IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 November 2024, accepted 22 November 2024, date of publication 25 November 2024, date of current version 9 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3506334

RESEARCH ARTICLE

A Lightweight Deep Convolutional Neural Network Implemented on FPGA and Android Devices for Detection of Breast Cancer Using Ultrasound Images

ADITYA VINOD, PRABHAV GUDDATI, AMIT KUMAR PANDA, (Senior Member, IEEE), AND RAJESH KUMAR TRIPATHY, (Senior Member, IEEE)

Department of Electronics and Electrical Engineering, Birla Institute of Technology and Science (BITS) Pilani, Hyderabad 500078, India

Corresponding author: Amit Kumar Panda (amit@hyderabad.bits-pilani.ac.in)

This work was supported by the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India, New Delhi under Start-up Research Grant (SRG) Project under Grant SRG/2021/001979.

ABSTRACT Breast cancer (BC) continues to be the primary cause of high mortality rates among women globally. Early and automated detection of this disease plays a significant role in clinical standards for better diagnosis and improved survival rates. Ultrasonography is a widely used non-invasive imaging test to diagnose BC. The traditional ultrasound-based automated diagnostic systems for detecting BC rely on cloud-based processing, which has high latency, requires constant internet connectivity, and raises the privacy of patient's ultrasound image data. This paper proposes a lightweight deep convolutional neural network (LWDCNN) implemented on edge devices (field programmable gate array (FPGA) and android devices) for real-time detection of BC using breast ultrasound (BUS) images. FPGA is used due to its parallel processing capability, low latency, and low power for the real-time processing of BUS images. Similarly, the Android device provides portability and a user-friendly system to process BUS images for automated detection of BC. The proposed LWDCNN is trained on the Google Cloud CPU-based framework to obtain the optimized model with weight and bias parameters. The compression methods (pruning and fixed-point precision-based representation of weight values) are applied to the LWDCNN model, and the inference of this model is performed on Android and PYNQ-Z2 FPGA-based edge devices to detect malignant tumor classes using BUS images. The proposed approach is evaluated using BUS images from Kaggle and breast ultrasound imaging databases. The experimental results reveal that the LWDCNN achieved average accuracy values of 94.15% on Google Cloud CPU and 93.16% on FPGA-based edge devices for detecting malignant tumors in the inference phase using hold-out validation. Similarly, the post-training quantization of LWDCNN with an integer 8-bit (INT8) case implemented on an Android device has yielded an accuracy value of 93.76% for detecting malignant tumors using BUS images. The presented deep-learning approach has demonstrated higher classification accuracy than different transfer-learning techniques and existing methods using the BUS images from the same database.

INDEX TERMS Ultrasound images, breast cancer, deep learning, edge computing, FPGA, android.

I. INTRODUCTION

Breast cancer (BC) is the most common cancer type among women, and it has a high death rate in India and other countries [1]. It occurs due to the mutation of breast cells and produces cancer cells, which further proliferate to create malignant tumors [2]. The diagnosis of BC is performed using different medical imaging techniques such as mammography, magnetic resonance imaging (MRI), ultrasonography, and positron emission tomography (PET) [2]. Mammography

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

is the most common and widely used procedure based on low-dose X-rays to provide detailed images of breast tissue for diagnosing BC [3]. It has shortcomings, such as prolonged exposure of breast tissue to non-ionizing radiation and limited sensitivity with dense breast tissue in younger women [4]. MRI and PET tests are expensive and require longer examination times to diagnose BC. On the other hand, ultrasonography is a non-invasive imaging procedure that utilizes sound waves to obtain images of breast tissue [5]. It has advantages, such as not requiring exposure to radiation on the breast tissue like mammography, helping distinguish solid masses from fluid-filled cysts, and being very useful for diagnosing breast cancer in dense breast tissue [6]. The manual procedure is time-consuming and requires experienced radiologists to provide accurate diagnostic reports for diagnosing malignant tumor classes using breast ultrasound (BUS) images. Artificial intelligence (AI)-based methods enable continuous monitoring and provide additional diagnostic decisions that help radiologists in diagnosing BC [2]. The advances in edge computing and the Internet of Things (IoT) are helpful for the development of intelligent healthcare frameworks for accurate and speedy diagnosis of various diseases using biomedical signals and images [7]. In such intelligent healthcare systems, the on-device implementation of privacy-preserving AI-based models is performed for the real-time detection of diseases using medical images. The advantages of such systems with respect to cloud computing are that the diagnosis is performed closer to the imaging system, ensuring data privacy and security through local processing, lowering costs associated with data transfer from the imaging device to the cloud, minimizing latency, and optimizing resource utilization [8]. The Android device for edge computing applications has benefits such as being widely available, performing on-device computation, having a user-friendly interface, and being cost-effective. Similarly, the field programmable gate array (FPGA) has shown significant advantages due to parallel processing, low power, low latency, adaptability, and function acceleration compared to other edge devices for healthcare applications [9], [10]. The development of a novel AI-based model and its implementation on resource-constrained edge (RCE) devices is essential for the automated detection of malignant tumors using BUS images.

In literature, various AI-based methods have been proposed for detecting using BUS images [11]. Moon et al. [12] have formulated an ensemble learning-based approach based on the outputs of different transfer learning techniques to categorize benign and malignant tumors using BUS images. They verified that the classification accuracy of the ensemble learning-based approach has been improved as as compared to the individual transfer learning-based methods for detecting BC. In [13], the authors have implemented a 15-layer deep convolutional neural network (CNN)-based model to detect malignant tumor classes using BUS images. The ResNet101-based transfer learning model has been utilized in [14] to classify benign and malignant tumors using BUS images. The accuracy of the ResNet101 model has been improved after the fine-tuning (training all layers) process for detecting malignant tumor classes. Similarly, in another study, the authors have compared the classification accuracy of various deep learning models, such as RetinaNet, faster R-CNN, YoLoV2, cascaded R-CNN, and BUSNet (formulated using Resnet50-based transfer learning models) to automatically categorize benign and malignant tumors by considering input as BUS images [5]. They have found that the average accuracy of BUSNet is 56.60% compared to the other four deep learning-based models. Gade et al. [2] have proposed the multiscale analysis domain deep learning (MSADDL) model for detecting BC using BUS images. They obtained the overall classification accuracy values of 84.96% and 79.88% to categorize benign and malignant tumors using BUS images from the Kaggle and BUSI databases. In another study, authors utilized a pre-trained version of a deep residual network model to extract features from BUS images and used a support vector machine (SVM) model to categorize benign and malignant tumors [15]. They have obtained an area under the receiver operating characteristics curve (AROC) value of 0.683. Similarly, Raza et al. [16] have used a 24-layer-based deep learning model to classify benign and malignant tumors using BUS images. They have obtained higher classification accuracy than various transfer learning models to detect BC using BUS images. Similarly, in another study, Lu et al. [17] utilized the ResNet-50 transfer learning-based pre-trained network as the feature extractor and optimized extreme learning machine (ELM) as a classifier to detect malignant tumors using BUS images. They have obtained an overall accuracy value of 93.97%. Mo et al. [18] have explored the HoVer-Transformer-based deep learning approach to detect malignant tumors using BUS images and reported an accuracy value of 92.40%. They have also shown that the HoVer-Transformer-based model has outperformed various transfer learning techniques, vision transformer, and swin transformer-based methods for detecting BC using BUS images. Balasubramaniam et al. [19] have proposed a modified LeNet-based deep learning model to detect malignant tumors using BUS images and obtained an accuracy value of 89.91%. Rao et al. [20] have formulated an ensemble deep-learning strategy by considering VGG16, VGG19, and InceptionV3-based transfer learning blocks as feature extractors and stacking-based multilayer perceptron models for detecting BC using BUS images. They have obtained an overall accuracy value of 85.80%. In another study, Lanjewar et al. [21] used MobileNetV2-based transfer learning block as a feature extractor and long short-term memory (LSTM) as a classifier to detect malignant tumors using BUS images and reported an accuracy value of 94%.

