. Further, the availability of test subjects, in context of medical images, is often limited. The former challenge is well known in deep learning, and there is very little that we can do here to address the problem. For the latter issue, a solution can be found via the paradigm of transfer learning. The rationale here is backed by previous literature, e.g. [23], [24], [25], [26], where multiple authors have argued that fine-tuning, as well as, using a pre-trained model, could improve the overall performance of the system. Consequently, pre-trained models quickly gained attention compared to trainedfrom-scratch models, and the so-called field of transfer learning expanded even further. In transfer learning based computation, we have two different application domains: the source domain and the target domain. A framework using transfer learning trains architecture in one area (source) and applies the learned knowledge in the other domain (target). This naturally has several advantages: 1) training time is significantly reduced; 2) the problem of imbalanced class classification is solved to great extent [27], [28].

In light of the challenges and the potential solution discussed in this section, we used the paradigm of transfer learning to classify histopathological images. The work presented in this study, outperforms the existing literature regarding IDC classification, e.g. [16], [29], [30]. We utilized the existing VGG-19 [21] as the pretrained model and applied the learned knowledge in the target domain consisting of histopathological images. Additionally, we modified the last layer of the CNN and tried

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

different permutations of classification methods to find out the best and numerically superior combination. By comprehensively inspecting the dataset taken from [14], [15], [16], we demonstate that the work presented in this paper enhances the existing knowledge in several ways. The accuracy of the proposed work is 90.3%, the F1 score is 93.22, sensitivity is 93.31, specificity is 82.75, and BAC is 88.03. These figures emphasise this work's superior performance. The following points summarise the contribution of this paper:

1) We used the paradigm of transfer learning to classify imbalanced histopathological images taken from [14], [15], [16].
2) We used VGG-19 as the base model and complement it via the use of different classification schemes.
3) Through extensive numerical simulations conducted on supercomputer, we analysed the framework in multiple ways. Based on this, we present additional guidelines for work involving image processing and imbalanced class classification.

II. RELATED WORK

This paper deals with imbalanced class classifications in the domain of histopathology. There is already a large amount of literature that attempts to address this issue, for example, the work proposed in [17], [31] used semisupervised learning in sentiment classification, while the authors of [32] used a support vector machine (SVM) to classify text. The authors of [20] used emsemble of classifiers to handle imbalanced classes. Variants of SVMs have also been used to try and classify data [33], [34]. The study presented in [18] used an ensemble of cost-insensitive trees (decision trees), and the work in [35] addressed the case pertaining to borderline instances in an imbalanced dataset. The ideas discussed in [36] went one step further and transformed an imbalanced classification problem into a two-class symmetrical problem. The work in [37] took the idea to the next level and discussed the behaviour of receiver operating characteristics (ROC) for selecting optimal classifiers. Along the same lines, the authors of [38] proposed a comprehensive framework that uses threshold metrics and rank metrics to evaluate datasets that are highly skewed. Based on their analysis, they recommended skew-normalised scores. Our work is similar as we also try to address the issue of imbalanced classes, however, we use the paradigm of transfer learning and apply the model to histopathological images.

For the past few years, deep learning (DL) has been significantly applied in image classification. It has also been applied to a wide range of other fields, including automatic speech recognition, image recognition, natural language processing, drug discovery, and bioinformatics [43], [44], [45]. Furthermore, the evolution of DL gave birth to CNNs, which have shown excellent results in image processing. In 2012, the famous AlexNet CNN was proposed [22] and showed remarkable results when used on an ImageNet dataset. Following this success, DLbased approaches using CNNs have begun to demonstrate impressive performance. This even extended to the field of medical tasks [44], [46]. In the literature, different CNN-based architectures have been used in the diagnosis of IDC cancer. Using CNNs and random forest as the classifier, the work in [16] differentiated IDC tissues from the cancerous breast area with a classification accuracy of 84.23. Early and late stage classification of IDCs with a success rate of 86% has been achieved using supervised machine learning methods in [47], and the histopathological image classification of five breast cancer subtypes using a pipeline of four convolutional networks and an SVM was reported as having achieved 55% accuracy [48]. The authors in [49] categorised invasive and noninvasive breast cancer histology images and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

following this work, deeper and more precise techniques were demanded.

