IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 October 2023, accepted 15 October 2023, date of publication 23 October 2023, date of current version 1 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327221

RESEARCH ARTICLE

Lightweight Histological Tumor Classification Using a Joint Sparsity-Quantization Aware Training Framework

DINA ABOUTAHOUN[1], (Member, IEEE), RAMI ZEWAIL[1], KEIJI KIMURA[2], AND MOSTAFA I. SOLIMAN[1,3]

[1]Department of Computer Science and Engineering, Egypt–Japan University of Science and Technology, New Borg El-Arab City, Alexandria 21934, Egypt
[2]Department of Computer Science and Engineering, Waseda University, Tokyo 169-8555, Japan
[3]Department of Electrical Engineering, Aswan University, Aswan 81528, Egypt

Corresponding author: Dina Aboutahoun (dina.aboutahoun@ejust.edu.eg)

ABSTRACT Cancer decision-making is a complex process that can be exacerbated by the limited availability of oncological expertise. This is particularly true in rural areas and settings with fewer resources. Recently, there has been an interest in the potential of artificial intelligence in reliable computer-aided diagnosis tools in such settings. Nevertheless, the majority of deep learning algorithms are resource hungry in terms of data and storage requirements. In this work, we propose a novel lightweight deep learning model for histological tumor classification through a Joint Sparsity-Quantization Aware Training framework. Extensive experiments were conducted to evaluate the proposed framework. Promising performance has been achieved compared to the most relevant state-of-the-art work with a classification accuracy of 94.26% and an average 5× reduction in the memory footprint. This work aims at opening doors toward efficient point-of-care diagnostic devices suitable for environments with limited resources.

INDEX TERMS Deep learning, histopathology, quantization, pruning, transfer learning, medical image analysis.

I. INTRODUCTION

The widespread use of deep learning models has many implications across multiple disciplines. Models are becoming larger in size with massive amounts of data collected and fed to them, requiring tremendous computing power and energy. There is a current need to counteract the negative consequences of traditionally large models by looking into methods that scale back resource consumption. Deep learning that is economical in size and resource consumption is an attractive prospect, as it cuts back on cost, is accessible to more devices, and reduces the carbon footprint. The lower cost of entry incentives deployment to constrained resource devices, which can reach communities that cannot afford the cost of higher-end devices and expensive training. This is especially important in medical devices that facilitate the clinical decision-making process which is needed in areas with a deficiency in medical expertise.

Model compression is an area that is receiving a lot of attention due to the potential gains in conserving memory and computation, coupled with the interest and traction the Internet of Things (IoT) and mobile devices have gained in recent years, hence expediting the demand for efficient models. Quantization, pruning, and knowledge distillation are the main directions concerned with balancing computational costs with the associated consequence of accuracy degradation which are crucial factors for enabling deployment to resource-constrained devices. Deep learning applied to medical data has the potential to improve diagnostic accuracy and subsequent patient prognosis through early detection. Deep learning has a track record of being used in many medical applications, for example in drug prediction [1], medical image segmentation [2], cancer classification [3], abnormal speech recognition [4], and others.

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeswari Sundararajan.

Examination of histopathological images remains one of the essential methods for diagnosing cancer. Despite it being an effective diagnostic procedure, expert knowledge about the disease is vital for a valid interpretation [5]. Moreover, histopathological examination is liable to the subjective interpretation of the pathologist which is dependent on their level of experience, thus leading to variations in patient assessment [6]. Therefore, the need for a second opinion is often merited, however, it may be difficult and time-consuming to obtain. This has created a niche for computer-aided diagnosis systems for rapid diagnosis in which deep learning has shown great promise [7], [8].

Despite the history of using deep learning in medical applications, the availability of medical data is severely restricted in a legal and natural sense. This is at odds with the data-driven nature of deep learning. Laws and ethics regulating patient data might be a hindrance to widespread data collection, alongside the privatization of medical datasets. Until now there is still no standardized protocol in place for data collection, nor consistency in data quality [9]. In conjunction, the fact that some diseases manifest more rarely in the population makes it impossible to create equal classes of diseases [10]. Additionally, medical data needs to undergo annotation by experts, making abundant, high quality and annotated medical data a cumbersome goal to achieve. We arrive at a point where traditional deep learning's superior abilities in pattern recognition, which are especially important in tasks dependent on recognizing abnormalities, sometimes reported to be arguably superior to that of human experts [11], [12], [13], are being undermined by the scarcity of medical data [14].

Motivated by the extensive computational and data resources required for deep learning, in this work we aim to provide an efficient framework that can aid medical professionals in cancer decision support which could be suitable for deployment to resource-constrained devices such as EDGE and IoT devices for healthcare applications. The contribution of this paper can be summarized as follows:

- A Joint Sparsity-Quantization Aware Training (JSQAT) framework is proposed for lightweight histopathological classification. The approach presented interleaves transfer learning, sparsity, and quantization techniques through the training process.
- Extensive experiments were conducted to evaluate the impact of sparsity and quantization within the context of malignant tumor detection in breast histopathological images.
- We empirically evaluate the proposed framework on the BreakHis dataset and observe that JSQAT results indicate the possibility of a performance/memory trade-off that balances the needs of a medical classifier with the memory bounds of a resource-constrained environment. We compare our results with other lightweight models oriented toward breast cancer histopathological classification.

II. RELATED WORK

The problem of histopathological classification is of great interest across multiple disciplines. Cancer diagnosis remains a challenging task for pathologists given cancer's heterogeneous nature [15]. Cancer is varied in its types and classifying malignant tumors in images is a nuanced problem. Diagnostic accuracy and interpretability are priorities in medical applications which also apply to histological classification [16], [17]. However, the majority of the literature on medical images is confined to the traditional deep learning paradigm that is both computationally and data intensive.

There is a need to address the demand for efficient models in histological imaging and medical imaging in general, to enable performant models using limited resources [18]. This can contribute to more cost-effective solutions and wider accessibility especially in resource-limited healthcare facilities that suffer from insufficient infrastructure to support large models and a shortage of medical specialists. We believe that compression methods such as pruning and quantization in addition to transfer learning can allow for resource-efficient models to be integrated into clinical workflows in low-resource settings.