The existing deep learning and transfer learning methods for the automated detection of malignant tumor classes using BUS images have more than 15 layers [2], [5], [14]. The

number of weight parameters for BUSNet and MSADDl models is 25.6 million and 12, 2038, respectively. The transfer learning techniques have fewer training parameters, but they have more than one million parameters in the inference phase for evaluating performance using unseen data. The inference times to process the BUS images using these techniques are high due to the larger model size and parameters in the inference phase. These models have not been implemented on RCE devices (Android and FPGAs) to detect BC. The real-time implementation of such deep layer models requires higher utilization of resources on FPGA. The existing deep learning models, such as RetinaNet, faster R-CNN, YoLoV2, cascaded R-CNN, BUSNet, and MSADIDL models [2], [5], have obtained classification accuracy values of less than 80% on the public database for detecting malignant tumor classes using BUS images. Hence, an accurate and lightweight deep learning model implemented in real-time using RCE devices is required to detect BC using BUS images. Recently, Guddati et al. [9] have proposed a lightweight CNN model and implemented this model on an FPGA-based RCE device for detecting pneumonia and tuberculosis diseases using chest X-ray images. They have verified that the inference time of their lightweight CNN-based method has been less than various transfer learning techniques to process chest X-ray images. The lightweight deep learning models have not been explored to detect benign and malignant tumors from BUS images. Therefore, the novelty of this work is developing a lightweight deep-learning model and implementing this model on RCE devices (FPGA and an Android device) to categorize benign and malignant tumors using BUS images. The essential contributions of this work are written as follows:

- A lightweight deep CNN (LWDCNN) model with less than 2000 parameters implemented on the cloud-based framework is proposed to detect malignant tumors using BUS images.
- The model is deployed on the Android device for automated categorization of benign and malignant tumors.
- The LWDCNN model is implemented on the PYNQ-Z2-based resource-constrained FPGA device for real-time detection of malignant tumors using BUS images.
- The inference times on the testing phase are compared for cloud, Android device, and FPGA-based implementation of the proposed LWDCNN model.

The remaining sections of this article are organized as follows. In Section II, we have written the details about the BUS image database used in this work. The proposed LWDCNN model architecture and its implementation on FPGA and Android are presented in Section III. We have written the details regarding the results and discussion in Section IV. Finally, the conclusions of this paper are written in Section V.

II. BUS IMAGE DATABASES

In this work, we have utilized two publicly available databases [22], [23] to evaluate the performance of the suggested LWDCNN model to detect malignant tumors using BUS images. The first database (augmented BUS image dataset) mainly consists of training and independent test datasets containing BUS images from benign and malignant tumors [23]. The training dataset contains 8116 BUS images in which 4074 and 4042 belong to benign and malignant tumors, respectively. Similarly, the independent test dataset in the first database (database 1) comprises 900 BUS images. Out of these 900 BUS images, 500 and 400 images are for benign and malignant tumors, respectively. The size of each BUS image in the first database is $224 \times 224 \times 3$. The second database (database 2) comprises 780 BUS images from women ages ranging from 25 to 75 [22]. Out of 780 BUS images, 437, 210, and 133 belong to benign, malignant, and normal classes, respectively. The existing methods in [2] and [5] have considered the BUS images from benign and malignant tumor classes to evaluate the classification performance of their developed deep-learning models. To make a fair comparison with the existing methods, we have used 437 and 210 (in total 647) BUS images from only benign and malignant tumor classes of the second database (database 2) to develop and evaluate the proposed LWDCNN model. The average size of the BUS images in the second database is $500 \times 500 \times 3$. In this work, the BUS images from the first and second databases are resized into $128 \times 128 \times 3$. The resizing of the BUS image helps reduce the size of the LWDCNN model for the implementation on Android and FPGA-based RCE devices.

III. PROPOSED APPROACH

The proposed approach to detect malignant tumors from BUS images mainly comprises two steps. These steps are (a) to develop the LWDCNN model architecture to categorize benign and malignant tumors using resized BUS images and (b) to implement the LWDCNN model on Android and FPGA-based RCE devices. We have described these two steps of the proposed approach in the following subsections.

A. PROPOSED LWDCNN MODEL ARCHITECTURE

The CNN model architecture to detect malignant tumor classes using BUS images is displayed in Fig. 1. It mainly consists of 3 convolution layers, 3 batch normalization (BN) layers, 3 max-pooling layers, and one dense layer. The BN layer is used after each convolution layer, and the total number of parameters for the LWDCNN model is only 1837. The input to the LWDCNN model is the resized BUS image of size as $\mathcal{X} \in R^{128\times128\times3}$. The first convolution layer contains four filters, with the size of each filter as $[3, 3]$, stride as 1, and the 'ReLU' activation function. Similarly, the second and third convolution layers have four and eight filters. The dense layer contains 16 neurons. The i^{th} feature

TABLE 1. Number of BUS images per class in training, validation, and testing from both databases.