TABLE 1: State-of-the-art techniques using the dataset given in [14],[15],[16].

Method	Data Sample Size	TL Employed	Imbalance / Balanced	Performance Metrics	
	Partial (p) / Complete (c)	Yes/ No	Dataset Type	Accuracy (%)	F1-Score
[26]	277524 (c)	No	Imbalanced	85.48	84.79
[39]	7500 (p)	No	Nearly Balanced	N.A.*	N.A.*
[40]	1650.98 (p)	No	Imbalanced	N.A.*	N.A.*
[29]	277524 (c)	No	Imbalanced	-	85.20
[30]	277524 (c)	No	Imbalanced	88.33	77.94
[41]	157072 (c)	No	Balanced	N.A.*	N.A.*
[42]	275601 (p)	No	Imbalanced	N.A.*	N.A.*

* The accuracy given in these studies is not applicable for comparison because these studies have used only partial datasets and not the complete datasets. Therefore, the accuracy reflected in these studies is biased and not suitable for comparison with studies who have worked upon the complete datasets.

This resulted in the VGG models, the architecture of which were introduced in [21]. The uniform architecture of VGG models make them very appealing and they also outperform baselines on many tasks outside of ImageNet. Recently, various VGG-based architectures have shown promise for histopathological image analysis [50], [51]. For example, VGG-19 has been investigated for human breast cancer [51], colorectal cancer [52], [53], and skin cancer [54]. VGG-19 complemented with principal component analysis (PCA) and singular value decomposition (SVD) has also been used in fundus image classification [55]. The authors of [56] used a VGG model to recognize structural damage, and the work presented in [57] used the network to classify high-res satellite images. In short, the idea of using VGG-19 is not new, and there is significant literature related to solving the issues related to this model. We therefore lay the foundation of our work within this network and try to classify IDC vs non-IDC images.

To summarize the contribution of this work in brief, Table 1 shows the comparison of the proposed method with the state-of-the-art. It should be noted here that the numbers for some of the techniques presented in the Table are excluded as their sample size is much smaller than ours. One can understand that scale disrupts performance. Therefore, the numbers are presented for techniques that have utilized the complete data presented in [14], [15], [16].

III. METHODS

This section give the detailed description of complete method in multiple subsections and the overall workflow of this section is as follows:

1) Brief paradigm of transfer learning is discussed.

2) We then present a discussion on CNNs.

3) Lastly, we discuss the framework followed in the paper.

3.1 Transfer Learning

The core of transfer learning revolves around training a model in the source domain (Ds) and applying the learned knowledge in the target domain (Dt) [58]. Both the source and target domains consist of labelled data. To understand the objective of transfer learning, consider the domain Ds. Here, Ds has the following data points: {x 1 , y1}, {x 2 , y2}, ..., {x n, yn}, where x i ∈ X is the training data. In our case, x i is the input image and yi ∈ Y is the classification label. It should be noted here that the problem addressed in this paper is a binary class classification problem. The purpose henceforth is to devise an automated mechanism to learn the conditional probability function, p(y t |x t). To do this, transfer learning uses the existing distribution function, fs(.), to find the value for ps(y t |x t). It should be noted that the function fs(.) is learned while training a network/model in the source domain. As expected, there can be a variety of situations in transfer learning. Therefore, it could happen that the source domain's input set of features is different than the target domain's features, e.g. ∀ix S i 6= x T i and pT (yt|xt) 6= pS(ys|xs). Other situations include the set of features in the source domain being equal to that of the target domain, e.g. ∀ix s i = x t i , however, ps(y t |x t) 6= ps(y s |x s), and the source domain and the target domain being exactly the same, e.g. ∀ix s i = x t i and ps(y t |x t) = ps(y s |x s). The latter is one of the classic situations of machine

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

learning and is extensively investigated in the literature [58].

To apply transfer learning in practice, we have to answer a few questions:

1) What should be transferred?

2) How should the knowledge be transferred?