The reviewed literature prominently features CNNs, which are exceptional in pattern detection and suited for navigating the challenging topology of histological images. Previous works that tackle the problem of breast cancer classification can be grouped into ensemble or fused models [19], [20], lightweight models, or other various deep learning or machine learning models. Models in the literature can be also classified according to being magnification agnostic [21], [22] or magnification dependent [23], or whether data augmentation was used [24].

Another prominent approach is transfer learning based methods, typically using ImageNet as the base dataset. As an example, an approach followed in [25], uses a pre-trained EfficientNetV2 [26] backbone with a modified dual squeeze and excitation network for binary classification of BreakHis images. The authors report a precision of 0.9858 and an F1-score of 0.9764 at 40X magnification. Using a pre-trained network is also attempted in [27], where the authors devise an approach based on using a MobileNet network [45] with a Support Vector Machine named MobileNet-SVM. This yielded an accuracy of 91% on the 400X magnification of BreakHis.

A subset of medical image classification research is taking resource efficiency into consideration, however, those that tackle histopathological images are significantly fewer in the literature as the main body of literature does not prioritize resource/performance trade-offs. This could be due to the anticipated impact on diagnostic accuracy from reducing computational resources. However, striking the balance between the two will allow computer-aided diagnosis systems to reach low-resource devices and communities.

We wish to highlight efforts in this direction as the following reviewed works are conceptually closer to the goal we aim to achieve. In [28], the authors used post-training

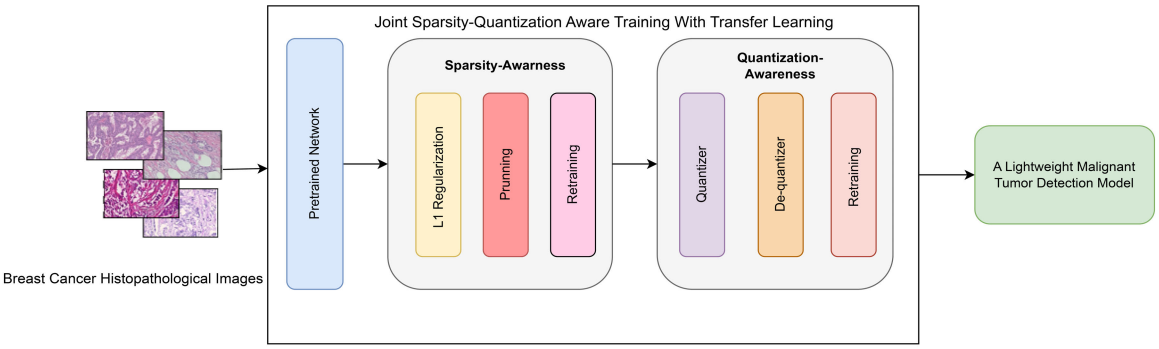

FIGURE 1. Diagram of the methodology of the system. The pre-trained model undergoes three stages of training; standard fine-tuning, sparsity-aware training, and quantization-aware training in order to adapt it to the target dataset with the goal of subjecting it to the compound effect of sparsity and quantization.

quantization (PTQ) to compress their model MobiHisNet. The results indicate a 3.3× compression at an accuracy of 87.31% on 40X magnification BreakHis images [29] and 89.94% accuracy for 16-bit PTQ. In another work [30], the authors followed a structured pruning approach by filtering less important filters based on their absolute sum. The pruned VGG19 [31] based model achieved 90.08% accuracy on the magnification-independent binary classification task. The authors also reported a 47.54% parameter reduction and 63.46% FLOPs reduction. In [32], the authors applied structured pruning on a hybrid Inception network [33]. Image pre-processing and data augmentation were applied to the dataset. The accuracy for 40X magnification is 85.7 ±1.9% when 50% of the channels are pruned.

While not widely attempted in the medical literature, there has been an interest in exploring ways to combine quantization and pruning to achieve lighter models mainly in the computer vision community. This combination has been implemented on medical images in [34]. The authors devise a toolbox for producing efficient medical models for constrained hardware where they provide customizable pruning and quantization options. Of the applications tested, the performance of the toolbox on 2D nuclei segmentation task achieved a Dice score of 0.567.

In computer vision, attempts were made to combine pruning and quantization on general datasets [35], [36], [37], such as MNIST [38], ImageNet [39], CIFAR-10 [40], and ILSVRC-2012 [41]. Results reported in these works include 1,910× and 210× size reduction [35] using LeNet-5 [42] and AlexNet [43] on MNIST and ImageNet respectively. With a 94.09% accuracy at 53.9% sparsity reported in [36] using ResNet [44] on CIFAR-10. Alongside 92.23% accuracy for 50% weight sparsity outlined in [37] using MobileNet [45].

III. METHODOLOGY

This section details the methodology of the proposed JSQAT framework and the associated experimental configurations. The proposed framework consists of three major stages, starting with transfer learning and fine-tuning [46] on histopathological images of breast cancer labeled as malignant or benign. The second stage is sparsity-aware training, and the final stage is quantization-aware training as shown in Figure 1. In this first stage of transfer learning, we started with two ImageNet [39] VGG19 [31] and ResNet-50 [44] pre-trained networks. Each network is appended by a global average pooling layer, a fully connected layer, and a classification (softmax) layer. We set the frozen/unfrozen layer ratio to around 70%/30% of total parameters. This results in most of the layers being frozen which preserves the pre-trained weights as shown in Figure 2, which demonstrates the state of the model layers at each of the three stages.

A. SPARSITY-AWARE TRAINING

Network pruning is a well-researched approach to model compression [47], with two main methods; unstructured and structured pruning, our approach to pruning falls under the former category. We incorporate network pruning into our training to reduce model complexity and overcome overfitting, with the goal of reducing the storage required in order to suit the capabilities of resource-constrained devices.

To maximize the efficacy of the pruning step, we adopt L_1 regularization as a preliminary step before pruning by integrating it into the loss function of the fine-tuning stage. L_1 regularization-based pruning is used for its effectiveness in minimizing weights and encouraging sparsity in the training layers thus facilitating the subsequent pruning step [48], [49]. When added to a training loss function, the L_1 regularization term imposes a penalty that encourages more weights to become zero or near zero by penalizing the sum of absolute weights, this allows the model to learn a sparse representation. The regularization term consists of a penalty hyperparameter (λ) that can be tuned to control the severity of the penalization.