	Database 1				Database 2		
Classes	Training	Validation	Test	Independent test set	Training	Validation	Test
Benign	2927	328	819	500	314	35	88
Malignant	2916	321	805	400	150	18	42
Total	5843	649	1624	900	474	53	130

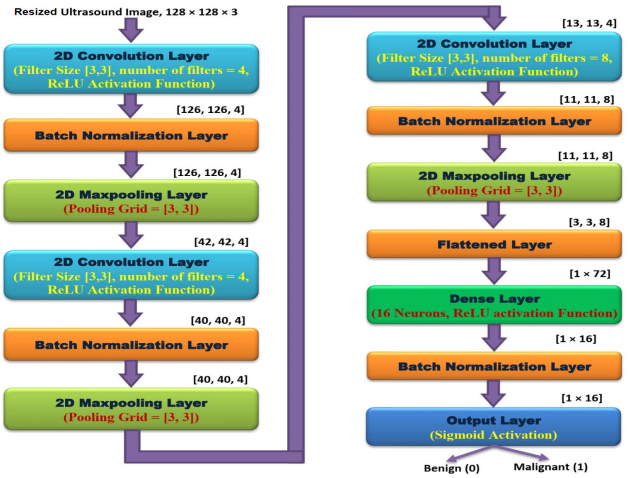

FIGURE 1. LWDCNN model architecture to detect BC using BUS images.

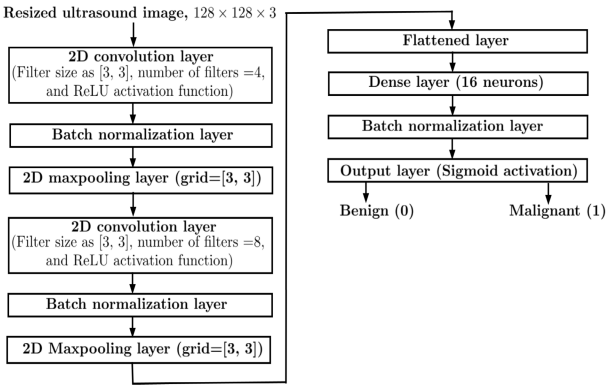

FIGURE 2. Deep CNN architecture 1 to detect BC using BUS images.

map computed in the p^{th} convolution layer is given using the following mathematical expression as [24],

$$\mathcal{X}^{p}_{\tilde{m}1,\tilde{m}2,i} = g\Big[\Big(\sum_{k1=1}^{K1}\sum_{k2=1}^{K2}\sum_{l=1}^{C} H^{i}_{k1,k2,l}\mathcal{X}^{p-1}_{m1+k1-1,m2+k2-1,l}\Big) + b^{i}\Big] \quad (1)$$

where $H^{i}_{k1,k2,l}$ is the kernel for i^{th} feature map and $\mathcal{X}^{p-1}_{(m1+k1-1,m2+k2-1,l)}$ is the tensor used as the input to the p^{th} convolution layer. The factor $g(.)$ is the activation function for the convolution layer. Similarly, the output of the dense layer in the proposed LWDCNN model is evaluated as $\mathbf{z} = f(\mathbf{Wx}+\mathbf{b1})$, where $\mathbf{x}$ is the feature vector evaluated in the flattened layer. The $\mathbf{W}$ and $\mathbf{b1}$ are the weight matrix and bias vector

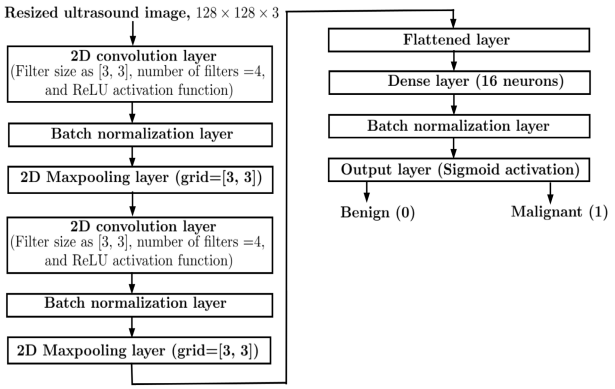

FIGURE 3. Deep CNN architecture 2 to detect BC using BUS images.

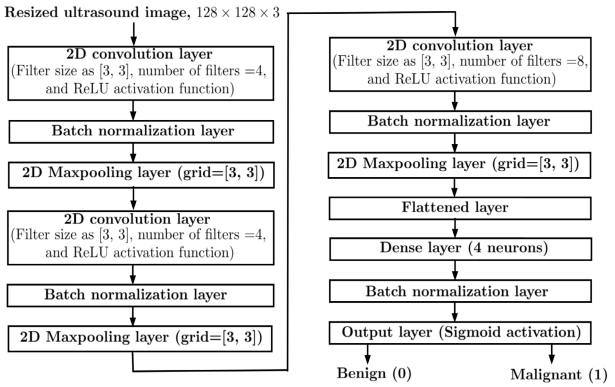

FIGURE 4. Deep CNN architecture 3 to detect BC using BUS images.

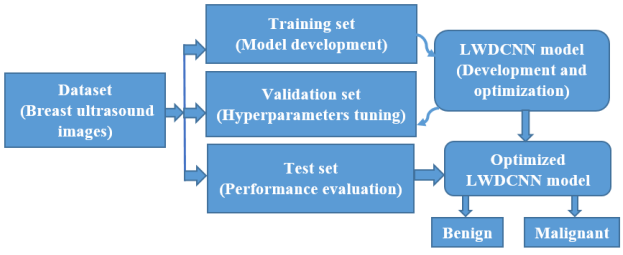

FIGURE 5. Training, validation, and testing procedure of the LWDCNN model to detect malignant tumors using BUS images.

for the dense layer. The factor $f(.)$ is the activation function for the dense layer. In this work, we have experimented with three types of CNN architectures and selected the lightweight model based on the model parameters and accuracy. The deep CNN architecture 1 is depicted in Fig. 2. In architecture 1, we have considered two convolution layers with 4 and

8 filters, two pooling layers with a pooling grid as [3, 3], and one dense layer with 16 neurons. The batch normalization is used after each convolution layer and dense layer. Similarly, the deep CNN architecture 2 considered in this work is shown in Fig. 3. In Architecture 2, we used two convolution layers, with each layer having 4 filters, two pooling layers with a pooling grid [3, 3], and one dense layer with 16 neurons. Likewise, the deep CNN architecture 3 is depicted in Fig. 4. In architecture 3, we have changed the dense layer neuron to 4 compared to the original LWDCNN model, with 16 neurons in the dense layer. All other layers of architecture 3 remain the same as that of the proposed LWDCNN model to detect malignant tumors using BUS images. We have also used transfer learning models such as ResNet50 [25], Densenet201 [26], MobileNetV2 [27], InceptionV3 [28], and XceptionNet [29] to classify benign and malignant tumors using BUS images. For these transfer learning techniques, the pre-trained models (the models trained in the ImageNet dataset for object classification tasks) are used as backbones or feature extractors. The last layer of the pre-trained model has been removed, and the new output layer containing one output neuron (sigmoid activation function) is connected. During the fine-tuning phase, the last layer weight and bias parameters are updated for each transfer learning model.