To answer the first question, a system designer has to use their judgement and apply extensive feature engineering. For the second question, the problem is model selection and how to complement it to make the predictions more accurate. For this purpose VGG-19 is used [21].

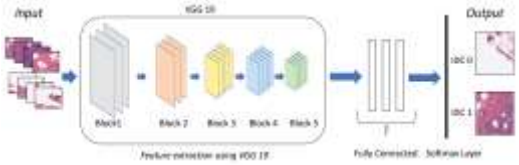

Fig. 1: VGG-19 Architecture

Proposed Framework for Feature Extraction and Classification

In this subsection, we discuss the proposed framework followed in the paper. It should be noted that the literature points to the fact that transfer learning could be used to improve the performance of the system [23], [24], therefore, we employ the idea of CNN-based transfer learning. Moreover, we try different permutations of classification schemes to find the best overall performance.

In Fig. 1 and Fig. 2, the overall architecture of the work is presented. As we used the paradigm of transfer learning, the existing architecture presented in [21] was used. This framework is commonly referred to as VGG19 and is a 19-layer deep neural network that has been used extensively in the literature [60], [61]. The base network of VGG-19 is presented in Fig. 1. This network was trained on 1.2 million images taken from the ImageNet database, therefore, the ImageNet database was the source domain (Ds) and the set of histopathological images we were trying to classify was the target domain (Dt). Due to VGG-19's popularity and its extensive use in transfer learning, we used this network as the pretrained framework in this paper. Moreover, we also proposed few modifications which are discussed in the following paragraph.

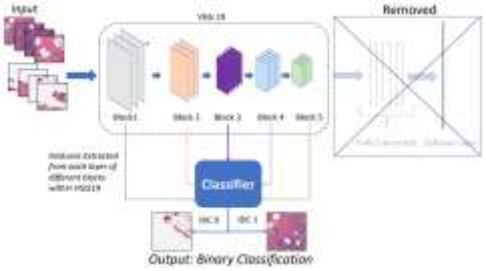

Fig. 2: Proposed framework based on VGG-19 architecture

IV. RESULTS AND DISCUSSION

In this section, we evaluate and compare the performance of the model against multiple existing framework.

It was specified in Section 1 that we experimented with the dataset taken from [14], [15], [16], which consists of 277,524 images in RGB format. The patch size of the images was 50 X 50. This dataset was well known for having imbalanced classes, and numerically speaking, the number of IDC classes was 198,738 compared to 78,786 healthy tissue classes. The major objective of this paper is to classify IDC vs non-IDC images by specially considering the imbalanced classes. A few sample images from the dataset are shown in Fig. 3.

To demonstrate the superiority of the method, we used the following metrics: F1 score; accuracy; sensitivity; specificity; BAC; G-means; and AUC. These were defined as follows:

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

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

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

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

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

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$BAC = 0.5 * \left(\frac{TP}{TP + TN} + \frac{TN}{TN + TP}\right)$$

$$G - Mean = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{TN + FP}}$$

where TP is true positive, FP is false positive, TN is true negative, and FN is false negative. We used the above evaluation measures as they were some of the standard metrics used in the literature for imbalanced class classification [16]. The experiments were conducted on an Nvidia DGX V-100 with the following specifications: 8X NVIDIA Tesla V100 16 GB/GPU 40,960, 5,120 Tensor Cores, 512 GB RAM, 4X 1.92 TB SSDs, and 20- Core Intel Xeon E5-2698 v4 2.2 GHz.

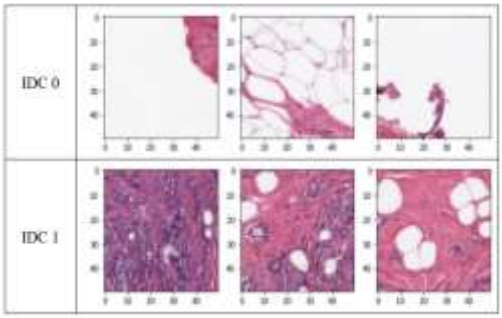

Fig. 3: Sample images in the dataset

TABLE 2: Comparison with state of the art techniques.