We also experiment with another regularization technique, L_1L_2. While L_1 regularization encourages sparsity in weights, L_2 regularization [50], [51] is primarily for controlling model complexity by distributing weights evenly throughout the model. The implication of using L_2 regularization is

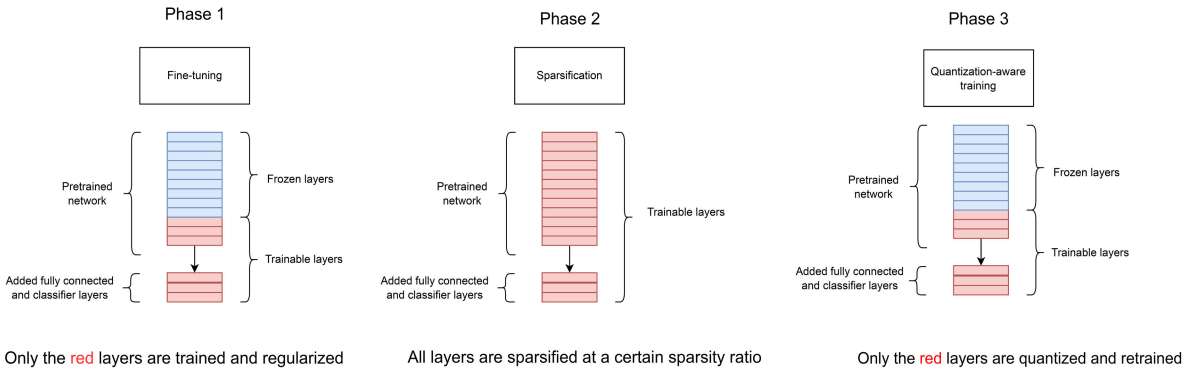

FIGURE 2. Changes to the network at each stage, where the layers colored red are trainable.

preventing overfitting and stark differences between weights by promoting weight decay. Typically it is used in conjunction with L_1 regularization for a more pronounced effect when it comes to sparsification and regularization-based pruning to improve the overall generalization ability of the model.

During the pruning step, we use unstructured magnitude-based pruning [52] in which the smallest weights are removed until the sparsity constraint is satisfied. This is followed by re-training to allow the network to recover from potential accuracy loss. The former steps are repeated iteratively, this is referred to as iterative pruning [53]. It is important to point out that the magnitude-based pruning schedule that is followed enforces the sparsity percentage layer-wise not globally to the entire network. We chose unstructured magnitude-based pruning as our method due to its flexibility and more effective compression without compromising accuracy [54].

Equations (1) and (2) show the loss function used. This includes the Cross-Entropy loss term (L_{CE}) and regularization penalty terms to encourage sparsity in the training layers, y represents the ground truth and $\hat{y}$ is the predicted output while N refers to the number of sample points.

$$L(\hat{y}, y) = L_{CE} + \lambda_1 \sum_{i=1}^{N} |w_i| + \lambda_2 \sum_{i=1}^{N} w_i^2 \tag{1}$$

$$L_{CE} = -\sum_{i}^{N} y_i log(\hat{y}_i) \tag{2}$$

The second and third terms of the loss function (Equation (1)) are the regularization terms which consist of the λ_1 and λ_2 hyperparameters and the L_1-norm and the L_2-norm respectively. The λ hyperparameter affects the severity of the regularization penalty. The regularization terms penalize the sum of absolute weights and squared weights. This incentivizes the loss function to minimize the magnitude of the weights [55], [56]. In our experiments, we report the results for the following selection of hyperparameters $< \lambda_1, \lambda_2 >=< \lambda_0, 0 >$. Going forward, this is referred to as L_1 regularization mode and $< \lambda_1, \lambda_2 >=< \lambda_0, \lambda_0 >$ is referred to as L_1L_2 regularization mode. The λ_0 values tested are $\{0, 0.001, 0.005, 0.01, 0.05\}$ for the first network and $\{0, 0.001, 0.005, 0.01\}$ for the second network.

B. QUANTIZATION-AWARE TRAINING

In order to obtain a more compact network, we use quantization to reduce the number of bits required for each parameter from 32-bit floating point to 8-bit integers after the sparsification methods previously mentioned. In addition to minimizing the required memory storage, weight quantization lessens the burden of computation by enabling fixed-point arithmetic. For the purposes of our application, we opt for quantization-aware training (QAT), which is able to achieve competitive performance when compared to post-training quantization (PTQ). This is due to accounting for quantization error resulting from 8-bit quantization in the training process [57], which is not considered in PTQ. As shown in Figure 1, this entails quantization and dequantization blocks in the forward propagation path of the training. No quantization takes place in the backpropagation. Quantization in the forward propagation occurs according to Equation (3):

$$w_q = min(q_{max}, max(q_{min}, \frac{w}{s} + z)) \tag{3}$$

where w is the input tensor/weight, w_q is the quantized weight, q_{max} and q_{min} are the maximum and minimum values for the desired bit quantization, s is the scaling factor and z is the zero-point. Then, dequantization is achieved by $\hat{w} = (w_q - z)s$. These operations are carried out by the quantizer and dequantizer nodes respectively. However, observing the dequantization equation, the recovered weight $\hat{w}$ is not exactly equal to the original input weight w. As a result, quantization is a process that triggers loss which the optimizer tries to minimize by adjusting the network's weights. The loss due to quantization for a network with w inputs, y labels, and θ parameters can be described as $L[f(w, q(\theta)), y]$ where q is the quantization operation. Adding this new loss to the loss function in Equation (1), then the overall objective function to be minimized from all prior loss functions becomes:

$$L_{CE} + \lambda_1 \sum_{i=1}^{N} |w_i| + \lambda_2 \sum_{i=1}^{N} w_i^2 + L[f(w, q(\theta)), y] \tag{4}$$

The steps in the framework are described in Algorithm 1, where the model undergoes sparsification and then quantization in the training pipeline.