B. TRAINING PROCEDURE AND HYPERPARAMETERS TUNING

The evaluation protocol of the proposed LWDCNN model is depicted in Fig. 5. The LWDCNN model is trained using the hyperparameters such as total epochs as 30 (early stopping with patience as 10 epochs), batch size as 32, cost function as binary cross-entropy, and initial learning rate as 3×10^{-3}, respectively. These values are selected based on the grid-search method based on the maximization of validation accuracy [24]. The search space for the hyperparameters is given as the learning rate grid $= [10^{-3}, 10^{-2}, 2 \times 10^{-3}, 3 \times 10^{-3}]$, batch size grid $= [32, 64, 128, 256]$, and epochs grid $= [30, 50, 100]$, respectively. In this work, we have utilized two validation approaches such as hold-out validation and 5-fold cross-validation (CV) [24], to select the training and test BUS images for the LWDCNN model to detect malignant tumors. The hold-out validation-based selection of the training, validation, and test sets from database 1 and database 2 is depicted in Table 1. For database 1, 80% of 8116 BUS images (6492 BUS images) are used for the training and validation of the LWDCNN model, and the remaining 20% of BUS images (1624 BUS images) are used in the testing phase of hold-out validation. We have selected 10% of 6492 BUS images (649 BUS images) as the validation set and 5843 BUS images as the training set for the proposed LWDCNN model in the hold-out validation. For database 2, the same procedure is followed for the hold-out validation of the LWDCNN model to detect malignant tumors using BUS images. The independent test dataset (900 BUS images) from database 1 is evaluated separately after the development and training of the LWDCNN model. For the 5-fold CV case, training and test instances are selected from the different portions of the dataset containing 8116 BUS images. The accuracy and other measures are evaluated for each fold, and the average and standard deviation values are shown in the results section for the 5-fold CV of the proposed LWDCNN model. For the second database, 559 BUS images and 138 BUS images are used in the training and testing phases of the LWDCNN model with hold-out validation. In this work, we have selected the same values of hyperparameters (batch size, cost function, number of epochs, and initial learning rate) for each transfer learning model as those of the proposed LWDCNN model to classify benign and malignant tumors using BUS images. The Google Cloud CPU-based framework is used to train and test the proposed LWDCNN model and existing transfer learning techniques for classifying benign and malignant tumors using BUS images.

C. FPGA IMPLEMENTATION OF LWDCNN MODEL

In this work, the PYNQ-Z2 FPGA device is used in the inference or testing phase of the LWDCNN model to detect malignant tumors using BUS images. After the training of the LWDCNN, the RTL, synthesis, and implementation of the trained model are generated using high-level synthesis (HLS) for machine learning (HLS4ML) framework [30], [31]. HLS4ML uses the HLS tools, optimization techniques, resource management, and testing to implement the deep learning model on an FPGA device [31]. The HLS tool helps analyze the deep learning model parameters and generate RTL code using Verilog or VHDL language [30]. The optimization techniques involve Parallelization, Pipelining, loop unrolling, resource sharing, and data reuse to minimize resource usage and latency [30]. Similarly, resource management mainly focuses on effectively implementing the deep learning model on the targeted FPGA device. It involves optimizing resource usage and balancing resource allocation in each layer of the deep learning model within the available resources of the FPGA device. The testing phase is important to know whether the resulting deep learning model performs accurately with low latency for classification or regression tasks on an FPGA device [31]. Before considering the HLS4ML tool, compression has been used to reduce the model size, energy consumption, resource utilization, and overtraining of the deep learning model. We have used pruning-aware training (PAT), which is based on eliminating low-magnitude weights in different layers of the proposed LWDCNN model in the training phase while maintaining or improving the accuracy. The advantages of PAT are efficiency (less memory and computational power due to reduction in model size) and speed (faster inference time) of the deep learning model, which is suitable for edge computing applications [32]. In this study, the PAT is applied for all convolution and dense layers of the LWDCNN model with the final sparsity as 0.50 (50% of the weight values are zero in the weight matrix of each convolution and dense

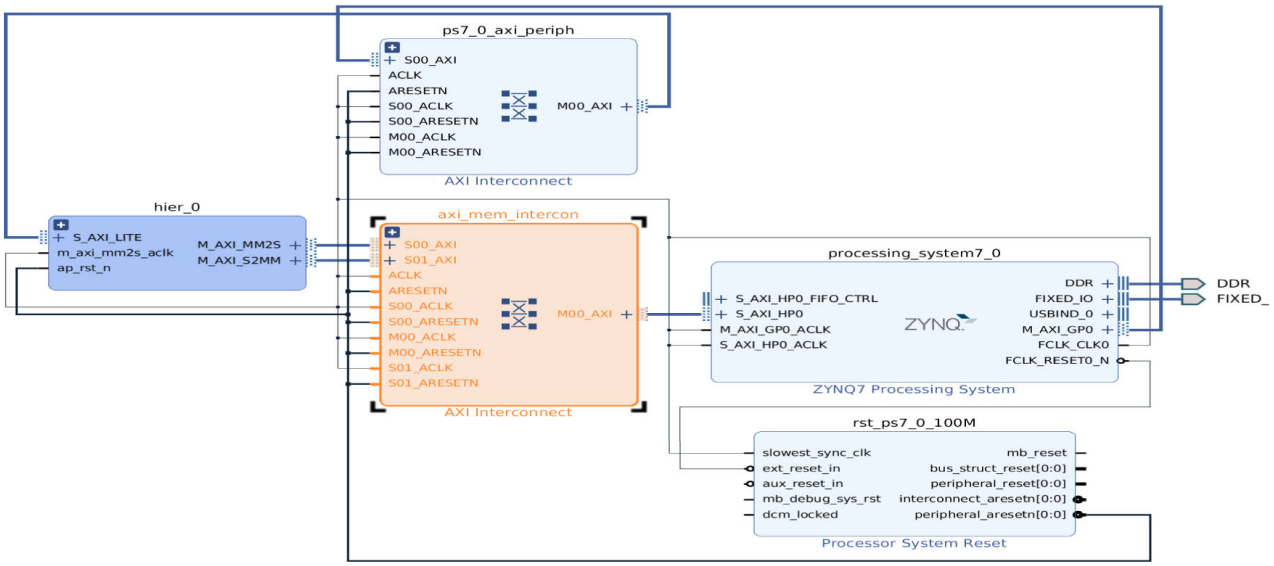

FIGURE 6. Block-diagram representation for implementing the LWDCNN model on PYNQ-Z2 FPGA system to detect malignant tumors using BUS images.