Methods	Sensitivity	Specificity	F-score	Accuracy	BAC
Imbalanced data [16]	0.9005	0.7396	0.8917	0.8501	0.8201
Under-Sampling [16]	0.7252	0.8667	0.7804	0.7959	0.796
Over-Sampling (WR) [16]	0.8223	0.9127	0.8613	0.8613	0.8675
ADASYN [16]	0.7753	0.8492	0.8	0.8137	0.8123
SMOTE [16]	0.909	0.8034	0.8634	0.8562	0.8562
[29]	0.86	0.85	0.85	-	0.8541
[30]	0.796	**0.9466**	77.94	88.33	83.54
Proposed Work	**0.9331**	0.8275	**0.9322**	**0.903**	**0.8803**

It should be noted that although several papers tried to classify the dataset presented in [14], [15], [16], however, they only considered a subset of the complete dataset. It is expected that considering the scale of the experiments, i.e considering all 277,524 images, the results are bound to deteriorate. In this regard, and to the best of our knowledge, the results presented in [16], [29], [30] are by far the closest to the work presented here in terms of scale of experimentation. In the following sections, we therefore attempt to show that the proposed model improves upon the work in [16], [29], [30] in several ways.

4.1 Comparison with State-of-the-art Models

To demonstrate the efficacy of the proposed technique, the comparative results are shown in Table 2. Furthermore, the number of features extracted from multiple layers of VGG-19 are presented in Table 2. From the evidence presented in Table 2, we can see that the work in this paper gave a superior performance considering the multiple evaluating criteria. It should be noted that some of the techniques used for comparison were comprehensively explained and used in the work presented in [16]. It can also be seen that the proposed model is able to outperform these existing methods in terms of multiple criteria. Although the specificity did not improve, the framework beat the previous work in four out of the five criteria. The results and figures discussed here show the superiority of the model compared to the existing work. This result is especially promising when we consider the imbalanced classes. Moreover, the scale of the experiment gave necessary insights into histopathological image classification, which are comprehensively discussed in Section 4.5.

4.2 Comprehensive Evaluation: A Random Forest Approach

Following the procedure discussed in Section 3.3, we present the comprehensive results of the experiment in Tables 4. The classification model

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

chosen was random forest as this is one of the best performing algorithms when dealing with high dimensional data [62]. Furthermore, random forest is robust to overfitting, and the parametrisation remains quite intuitive and straightforward. Therefore, the results with random forest as the base framework for classification are presented in Table 4. This table demonstrate the overall good performance of the methodology. Although the improvement was small, the proposed work was better than previous literature in a few criteria. In addition, the ROC curves for a few layers are presented in Fig. 4. According to the ROC analysis, the AUC value of the proposed work ranged from 0.7929 to 0.8893, which is a good indication that the technique performed well. Moreover, the evidence emphasises the ability of the framework to distinguish between the two classes.

TABLE 3: Number of Features extracted from different layers of VGG19

Intermediate Layers of VGG19	Number of Extracted Features	Intermediate Layers of VGG19	Number of Extracted Features
block1_conv1	160000	block4_conv1	18432
block1_conv2	160000	block4_conv2	18432
block1_pool	40000	block4_conv3	18432
block2_conv1	80000	block4_conv4	18432
block2_conv2	80000	block4_pool	4608
block2_pool	18432	block5_conv1	4608
block3_conv1	36864	block5_conv2	4608
block3_conv2	36864	block5_conv3	4608
block3_conv3	36864	block5_conv4	4608
block3_conv4	36864	block5_pool	512
block3_pool	9216		

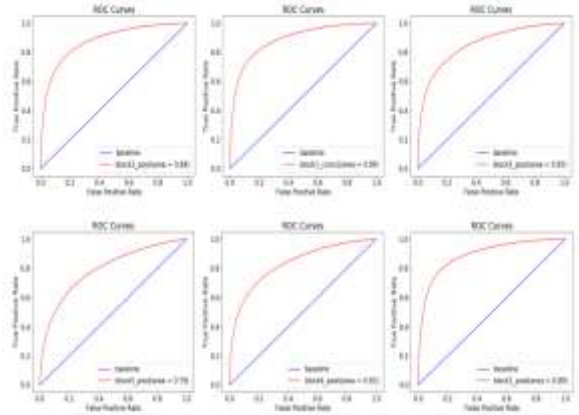

Fig. 4: ROC curves for a few layers. The layers are as follows: i) Block2 pool layer. ii) Block1 conv. iii) Block3 pool. iv) Block5 pool. v) Block4 pool. vi) Block1 pool.