Algorithm 1 Algorithm for Sparse-Quantization Aware Training

Input: Pre-trained network with parameters θ, total number of layers L, number of trainable layers k, number of iterations n, pruning mask m and pruning rate p

Output: Lightweight classifier for histopathological images

Stage I : Transfer Learning (Fine-tuning)

for one epoch **do**
  Minimize the loss function L_{CE}
end for
Freeze the first $L-k$ layers of the pre-trained model
for each epoch **do**
  Minimize the loss function $L_{CE} + \lambda_1 \sum_{i=1}^{N} |w_i| + \lambda_2 \sum_{i=1}^{N} w_i^2$
end for

Stage II : Sparse-Aware Training

Freeze the first $L-k$ layers of the pre-trained model
for each epoch **do**
  Prune the smallest absolute weights $p^{\frac{1}{n}}\%$ layer-wise. The resulting mask m is applied to the network parameters which become $\rightarrow \theta \odot m$
  Re-train the network
  Repeat for n iterations
end for

Stage III: Quantization-Aware Training

for each epoch **do**
  Train in Simulated Quantization
  Minimize the objective function $L_{CE} + \lambda_1 \sum_{i=1}^{N} |w_i| + \lambda_2 \sum_{i=1}^{N} w_i^2 + L[f(w, q(\theta)), y]$
end for
Quantize the k layers to 8-bits

C. PRACTICAL CONSIDERATIONS

When performing transfer learning we opted for the fine-tuning approach where the network's upper layers plus the classifier layers are trainable, and the lower layers are frozen as demonstrated by the starting configuration in Figure 2. Results reported in the following section are of VGG19 and ResNet-50 pre-trained networks. As previously mentioned, only about 30% of the parameters are trainable, which reduces the number of total parameters of VGG19 from 20,576,466 to 7,631,506 trainable parameters. Similarly, reducing ResNet-50 from 25,786,386 total parameters to 7,720,082 trainable parameters.

Experiments were conducted to test different configurations of the loss function presented in Equation (4). Namely, the following modes of the framework were tested: Quantization-only Aware Training (QAT), Sparsity-only Aware Training (SAT), and Joint Sparsity-Quantization Aware Training (JSQAT). We experimented with enforcing the sparsity using the two modes introduced in Section III. These are namely L_1 regularization mode, and L_1L_2 regularization mode.

We test the effect of pruning at 50%, 70%, and 90% sparsity on performance. We empirically experiment with different λ values of the regularization term to assess its effect Joint Sparsity-Quantization Aware Training schedule. The training schedule of the framework follows the prune first then quantize streamline as this is the recommended course for classification tasks [37]. Quantization is applied to the trainable layers only, in order to minimize training time. For both networks tested (VGG19 and ResNet-50), the network was trained for 50 epochs at each stage.

D. DATASET

The dataset used is the Breast Cancer Histopathological Database (BreakHis) dataset [29], which is the largest dataset for histopathological breast cancer images. Even though it is the most comprehensive medical dataset for this condition with a total of 7909 images, it dwarfs in comparison to conventional image datasets like ImageNet. Originally, the dataset has eight classes divided into malignant and benign tumors as shown in Figure 3. We divide the dataset into two classes by merging the malignant tumor samples into one class and merging the benign classes at 40X magnification. The dataset is highly imbalanced with the number of malignant samples being twice the number of benign samples; therefore, we use class weights in training to balance any bias in the results. At 40X magnification, there are 1,370 malignant samples and 625 benign samples which is a total of 1,995 samples. The breakdown of the entire dataset is as follows:

TABLE 1. Statistics of the BreakHis dataset.

Category	Magnification				Total
	40X	100X	200X	400X	
Benign	625	644	623	588	2480
Malignant	1370	1437	1390	1232	5429
Total	**1995**	2081	2013	1820	7909

IV. RESULTS AND DISCUSSION

We present the results for the following networks VGG19 and ResNet-50 when applying the proposed Joint Sparsity-Quantization Aware Training framework. We include the accuracies of the fine-tuned network at different λ values. The accuracies are averaged over three test runs. Firstly, the results of VGG19 are presented in Table 2, which includes the accuracies at different sparsity levels for different λ followed by the results of ResNet-50 in Table 3. Both networks are fine-tuned on the 40X magnification partition of the BreakHis dataset. The results are divided by the mode of regularization and the framework configuration used. This is outlined for both networks. Moreover, after displaying the test accuracies, the resultant model sizes are included in Table 4. The sizes of all networks are of *.tflite* [58] files with the exception of the baseline sizes of the model (underlined) which are *.h5* [59] files in MegaBytes.

Results on the 40X binarized Breakhis dataset at 0% and 50% sparsity levels are very close when Joint Sparsity-Quantization is incorporated into the training

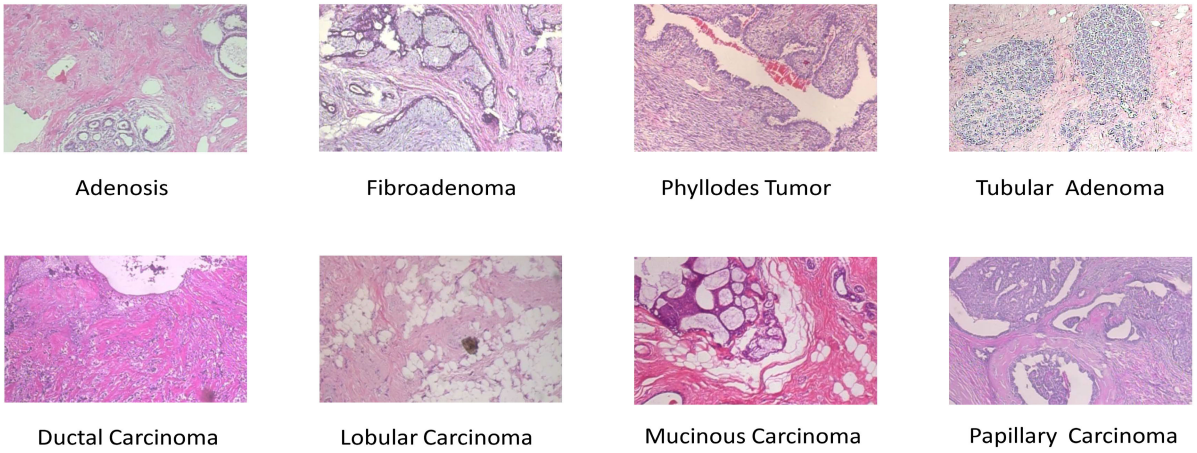

FIGURE 3. Sample images from each class in the BreakHis dataset. The first row consists of benign tumors, the second row is of malignant tumors. The images are 40X magnification.