layers). The number of steps to monitor the pruning operation for the LWDCNN model during training varies from $2n_s$ to $10n_s$, where $n_s = \frac{m \times 0.8}{\text{bs}}$ and m is the number of training instances or BUS images. After PAT, the fixed point (FDP) precision-based representation of weight and bias parameters is performed for convolution and dense layers of the proposed LWDCNN model on the HLS4ML framework. The FDP representation of the weight values of each layer of the deep learning model has several advantages, such as less resource utilization on FPGA, low latency due to simpler calculations, consistent precision, and higher throughput than the floating-point (FP) representations [9]. The <20,8>, <18, 8>, <16, 6>, <14,6> and <12, 4>-based FDP precision cases are considered. In <20, 8>-based FDP case, 8 bits are allocated for the integer part, and 12 bits are used to represent the fraction part of the FP number. The optimal FDP precision case for the weight parameters of the LWDCNN model is determined based on the accuracy measure. The direct implementation of the 2D convolution layer, as in equation 1, requires six nested for loops over image width, image height, input channels, number of output filters, height of filter, and width of filter. The loop pipelining concept (parallelizing or unrolling all nested inner loops while executing the outer loop) has been used for implementing a 2D convolution layer [30]. However, the HLS4ML implements a convolution layer using streams. The streams are the first in, first out (FIFO) buffers, which do not require an extra address and consume less storage on FPGA. The parallelization of the calculation in each hidden layer response of the deep learning model includes the tradeoff between latency and resource utilization on FPGA. The reuse factor helps know the number of times the multiplier has been used to evaluate the multiplication of the weight matrix with the input feature matrix or feature vector in each layer of the deep learning model [30]. The full serial multiplication operation results in high latency, low throughput, and low resource utilization. Similarly, a fully parallel operation has low latency and high throughput and thus requires more resources for FPGA. We have selected the reuse factor as 16 for implementing the proposed LWDCNN model on FPGA using HLS4ML. The DSP consumption reduces with an increase in the reuse factor. Similarly, the latency increases with the increase in the reuse factor [30].

The hardware configuration of the PYNQ-Z2 FPGA board is a dual-core ARM-cortex A9 processor, Xilinx Artix-7 (ZYNQ-7000 SOC) FPGA, 512MB DDR3 RAM, and 100MHz clock for the FPGA, respectively. Similarly, software environments such as Vivado, PYNQ image, and Python API are used to implement the LWDCNN model on PYNQ-Z2 FPGA. In this work, we have generated the register-transfer logic (RTL) and post-implementation parts of the proposed LWDCNN model using the HLS4ML tool in Python with VIVADO as the backend [9], [30]. The bitstream and hardware hand-off (HWH) files are generated from VIVADO, and these two files contain the internal logic configuration and address space for implementing the proposed LWDCNN model on the FPGA system. The block diagram containing the VIVADO IP blocks for the design of the proposed LWDCNN model is shown in Fig. 6. It mainly consists of the ZYNQ7 processor system, advanced extensible interface (AXI) memory interconnect, AXI peripheral connection, and processor system reset, respectively [33]. After obtaining the bitstream file containing the design of the suggested LWDCNN model from VIVADO in the post-implementation stage, we have utilized the neural network overlay package to transfer the bitstream file to

the programmable logic (PL) part of the PYNQ-Z2 FPGA framework. Similarly, the test data (BUS images used in the testing phase on FPGA) is transferred from the programmable system (PS) part to the LWDCNN model implementation core part on the PYNQ-Z2 FPGA framework. The predicted outputs or class labels are evaluated on the FPGA core part (LWDCNN model implementation core part) using the test BUS images. The class labels obtained in the testing phase on FPGA are then transferred to memory using AXI memory interconnect [9]. After evaluating the predicted outputs, the classification metrics such as accuracy, precision, recall, F1-score, and kappa score [24] are used to evaluate the performance of the LWDCNN model on hardware to detect malignant tumors using BUS images.

D. ANDROID AND WAPP IMPLEMENTATIONS OF LWDCNN MODEL

In this work, we have also deployed the trained LWDCNN model on Android device and web applications (WAPP) to detect malignant tumors using BUS images in the inference phase. The SAMSUNG Galaxy M33 mobile phone (8GB RAM, 128GB storage, and octa-core processor with 2GHz)-based Android (version 12) device has been used for deploying the quantized version of the LWDCNN model through mobile application (MAPP). The Androidstudio-based framework is used to create a MAPP in which the input to the application is the BUS image, and the output is the predicted class label (benign or malignant). To implement the LWDCNN model on the Android framework, we have used TensorFlow lite to reduce the model size. TensorFlow lite has been designed to implement machine learning models on mobile and embedded devices [34]. It has used different techniques to optimize and compress the machine learning model. It converts the optimized model into TensorFlow lite format, which is used for deployment on embedded devices. The optimization techniques on TensorFlow lite involve quantization and pruning. In this study, We have used only the post-training quantization (PTQ) to obtain the quantized or reduced precision-based LWDCNN model using TensorFlow lite [35]. The quantization is achieved by converting the precision of the model parameters (weights and activation function) from 32-bit FP numbers (FP32) to 8-bit integers (INT8). The quantized LWDCNN model is deployed on the Android device to automatically detect malignant tumors using BUS images. Similarly, the streamlit framework [36] is used to prepare the WAPP for the real-time inference of the proposed LWDCNN model for detecting malignant tumors using BUS images. Streamlit is an open-source, freely available cloud-based framework for deploying machine learning models and creating WAPP for IoT applications [36]. The advantages of considering WAPP for detecting BC are its accessibility at any location with internet connectivity, its usefulness for telemedicine applications, and its platform independence (WAPP can run on computer and Android devices). The unquantized version (FP32) of the LWDCNN model is deployed on the streamlit-based WAPP to detect malignant tumors using BUS images. The classifier parameters are utilized to evaluate the performance of the proposed LWDCNN model inference on Android and WAPP to detect malignant tumors using BUS images.

IV. RESULTS AND DISCUSSION

The proposed LWDCNN model inference results on CPU and FPGA frameworks are presented in section IV-A for different validation strategies to detect malignant tumors using BUS images. Similarly, the resource utilization on FPGA and the interpretability of the proposed LWDCNN model are illustrated in Sections IV-B and IV-C, respectively. The results for the inference of the LWDCNN model on Android devices and WAPP frameworks to detect malignant tumors are presented in Section IV-D. The comparison of the classification performance of the proposed LWDCNN model with transfer learning techniques and other existing models is shown in Section IV-E and IV-F, respectively. The limitations and future directions are written in Section IV-G.

A. LWDCNN RESULTS FOR DIFFERENT VALIDATIONS

In Table 2, we have shown the classification results computed using the proposed LWDCNN model inference on CPU framework to classify benign and malignant tumors using BUS images with different validation strategies (hold-out validation, 5-fold CV, independent test set) for both databases. It is noted that the accuracy of the proposed LWDCNN model is 94.15% when the inference is performed using the CPU to detect BC using BUS images from database 1 for the hold-out validation case. Similarly, the average accuracy value of the LWDCNN model for 5-fold CV is 91.89%. Similar variations are observed for other performance measures such as precision, recall, F1-score, and Kappa score for the inference of the LWDCNN model on CPU framework using hold-out validation and 5-fold CV cases to categorize benign and malignant tumors using BUS images from database 1. In the 5-fold CV, the LWDCNN model is evaluated using different portions of the BUS image dataset. The average values of all fold accuracy and other measures are shown in the table. When the independent test set containing BUS images from database 1 is used, the trained LWDCNN model produced an accuracy value of 86.88%. The accuracy value of the LWDCNN model inference on CPU is obtained as 80.76% using BUS images from database 2 to detect malignant tumor classes.