4.3 Additional Experiments and Analysis I

As this study used the idea of transfer learning [63], [64]. In these papers, the authors argued that there is no guarantee that transfer learning will improve performance. Indeed, there are documented instances where transfer learning actually degrades performance. In this context, and considering the work presented here the values presented in Tables 4 shows that stacking additional blocks leads to performance degradation. To be specific, and considering the metric of precision, after block5 conv2 the performance worsens. Similar cases can also be seen for the metrics of recall (after block conv4) and specificity (after block4 conv1). During the analysis, we found that this was due to the classic issue of negative transfer [63], [64]. As discussed in this section, negative transfer is a well-accepted notion of literature where one sees performance degradation. This is acceptable, however, as we are able to complement the existing work by improving the performance of the dataset.

Another possible reason for a poor performance could be the number of features. According to Tables 3 and 4, block1 conv1 had the maximum number of features, while block5 pool had the lowest number of features. At these layers, we did not see the best performance. Instead, the best figures were obtained at an intermediate layer. When considering all the performance criteria, the best results for sensitivity, G-means, etc. were obtained at different layers. This indicates that the balance between the performance, the set of features, and the computation time needs to be maintained.

TABLE 4: Performance evaluation after feature extraction from different layers of VGG19 +

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Random Forest

Layer	TP	FP	FN	TN	Precision	Recall/Sensitivity	Specificity	AUC	GMean	F1 Score	Accuracy	BAC
block1_conv1	5583	3801	9966	14600	0.9563	0.8617	0.8617	0.8872	0.7774	0.8974	0.8467	0.8279
block1_conv2	5560	3831	8649	14781	0.9559	0.8622	0.8622	0.8893	0.7779	0.8980	0.8479	0.8287
block1_pool	5564	3795	8696	14930	0.9566	0.8653	0.8653	0.8925	0.7517	0.8995	0.8501	0.8316
block2_conv1	5609	3542	9606	14022	0.9567	0.8558	0.8558	0.8912	0.7514	0.8951	0.8421	0.8262
block2_conv2	5639	3414	9959	13967	0.9425	0.8495	0.8495	0.8806	0.7566	0.8956	0.8392	0.8204
block2_pool	5648	3294	10367	12259	0.9415	0.8447	0.8447	0.8850	0.7385	0.8902	0.8362	0.8222
block3_conv1	5806	2826	11465	12161	0.9527	0.8321	0.8321	0.8694	0.7276	0.8885	0.8294	0.8218
block3_conv2	5696	2636	12218	11408	0.9558	0.8276	0.8276	0.8871	0.7123	0.8868	0.8256	0.8179
block3_conv3	5702	2620	12081	11545	0.9561	0.8252	0.8252	0.8594	0.7153	0.8858	0.8235	0.8200
block3_conv4	5707	2603	12263	11463	0.9564	0.8233	0.8240	0.8853	0.7136	0.8854	0.8227	0.8196
block3_pool	5697	2657	12266	11360	0.9555	0.8229	0.8229	0.8829	0.7138	0.8843	0.8208	0.8167
block4_conv1	5692	2728	12674	10952	0.9544	0.8179	0.8179	0.8806	0.7101	0.8809	0.8172	0.8095
block4_conv2	5665	2667	12611	11015	0.9555	0.8188	0.8188	0.8426	0.7029	0.8815	0.8165	0.812
block4_conv3	5703	2619	13280	10346	0.9561	0.8111	0.8111	0.8315	0.688	0.8777	0.8091	0.8066
block4_conv4	5718	2451	13529	10097	0.9589	0.8087	0.8087	0.8506	0.6859	0.8774	0.8081	0.8067
block4_pool	5715	2517	13361	10265	0.9578	0.8035	0.8035	0.832	0.6869	0.875	0.8091	0.8068
block5_conv1	5694	2638	13680	9946	0.9559	0.8065	0.8065	0.8219	0.6796	0.8748	0.8041	0.7985
block5_conv2	5686	2826	13485	10141	0.9527	0.8082	0.8082	0.8211	0.6846	0.8745	0.8041	0.7952
block5_conv3	5658	3064	13141	10485	0.9487	0.8015	0.8015	0.8225	0.6876	0.8746	0.8054	0.7927
block5_conv4	5661	3021	13861	9765	0.9494	0.8036	0.8034	0.8115	0.6718	0.8700	0.7973	0.7936
block5_pool	5675	3407	14298	9328	0.9421	0.7972	0.7972	0.7929	0.6579	0.8656	0.7868	0.7634