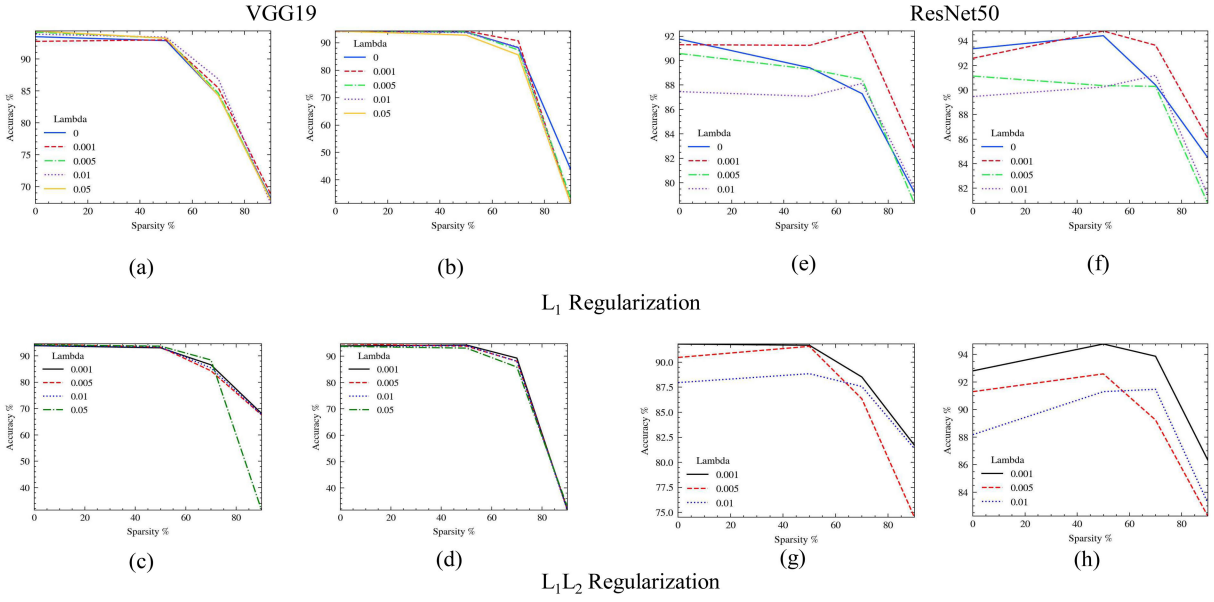

FIGURE 4. Accuracy vs. sparsity for the two configurations, Sparsity-only Aware Training (SAT) *(a,e,c,g)* and Joint Sparsity-Quantization Aware Training framework (JSQAT) *(b,f,d,h)*, with either L_1 and L_1L_2 regularization modes applied for both networks, VGG19 *(a,b,c,d)* and ResNet-50 *(e,f,g,h)*.

schedule, with L_1 and L_1L_2 regularization modes on VGG19 and ResNet-50. The accuracy at 50% sparsity exceeds 0% sparsity in some cases. At 50% sparsity in Joint Sparsity-Quantization Aware Training, there is the advantage of decreased storage requirement which makes it the configuration that achieves better accuracy/size trade-off. This can be due to the regulatory effect on the network from the sparsification process, in addition to the hypothesis [53] that introducing sparsity to networks allows for better generalization which can translate to better performance. Considering the trade-off between sparsity and accuracy, Hoefler et al. [56] observe that moderate sparsity targets (defined as lower than 90%) benefit from magnitude-based pruning, especially iterative pruning, in contrast to high sparsity targets, which is the effect we have observed in our experiments where the sparsity and accuracy trade-off are best at 50% and 70% sparsity as shown in Figure 4.

VGG19 performs best at $\lambda = 0.001$ for both types of regularizations, with an accuracy of 94.26% at 50% sparsity. This combination results in a model size of 46.6 MB, which is a compression of 5× when compared to the original model size. We observe a similar occurrence with regards to ResNet-50 when applying Joint Sparsity-Quantization Aware Training at 50% or 70% sparsity alongside L_1 or L_1L_2 regularization mode. In most cases, the accuracy here exceeds the baseline with more than a 2× decrease in size when compared to the baseline network at 0% sparsity. Overall, ResNet-50 demonstrates better relative performance

TABLE 2. Binary classification accuracy results for applying L_1 and L_1L_2 regularization modes at different sparsities (0%, 50%, 70%, 90%) using a VGG19 backbone. Baseline accuracies are underlined.

VGG19 Sparsity-only Aware Training (SAT)					VGG19 Joint Sparsity-Quantization Aware Training (JSQAT)			
L_1 Regularization								
Lambda (λ)	0%	50%	70%	90%	0% (QAT)	50%	70%	90%
0	93.48%	92.86%	84.28%	68.17%	94.20%	93.92%	88.18%	43.92%
0.001	92.75%	92.98%	85.45%	68.84%	94.09%	**94.26%**	90.69%	31.48%
0.005	94.20%	93.25%	84.62%	67.95%	94.20%	93.98%	87.35%	33.45%
0.01	93.87%	93.42%	86.84%	67.39%	94.20%	93.68%	88.41%	31.48%
0.05	94.37%	93.25%	84.28%	67.89%	94.26%	92.75%	85.56%	31.59%
L_1L_2 Regularization								
0.001	93.87%	93.03%	86.57%	68.17%	93.92%	**94.26%**	89.30%	31.77%
0.005	94.54%	93.31%	84.28%	67.61%	94.70%	93.81%	88.13%	32.05%
0.01	93.82%	93.36%	85.40%	67.73%	93.65%	94.03%	88.07%	31.61%
0.05	93.92%	93.14%	83.67%	68.50%	93.76%	93.09%	85.90%	32.95%

TABLE 3. Binary classification accuracy results for applying L_1 and L_1L_2 regularization modes at different sparsities (0%, 50%, 70%, 90%) using a ResNet-50 backbone. Baseline accuracies are underlined.