Moreover, the classification results of the LWDCNN model inference on the PYNQ-Z2-based FPGA framework using the BUS images are displayed in Table 3. The classification accuracy of the LWDCNN model is reduced to 93.16% for the automated categorization of benign and malignant tumors using BUS images using the hold-out validation case on the FPGA device. The average accuracy of the LWDCNN model has dropped from 91.89% to 85.36% when the model inference is performed on the PYNQ-Z2

TABLE 2. Classification performance of LWDCNN inference on CPU framework to detect malignant tumor classes using BUS images with different validation strategies.

Validation method	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	Kappa
Hold-out (Database 1)	94.15	94.15	94.15	94.15	0.883
5-fold CV (Database 1)	91.89 ± 1.82	91.80 ± 1.78	91.80 ± 1.78	91.80 ± 1.78	0.837 ± 0.036
Independent test set (Database 1)	86.88	86.72	87.12	86.81	0.736
Hold-out (Database 2)	80.76	79.31	80.00	79.60	0.592

TABLE 3. Classification performance of LWDCNN inference on PYNQ-Z2 FPGA framework to detect malignant tumor classes using BUS images with different validation strategies.

Validation method	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	Kappa
Hold-out (Database 1)	93.16	93.16	93.16	93.16	0.863
5-fold CV (Database 1)	85.36 ± 4.06	87.20 ± 2.38	85.60 ± 4.03	85.20 ± 4.43	0.707 ± 0.081
Independent test set (Database 1)	88.11	87.91	88.12	87.99	0.760
Hold-out (Database 2)	80.76	79.45	80.86	79.90	0.599

FPGA device for the 5-fold CV case. For implementing the trained LWDCNN model on the PYNQ-Z2 FPGA device, the FDP-reduced precision-based representation is used to reduce the model size, resource utilization, computational complexity, and power consumption. However, the CPU-based framework uses the original FP32 representation of the weight parameters of the LWDCNN model to detect BC using BUS images in the inference phase. In this study, the quantization of the LWDCNN model parameters, such as weight, bias, and activation function, is performed using <20,8>-based FDP-reduced precision representation. Hence, there is a reduction in the classification performance of the LWDCNN model when the inference is performed on the PYNQ-Z2 FPGA device compared to the CPU framework. The accuracy of LWDCNN has been improved to 88.11% when the inference is performed on the PYNQ-Z2 FPGA device to classify benign and malignant tumors using BUS images from an independent test set of database 1. The reduced precision <20,8>-based FDP representation works like a regularizer by limiting the LWDCNN model's capacity to memorize the training BUS images [30]. It helps prevent overfitting and improves the classification accuracy of the LWDCNN model by using the independent test set of BUS images. The LWDCNN model considered only 130 BUS images from database 2 during the inference phase, whereas the remaining 474 BUS images have been used to train the model. Due to the limited sample size in the training phase, the LWDCNN model has demonstrated less classification accuracy for the second database than the BUS images from the first database case to detect malignant tumor classes.

B. RESOURCE UTILIZATION OF LWDCNN ON FPGA

Furthermore, we have evaluated the accuracy, resource utilization, and on-chip power of the proposed LWDCNN model with different FDP precision cases for the weight parameters deployed on the PYNQ-Z2 FPGA device to classify benign and malignant tumor classes using BUS images. These results are displayed in Table 4. It is seen that the accuracy of the LWDCNN model is reduced with a decrease in the FDP precision cases for implementing the proposed LWDCNN model on the PYNQ-Z2 FPGA framework. The resources available for PYNQ-Z2 FPGA are lookup table (LUT) as 53200, LUTRAM as 17400, flip-flop (FF) as 106400, BRAM as 140, DSP block as 220 and BUFG as 32. It is noted that the proposed LWDCNN model with <20,8>-based FDP precision case of weight parameters requires a higher percentage in the utilization of resources such as LUT, LUTRAM, FF, BRAM, and DSP blocks as compared to the LWDCNN models with <18,8>, <16,6>, <14,6> and <12,4>-based FDP precision cases. The total on-chip power values obtained after the implementation of the proposed LWDCNN models on FPGA device using <20,8>, <16,6>, and <12,4>-based FDP precision cases of weight parameters are 1.86 watt, 1.79 watt, and 1.68 watt, respectively. Though the LWDCNN model with <20,8>-based FDP precision case of weight parameters requires higher resources and on-chip power on PYNQ-Z2 FPGA device, it has demonstrated the accuracy value of 93.16%, which is highest as compared to other four FDP precision cases of weight parameters cases to categorize benign and malignant tumors using BUS images. Though the utilization of resources is less for lower FDP precision cases of the weight parameters of the proposed LWDCNN model, the accuracy is obtained as 93.16% using <20,8> bit-width FDP precision case. Hence, we have selected <20,8> bit-width FDP precision case of weight parameters for the LWDCNN model to detect malignant tumor classes using BUS images.

C. INTERPRETABILITY OF LWDCNN MODEL

To enhance the interpretability of the proposed LWDCNN model, we have evaluated the class activation map (CAM) and t-distributed stochastic neighbor embedding (t-SNE) plots [37], [38]. The CAMs for benign and malignant tumors are displayed in Fig. 7 (b) and Fig. 7 (d), respectively. Here, the CAM plot is drawn using the feature map of one BUS image of each class (benign or malignant). The original BUS images from database 2 for benign and malignant tumors are visualized in Fig. 7 (a) and Fig. 7 (c), respectively. It is seen

TABLE 4. Accuracy and resource utilization of different FDP precision of weight parameters for the LWDCNN model referenced on PYNQ-Z2 FPGA framework.

FDP precision case	Accuracy(%)	LUT	LUTRAM	FF	BRAM	DSP	On chip power (Watt)
<20,8>	93.16	64% (34048)	11% (1914)	44% (46816)	80% (112)	10% (22)	1.86
<18,8>	64.71	60% (31920)	10% (1740)	40% (42560)	75% (105)	8% (17)	1.82
<16,6>	77.89	55% (29260)	9% (1566)	36% (38304)	70% (98)	7% (15)	1.79
<14,6>	54.98	49% (26068)	8% (1392)	32% (34048)	65% (91)	7% (15)	1.73
<12,4>	50.36	44% (23408)	7% (1218)	28% (29792)	60% (84)	7% (15)	1.68

TABLE 5. Variation of the classification performance of LWDCNN with different architectures and parameters.