TABLE 5: Comparison with different classifiers

Frameworks	VGG19 + Logistic Regression	VGG19 + SVM(Linear)	VGG19 +Random Forest	VGG19 with training the dense layers	**Retraining the VGG19 with dense layers**
TP	5642	5934	5675	5190	5556
FP	3279	479	3657	7591	496
FN	18753	22993	16296	6963	3988
TN	4673	711	9528	16663	19628
Precision	0.9452	0.992	0.9425	0.8711	0.9314
Recall/Sensitivity	0.7504	0.721	0.7972	0.8818	0.9331
Specificity	0.5985	0.6647	0.7297	0.8962	0.8275
F1 Score	0.8366	0.8355	0.8638	0.8764	0.9322
Accuracy	0.7355	0.7583	0.7668	0.826	0.903
BAC	0.6744	0.6629	0.7634	0.765	0.8803

4.4 Additional Experiments and Analysis II

In Section 4.3, we applied transfer learning with VGG-19 as the base model and complemented it with random forest. In this subsection, additional experimentation with multiple techniques are performed. The classification model was changed to SVM, fully connected layers, and logistic regression, and we retrained VGG-19 with deep layers. The rest of the setup remained the same. The results of the experiment using this architecture are presented in Table 5. These results show that the best performance was obtained when VGG-19 was complemented with deep layers. Specifically, performance improvement was 5.58% in terms of F1 score, 9.78% for specificity, 5.13% for sensitivity, 7.9% for accuracy, and 9.73% for BAC. Although the level of precision did not improve, this is acceptable as we are dealing with imbalanced classes. These results show the superiority of deep layers compared to other classification methods.

In Fig. 5, we have shown the result regarding accuracy. The figure shown here was obtained via fully connected layers and by retraining the model on histopathological images (VGG-19 was originally trained on the ImageNet dataset). Furthermore, in the subsequent mode without retraining, the original weights of VGG-19 were left untouched (no retraining on histopathological images). As shown in Fig. 5, the best results were obtained when the model was retrained on histopathological images. Moreover, there was quite a big difference between the performances of the two modes, which again shows that retraining the model gave good results.

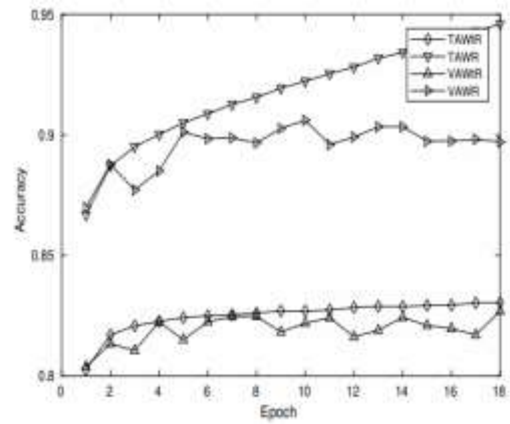

Fig. 5: Accuracy vs Epoch. TAWR: Training Accuracy with Retraining. TAWtR: Training Accuracy without Retraining. VAWR: Validation Accuracy with Retraining. VAWtR: Validation Accuracy without Retraining.