ResNet-50 Sparsity-only Aware Training (SAT)					ResNet-50 Joint Sparsity-Quantization Aware Training (JSQAT)			
L_1 Regularization								
Lambda (λ)	0%	50%	70%	90%	0% (QAT)	50%	70%	90%
0	91.75%	89.41%	87.29%	79.21%	93.37%	94.43%	90.41%	84.50%
0.001	91.30%	91.25%	92.42%	82.78%	92.59%	**94.81%**	93.65%	86.07%
0.005	90.58%	89.30%	88.46%	78.32%	91.14%	90.36%	90.30%	80.77%
0.01	87.46%	87.08%	88.13%	79.54%	89.47%	90.25%	91.19%	81.44%
L_1L_2 Regularization								
0.001	91.81%	91.69%	88.52%	81.72%	92.81%	**94.76%**	93.87%	86.34%
0.005	90.47%	91.58%	86.34%	74.53%	91.30%	92.59%	89.24%	82.27%
0.01	87.96%	88.85%	87.57%	81.38%	88.18%	91.30%	91.47%	83.22%

TABLE 4. Model sizes for different pruning percentages for the proposed Joint Sparsity-Quantization Aware Training framework (JSQAT) and the Sparsity-only Aware Training (SAT) configuration in MegaBytes. Baseline sizes are underlined.

	VGG19 Sizes (MB)				ResNet-50 Sizes (MB)			
	No Pruning	Pruning 50%	Pruning 70%	Pruning 90%	No Pruning	Pruning 50%	Pruning 70%	Pruning 90%
SAT	235	57.4	34.5	11.6	157	70.8	42.6	14.4
JSQAT	56.6	46.6	28.0	9.5	79.1	62.4	39.2	16.0

TABLE 5. Comparison with related work on the BreakHis dataset.

Reference	Methods	Backbone network	Accuracy (%)	Resource reduction
[28]	Model is quantized using post-training quantization	MobileNet	87.3	3.3× reduction in size
[30]	Structured pruning with transfer learning	VGG19, ResNet34, ResNet50	90.08	47.54% parameter reduction and 63.46% FLOPs reduction
[32]	Hybrid model with a structured pruning module	Inception and ResNet	85.7±1.9	50% of channels are pruned
Ours	Joint Sparsity-Quantization Aware Training	VGG19	**94.26**	5× reduction in size at 50% sparsity
Ours	Joint Sparsity-Quantization Aware Training	ResNet-50	**94.81**	2.5× reduction in size at 50% sparsity

at higher sparsity percentages when compared to VGG19. On the other hand, VGG19 displays higher compression ratios. ResNet-50's performance is best at $\lambda = 0.001$ using L_1 regularization mode, where 50% sparsity gives an accuracy of 94.81% and size reduction of 2.5×. Notably, 70% and 90% sparsity at the same λ give an accuracy of 93.87% and 86.34% respectively. In addition to achieving a size decrease of 4× and 9.8× for 70% and 90% sparsities when the Joint Sparsity-Quantization Aware Training framework is used.

To summarize our results, it is worth pointing to the enhanced accuracy at 50% sparsity when using the Joint Sparsity-Quantization Aware Training framework configuration for both VGG19 and ResNet-50 at most λ values used. With the addition of a 2.5× and 5× decrease in size for ResNet-50 and VGG19 whilst using only 30% of the total parameters. We note that VGG19 displays a reduced memory footprint with higher compression ratios but slightly lower accuracy than ResNet-50. In low-resource settings, using VGG19 as a backbone would be a better practical choice

as it provides better compression ratios than ResNet-50 and competitive performance accuracy. Given that VGG19 yields almost double the compression ratio for only a 0.55% decrease in accuracy. Thus it proves to be a better choice when it comes to practical resource considerations in resource-constrained environments.

To provide a point of comparison, we list our results with those of other lightweight models trained on the BreakHis dataset as shown in Table 5. We compare our best performing model against works that align with resource awareness, focusing specifically on pruning or quantization as compression methods. We limit the works to those that test on the BreakHis dataset in order to provide commonality when comparing since providing a fair and unbiased comparison is challenging due to the lack of standardization when reporting performance results on medical data. We note a 1.7 × more reduction in size and an increase of 6.96% in accuracy when comparing JSQAT (VGG19) to MobiHisNet [28]. One of the resource gains reported in [30] is a 47.54% reduction in model parameters compared to our 50% model sparsity ratio, where JSQAT (VGG19) yielded a 4.18% increase in classification accuracy. Additionally, our unstructured sparsity approach at 50% sparsity resulted in an increase of 9.77% accuracy compared to the unstructured pruning and hybrid model amalgamation followed in [32].

The results are demonstrative that it is possible to achieve a fair compromise between accuracy and size to meet the criteria of both resource-constrained environments and medical image classification needs with the proposed Joint Sparsity-Quantization Aware Training framework.

V. CONCLUSION

Clinical integration of computer-aided diagnostic devices depends on a multitude of factors. In order to decrease the dependency on human observers for a faster and more streamlined process of early breast cancer detection, the subject of deep learning in computer-aided diagnosis has become a heavily investigated research area. Deep learning in the medical domain has enjoyed much success, however, there remain difficulties that hinder its progress, namely massive resource and data consumption.

To navigate this issue, we explore the use of transfer learning and fine-tuning on the BreakHis dataset to counteract the effect of limited data samples. To meet this end of classifying histopathological tumors, we develop a Joint Sparsity-Quantization Aware Training framework that integrates model compression techniques such as quantization-aware training, regularization, and magnitude-based pruning for the benefit of balancing accuracy and the memory footprint. We empirically investigate the effectiveness of the approach on different pre-trained networks to assess the resilience of different networks against the introduction of reduced precision and removal of weights.

For future evaluation in extremely data-limited settings, few-shot learning has emerged in recent times as a promising method to address limited size classes, with significant implications pertaining to generalization when paired with compression, specifically sparsifying techniques, especially in medical applications, we leave this as future work.

Accordingly, focusing on learning with limited resources and data is essential for medical applications as overcoming these two constraints allows for designing better cost-effective applications to be used in low-resource clinical settings.