CNN with different architectures	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	Parameters
Architecture 1	94.03	94.35	93.96	94.01	22157
Architecture 2	95.38	95.39	95.37	95.38	11181
Architecture 3	88.40	88.58	88.46	88.40	913
LWDCNN	94.15	94.15	94.15	94.15	1837

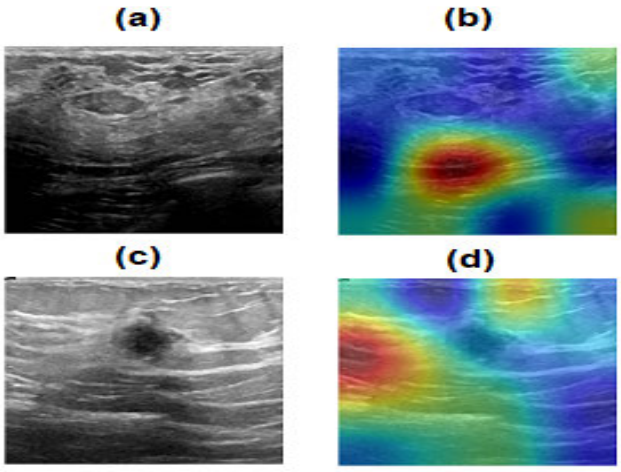

FIGURE 7. (a) BUS image for benign class. (b) LWDCNN-based Class activation map (CAM) for the benign class. (c) BUS image for malignant class. (d) LWDCNN-based CAM for the malignant class.

that the characteristics of CAMs are different for benign and malignant tumors. Hence, the LWDCNN model has learned different representations before the flattened layer for benign and malignant tumors. The high activation regions (marked in red or yellow) in the CAM plot indicate that these regions are most significant for classifying benign and malignant tumors using BUS images. Similarly, the regions marked in blue or green colors are less significant, as predicted by the proposed LWDCNN model. The t-SNE plot evaluated using original BUS images for benign and malignant tumors is shown in Fig. 8 (a). In this work, the t-SNE plot is drawn using 559 BUS images from database 2. Similarly, the t-SNE plot computed using the dense layer of the LWDCNN model for the test data (138 BUS images) is depicted in Fig. 8 (b). The blue color-filled circles represent the feature vectors for benign tumor classes. Similarly, the feature vectors for malignant tumor classes are shown in red color-filled circles. The learned features obtained in the dense layer of the LWDCNN are more discriminative than the features from the original BUS images using t-SNE, as shown in Fig. 8 (a). Hence, from CAM and t-SNE plots, it is observed that the proposed LWDCNN learns discriminative features from BUS images for the automated classification of benign and malignant tumors.

We have evaluated the classification performance of LWDCNN by considering different CNN architectures to categorize benign and malignant tumors using BUS images, and these results are displayed in Table 5. It is noted that when we have considered architecture 1, the accuracy has been reduced to 94.03%. The number of parameters has been increased to 11181. Similarly, the accuracy value of the CNN architecture 2 is obtained as 95.38%. Removing layers increases the number of parameters of the deep CNN models. Hence, it is impractical to implement these higher parameters-based deep CNN models on PYNQ-Z2-based RCE devices to categorize benign and malignant tumors using BUS images. Moreover, when we vary the number of neurons in the dense layer (architecture 3), the accuracy of the LWDCNN model has been reduced to 88.40%. Therefore, we have selected a lightweight LWDCNN model (only 1837 parameters) with three convolutions, three batch-normalization layers, three max-pooling layers, and one dense layer to classify BUS images.

D. IMPLEMENTATION OF LWDCNN ON ANDROID AND WAPP

We have deployed the FP32 version of the trained LWDCNN model on the streamlit cloud and prepared a WAPP for the real-time prediction of benign and malignant tumors using BUS images. The inference of the LWDCNN model on streamlit-based WAPP is shown in Fig. 9. It is observed that both BUS images are correctly classified using the LWDCNN model on WAPP. Furthermore, we have quantized the proposed LWDCNN model parameters using FP32 and INT8-based quantization cases using Tensorflow-lite. The classification results of the LWDCNN using unquantized (FP32), FP16, and INT8 cases are shown in Table 6. For the INT8 case, the accuracy of the LWDCNN model is obtained as 93.76%, which is lower than the unquantized version FP32

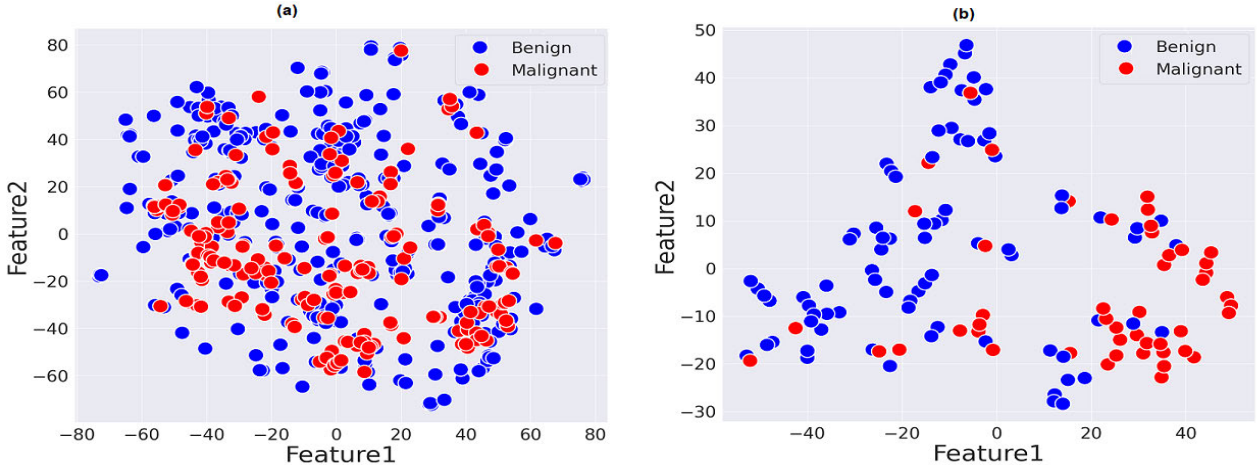

FIGURE 8. (a) t-SNE plot for the original BUS images for benign and malignant tumors using the second database. (b) t-SNE plot for the feature vector evaluated in the dense layer (16 neurons) of the LWDCNN model for the test data (BUS images used in the testing phase) of benign and malignant tumors using the second database.

TABLE 6. Variations of the accuracy of LWDCNN with different quantization cases to detect malignant tumors using BUS images.