V. CONCLUSION

In this paper, we attempted to classify large-scale histopathological images using automated machinedriven procedures. This task was complicated by the fact that the dataset [14], [15], [16] was highly imbalanced. To address this issue, we used the paradigm of transfer learning and trained the existing VGG-19 on more than a million ImageNet images before applying the learned knowledge to the dataset of histopathological images. The model was further complemented with different classification methods at the output layer. Through extensive numerical simulations, it was shown that the work in this paper augments the existing knowledge of classifying imbalanced datasets in

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

multiple ways. Furthermore, the classification performance of the technique was compared with existing frameworks. Considering the scale of the experiments, it is expected that future work, whether in TL or not, could improve upon the base performance following the guidelines presented in the article. Further studies are required to investigate the effect of stain normalization, data augmentation and the use of different classifiers on the performance of the proposed framework. Having established the base performance of the proposed transfer learning based approach, future studies needs to be directed towards further improvement in the performance of the method and validation of its efficacy on more number of datasets.

REFERENCES

1. I. A. of Cancer Research (2018) Latest Global Cancer Data:, "Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018." World Health Organization, Geneva. Available at https://www.who.int/cancer/PRGlobocanFinal.pdf (2019/11/11).
2. G. K. Malhotra, X. Zhao, H. Band, and V. Band, "Histological, molecular and functional subtypes of breast cancers," Cancer biology & therapy, vol. 10, no. 10, pp. 955–960, 2010.
3. F. A. Tavassoli, "Pathology and genetics of tumours of the breast and female genital organs," World Hhealth Organization Classification of Tumours, 2003.
4. M. F. Lerwill, "Current practical applications of diagnostic immunohistochemistry in breast pathology," The American journal of surgical pathology, vol. 28, no. 8, pp. 1076–1091, 2004.
5. B.-N. Zhang, X.-C. Cao, J.-Y. Chen, J. Chen, L. Fu, X.-C. Hu, Z.- F. Jiang, H.-Y. Li, N. Liao, D.-G. Liu et al., "Guidelines on the diagnosis and treatment of breast cancer (2011 edition)," Gland surgery, vol. 1, no. 1, p. 39, 2012.
6. C. Pagani, D. Coscia, C. Dellabianca, M. Bonardi, S. Alessi, and F. Calliada, "Ultrasound guided fine-needle aspiration cytology of breast lesions," Journal of ultrasound, vol. 14, no. 4, pp. 182–187, 2011.
7. S. Hussain, S. Saxena, S. Shrivastava, R. Arora, R. J. Singh, S. C. Jena, N. Kumar, A. K. Sharma, M. Sahoo, A. K. Tiwari et al., "Multiplexed autoantibody signature for serological detection of canine mammary tumours," Scientific reports, vol. 8, no. 1, p. 15785, 2018.
8. S. Mittal, H. Kaur, N. Gautam, and A. K. Mantha, "Biosensors for breast cancer diagnosis: A review of bioreceptors, biotransducers and signal amplification strategies," Biosensors and Bioelectronics, vol. 88, pp. 217–231, 2017.
9. S. C. Jena, S. Shrivastava, S. Saxena, N. Kumar, S. K. Maiti, B. P. Mishra, and R. K. Singh, "Surface plasmon resonance immunosensor for label-free detection of birc5 biomarker in spontaneously occurring canine mammary tumours," Scientific reports, vol. 9, no. 1, pp. 1–12, 2019.
10. A. F. Taktak and A. C. Fisher, Outcome prediction in cancer. Elsevier, 2006.
11. W. H. Wolberg, W. N. Street, and O. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," Cancer letters, vol. 77, no. 2-3, pp. 163–171, 1994.
12. H. R. Roth, C. T. Lee, H.-C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, and R. M. Summers, "Anatomy-specific classification of medical images using deep convolutional nets," in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE, 2015, pp. 101–104.

13. G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin et al., "Interactive medical image segmentation using deep learning with imagespecific fine tuning," IEEE transactions on medical imaging, vol. 37, no. 7, pp. 1562–1573, 2018.
14. A. Cruz-Roa, A. Basavanhally, F. Gonzalez, H. Gilmore, M. Feld- ´man, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in Medical Imaging 2014: Digital Pathology, vol. 9041. International Society for Optics and Photonics, 2014, p. 904103.
15. 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.