REFERENCES

[1] Z. Zuo, P. Wang, X. Chen, L. Tian, H. Ge, and D. Qian, "SWnet: A deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures," *BMC Bioinf.*, vol. 22, no. 1, p. 434, Sep. 2021, doi: 10.1186/s12859-021-04352-9.

[2] R. Ranjbarzadeh, A. B. Kasgari, S. J. Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Sci. Rep.*, vol. 11, no. 1, p. 1, May 2021, doi: 10.1038/s41598-021-90428-8.

[3] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Med.*, vol. 13, no. 1, p. 152, Sep. 2021, doi: 10.1186/s13073-021-00968-x.

[4] Q. Yang, X. Li, X. Ding, F. Xu, and Z. Ling, "Deep learning-based speech analysis for Alzheimer's disease detection: A literature review," *Alzheimer's Res. Therapy*, vol. 14, no. 1, p. 186, Dec. 2022, doi: 10.1186/s13195-022-01131-3.

[5] L.-J. Tseng, A. Matsuyama, and V. MacDonald-Dickinson, "Histology: The gold standard for diagnosis?" *Can. Veterinary J.*, vol. 64, no. 4, p. 389, Apr. 2023.

[6] C. van Dooijeweert, P. J. van Diest, S. M. Willems, C. C. H. J. Kuijpers, E. van der Wall, L. I. H. Overbeek, and I. A. G. Deckers, "Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in The Netherlands," *Int. J. Cancer*, vol. 146, no. 3, pp. 769–780, Feb. 2020, doi: 10.1002/ijc.32330.

[7] C. S. Vikranth, B. Jagadeesh, K. Rakesh, D. Mohammad, and S. Krishna, "Computer assisted diagnosis of breast cancer using histopathology images and convolutional neural networks," in *Proc. 2nd Int. Conf. Artif. Intell. Signal Process. (AISP)*, Feb. 2022, pp. 1–6, doi: 10.1109/AISP53593.2022.9760669.

[8] Z. Guo, J. Xie, Y. Wan, M. Zhang, L. Qiao, J. Yu, S. Chen, B. Li, and Y. Yao, "A review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis," *Open Life Sci.*, vol. 17, no. 1, pp. 1600–1611, Dec. 2022, doi: 10.1515/biol-2022-0517.

[9] H. Kim, D. Kim, and K. Yoon, "Medical big data is not yet available: Why we need realism rather than exaggeration," *Endocrinol. Metabolism*, vol. 34, no. 4, pp. 349–354, Dec. 2019, doi: 10.3803/EnM.2019.34.4.349.

[10] W.-H. Weng, J. Deaton, V. Natarajan, G. F. Elsayed, and Y. Liu, "Addressing the real-world class imbalance problem in dermatology," 2020, *arXiv:2010.04308*.

[11] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. A. Ioannidis, G. S. Collins, and M. Maruthappu, "Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 2020, p. m689, Mar. 2020, doi: 10.1136/bmj.m689.

[12] H. A. Haenssle et al., "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[13] T. J. Brinker, A. Hekler, A. H. Enk, C. Berking, S. Haferkamp, A. Hauschild, M. Weichenthal, J. Klode, D. Schadendorf, T. Holland-Letz, C. von Kalle, S. Fröhling, B. Schilling, and J. S. Utikal, "Deep neural networks are superior to dermatologists in melanoma image classification," *Eur. J. Cancer*, vol. 119, pp. 11–17, Sep. 2019, doi: 10.1016/j.ejca.2019.05.023.

[14] A. R. Luca, T. F. Ursuleanu, L. Gheorghe, R. Grigorovici, S. Iancu, M. Hlusneac, and A. Grigorovici, "Impact of quality, type and volume of data used by deep learning models in the analysis of medical images," *Informat. Med. Unlocked*, vol. 29, 2022, Art. no. 100911, doi: 10.1016/j.imu.2022.100911.

[15] R. Fisher, L. Pusztai, and C. Swanton, "Cancer heterogeneity: Implications for targeted therapeutics," *Brit. J. Cancer*, vol. 108, no. 3, pp. 479–485, Feb. 2013, doi: 10.1038/bjc.2012.581.

[16] T. Dhar, N. Dey, S. Borra, and R. S. Sherratt, "Challenges of deep learning in medical image analysisimproving explainability and trust," *IEEE Trans. Technol. Soc.*, vol. 4, no. 1, pp. 68–75, Mar. 2023.

[17] C. H. Yoon, R. Torrance, and N. Scheinerman, "Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned?" *J. Med. Ethics*, vol. 48, no. 9, pp. 581–585, Sep. 2022, doi: 10.1136/medethics-2020-107102.

[18] Z. Jia, J. Chen, X. Xu, J. Kheir, J. Hu, H. Xiao, S. Peng, X. S. Hu, D. Chen, and Y. Shi, "The importance of resource awareness in artificial intelligence for healthcare," *Nature Mach. Intell.*, vol. 5, no. 7, pp. 687–698, Jun. 2023, doi: 10.1038/s42256-023-00670-0.

[19] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, "Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning," *Comput. Methods Programs Biomed.*, vol. 223, Aug. 2022, Art. no. 106951, doi: 10.1016/j.cmpb.2022.106951.

[20] C. Wang, W. Gong, J. Cheng, and Y. Qian, "DBLCNN: Dependency-based lightweight convolutional neural network for multi-classification of breast histopathology images," *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103451, doi: 10.1016/j.bspc.2021.103451.

[21] Q. A. Al-Haija and A. Adebanjo, "Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network," in *Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS)*, Vancouver, BC, Canada, Sep. 2020, pp. 1–7, doi: 10.1109/IEMTRONICS51293.2020.9216455.

[22] M. F. I. Soumik, A. Z. B. Aziz, and M. A. Hossain, "Improved transfer learning based deep learning model for breast cancer histopathological image classification," in *Proc. Int. Conf. Autom., Control Mechatronics Ind. (ACMI)*, Rajshahi, Bangladesh, Jul. 2021, pp. 1–4, doi: 10.1109/ACMI53878.2021.9528263.

[23] Y. Benhammou, B. Achchab, F. Herrera, and S. Tabik, "BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights," *Neurocomputing*, vol. 375, pp. 9–24, Jan. 2020, doi: 10.1016/j.neucom.2019.09.044.

[24] S. H. Gheshlaghi, C. N. E. Kan, and D. H. Ye, "Breast cancer histopathological image classification with adversarial image synthesis," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 3387–3390, doi: 10.1109/EMBC46164.2021.9630678.

[25] M. M. K. Sarker, F. Akram, M. Alsharid, V. K. Singh, R. Yasrab, and E. Elyan, "Efficient breast cancer classification network with dual squeeze and excitation in histopathological images," *Diagnostics*, vol. 13, no. 1, p. 103, Dec. 2022, doi: 10.3390/diagnostics13010103.

[26] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," 2021, *arXiv:2104.00298*.

[27] R. O. Ogundokun, S. Misra, A. O. Akinrotimi, and H. Ogul, "MobileNet-SVM: A lightweight deep transfer learning model to diagnose BCH scans for IoMT-based imaging sensors," *Sensors*, vol. 23, no. 2, p. 656, Jan. 2023, doi: 10.3390/s23020656.

[28] A. Kumar, A. Sharma, V. Bharti, A. K. Singh, S. K. Singh, and S. Saxena, "MobiHisNet: A lightweight CNN in mobile edge computing for histopathological image classification," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17778–17789, Dec. 2021, doi: 10.1109/JIOT.2021.3119520.

[29] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.

[30] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A transfer learning with structured filter pruning approach for improved breast cancer classification on point-of-care devices," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104432, doi: 10.1016/j.compbiomed.2021.104432.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[32] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact CNNs," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 198, Oct. 2019, doi: 10.1186/s12911-019-0913-x.

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[34] Y. Zhou, J. Sonneck, S. Banerjee, S. Dörr, A. Grüneboom, K. Lorenz, and J. Chen, "EfficientBioAI: Making bioimaging AI models efficient in energy, latency and representation," 2023, *arXiv:2306.06152*.

[35] S. Ye, T. Zhang, K. Zhang, J. Li, J. Xie, Y. Liang, S. Liu, X. Lin, and Y. Wang, "A unified framework of DNN weight pruning and weight clustering/quantization using ADMM," 2018, *arXiv:1811.01907*.

[36] P.-H. Yu, S.-S. Wu, J. P. Klopp, L.-G. Chen, and S.-Y. Chien, "Joint pruning & quantization for extremely sparse neural networks," 2020, *arXiv:2010.01892*.

[37] X. Zhang, I. Colbert, K. Kreutz-Delgado, and S. Das, "Training deep neural networks with joint quantization and pruning of weights and activations," 2021, *arXiv:2110.08271*.

[38] Y. LeCun, C. Cortez, and C. C. J. Burges. *The MNIST Handwritten Digit Database*. Accessed: Jan. 26, 2023. [Online]. Available: https://yann.lecun.com

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[40] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015, doi: 10.1007/s11263-015-0816-y.

[42] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[45] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[46] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.

[47] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 598–605.

[48] M. D. Collins and P. Kohli, "Memory bounded deep convolutional networks," 2014, *arXiv:1412.1442*.

[49] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," 2017, *arXiv:1708.06519*.

[50] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," 2015, *arXiv:1506.02626*.

[51] H. Wang, C. Qin, Y. Zhang, and Y. Fu, "Neural pruning via growing regularization," 2020, *arXiv:2012.09243*.

[52] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.

[53] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," 2018, *arXiv:1803.03635*.

[54] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," 2016, *arXiv:1611.06440*.

[55] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," 2017, *arXiv:1710.09282*.

[56] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, "Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks," 2021, *arXiv:2102.00554*.

[57] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," 2017, *arXiv:1712.05877*.

[58] *TensorFlow Lite*. Accessed: Jan. 26, 2023. [Online]. Available: https://www.tensorflow.org/lite/guide

[59] *Save, Serialize, and Export Models*. Accessed: Jan. 26, 2023. [Online]. Available: https://www.tensorflow.org/guide/keras/serialization_and_saving

DINA ABOUTAHOUN (Member, IEEE) received the B.Sc. degree in computer and communications engineering from Alexandria University. She is currently pursuing the M.Sc. degree with the Egypt–Japan University of Science and Technology. She is also a Teaching Assistant with the Egypt–Japan University of Science and Technology. Her research interest includes artificial intelligence in medicine.

RAMI ZEWAIL received the B.Sc. and M.Sc. degrees from the Arab Academy for Science and Technology, Egypt, and the Ph.D. degree from the University of Alberta, Canada. He has over 15 years of academic and industrial research and development experience in the areas of computer vision, machine learning, and embedded intelligence. He is an Assistant Professor with the Department of Computer Science and Engineering, Egypt–Japan University of Science and Technology, Alexandria, Egypt. His research experience spans different fields, such as healthcare, the oil and gas industry, and biometrics authentication. His research interests include computer vision, machine learning, and embedded intelligence, with a focus on learning in resource-constrained settings. He is a member of the Canadian Association for Artificial Intelligence. He serves as a reviewer for a number of scientific journals.

KEIJI KIMURA received the Ph.D. degree in electrical engineering from Waseda University, in 2001. He was an Assistant Professor (2004), an Associate Professor (2005), and a Professor (2012) with Waseda University, where he has been the Director of the Green Computing System Research Organization, since 2019. His research interests include multi-core processor architecture and parallelizing compiler technologies. He is a member of IPSJ and ACM. He was a recipient of the 2014 Ministry of Education, Culture, Sports, Science, and Technology in Japan (MEXT) Award. He has served on program committee for many conferences.

MOSTAFA I. SOLIMAN received the B.Sc. and M.Sc. degrees in computer science and engineering from the University of Assiut, Egypt, in 1994 and 1998, respectively, and the Ph.D. degree in computer science and engineering from the University of Aizu, Japan, in 2004. He is currently a Full Professor with Aswan University, Egypt. He is also the General Director of the Computer Science and Information Technology (CSIT) Programs, Egypt–Japan University of Science and Technology. His research interests include computer architecture, parallel processing, vector/matrix processing, performance evaluation, parallel algorithms, FPGA, and SystemC implementations.