Quantization case	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa (%)	Model size
FP32	94.15	94.15	94.15	94.15	0.883	11.13KB
FP16	93.86	93.82	93.82	93.82	0.872	9.10KB
INT8	93.76	93.72	93.70	93.72	0.870	7.75KB

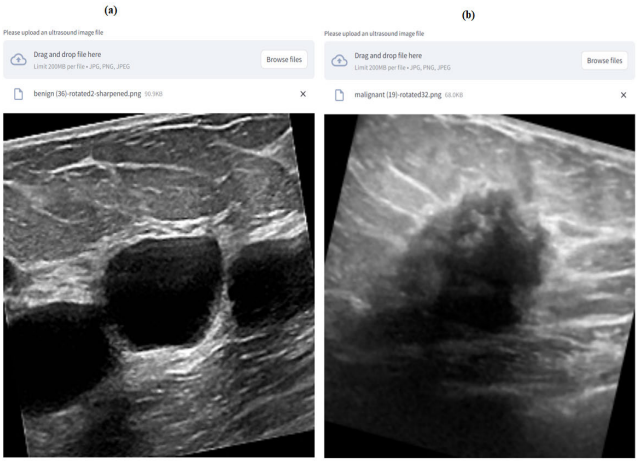

FIGURE 9. Inference of LWDCNN model on streamlit cloud to detect (a) benign class and (b) Malignant class, using BUS images.

FIGURE 10. Inference of LWDCNN model on android device to detect (a) benign class (b) Malignant class, using BUS images.

TABLE 7. Inference time values of the LWDCNN model tested on different frameworks using 1624 BUS images.

Inference platforms	Inference time
Streamlit cloud	1331.68 sec
Android device (Mobile phone)	1071.84 sec
PYNQ-Z2 FPGA	6.44 sec

case to detect malignant tumors using BUS images. Though the accuracy value of the LWDCNN model is reduced by 0.39% for INT8-case, the LWDCNN model size has been reduced for implementing the model on an Android device to detect malignant tumors using BUS images. The PTQ of the LWDCNN model uses the reduced precision-based representation of weight parameters and activation functions. Due to this, the accuracy of the LWDCNN model has been reduced for FP16 and INT8 cases, as it is for the FP32 case for the classification of BUS images in the inference phase. The Android-implemented LWDCNN model for categorizing benign and malignant tumors using BUS images is shown in Fig. 10 (a) and Fig. 10 (b), respectively. We have shown

TABLE 8. Comparison with different transfer learning models to detect malignant tumors using BUS images (each model is evaluated using 1624 BUS images in the inference phase).

Models used	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa (%)	Inference time	Parameters
ResNet50 [25]	73.95	82.21	73.92	72.15	47.87	82.08sec	23.59 Million
DenseNet201 [26]	90.82	91.09	90.82	90.80	0.816	11.97sec	18.35 Million
MobileNetV2 [27]	86.45	86.46	86.45	86.45	0.729	10.29sec	2.27 Million
InceptionV3 [28]	75.18	81.91	75.21	73.81	0.503	41.07sec	21.81 Million
XceptionNet [29]	90.88	91.07	90.88	90.87	0.817	142.06sec	20.89 Million
LWDCNN (CPU)	94.15	94.15	94.15	94.15	0.883	9.83sec	1837
LWDCNN (PYNQ-Z2 FPGA)	93.16	93.16	93.16	93.16	0.863	6.44sec	1837

TABLE 9. Comparison with existing deep learning-based methods to detect malignant tumors using BUS images from database 2.

Method considered	Accuracy(%)	Precision(%)	Recall(%)	Parameters	Simulation time
BusNet model [5]	56.60	62.90	66.40	25.6 Million	6.97 sec
MSADL Model [2]	79.88	79.88	79.91	12,2038	5.21 sec
Proposed LWDCNN model	80.76	79.45	80.86	1837	0.83 sec

TABLE 10. Comparison with an existing deep CNN-based method on PYNQ-Z2 FPGA to detect malignant tumors using BUS images from database 1.

Methods used	Accuracy (%)	LUT	LUTRAM	FF	BRAM	DSP
Guddati et al. [9]	87.05	68% (36189)	12% (2084)	51% (53793)	83% (116)	12% (27)
Proposed work	93.16	64% (34048)	11% (1914)	44% (46816)	80% (112)	10% (22)

the variations in the latency values for the implementation of LWDCNN using different platforms such as streamlit cloud, Android device, and PYNQ-Z2 FPGA device, respectively, with 1624 BUS images in the inference phase in Table 7. The inference times of the proposed LWDCNN model are more than 1000 seconds for processing 1624 BUS images on the cloud and an Android device. The latency is reduced when the inference of the LWDCNN model is performed using an Android device compared to the cloud-based framework used to detect BC using BUS images. The latency has been reduced to 6.44 sec when the inference of the LWDCNN model is performed on the PYNQ-Z2 FPGA device. Hence, the PYNQ-Z2-based resource-constrained FPGA device provides low power and low latency and is suitable for healthcare edge computing applications such as the automated detection of malignant tumors using BUS images.

E. COMPARISON WITH TRANSFER LEARNING

Furthermore, we have compared the classification performance of the proposed LWDCNN model with ResNet50, InceptionV3, XceptionNet, DenseNet201, and MobileNetV2-based transfer learning techniques to categorize benign and malignant tumors using BUS images, and these results are shown in Table 8. It is noted that the ResNet50 and InceptionV3 models have obtained accuracy values of less than 80% for classifying BUS images. The Densenet201, MobileNetV2, and XceptionNet have yielded accuracy values of less than 90% for BC detection using BUS images. The proposed LWDCNN inference on CPU and FPGA-based edge devices has demonstrated higher classification accuracy than various transfer learning models. The inference time of the deep learning model increases with the number of layers and weight parameters. The time complexity in a dense layer is given as $O(Ln_in_o)$ [2] where L is the number of dense layers. The parameters n_i and n_o are the number of input and output neurons in each dense layer. Similarly, the time complexity in a convolution layer is given as $O(nmk^2c)$ [2]. where n, and m are the dimensions of the output feature map. k and c are the dimensions of the kernel and number of channels, respectively. The total inference time is calculated by adding the time complexity of each layer. The number of layers in ResNet50, DenseNet201, MobileNetV2, InceptionV3 and XceptionNet models are 50, 201, 46, 104, and 71, respectively. Due to the higher number of layers, the inference times of these transfer learning models are high. The proposed LWDCNN model has only three convolution layers and one dense layer. Hence, it has less inference time than ResNet50, DenseNet201, MobileNetV2, InceptionV3, and XceptionNet models for classifying benign and malignant tumors using BUS images. The total number of parameters for MobileNetV2, ResNet50, DenseNet201, InceptionV3, and XceptionNet models in the inference phase is more than 2 Million. However, the proposed LWDCNN model has only 1837 total parameters. Due to fewer total parameters, the proposed LWDCNN model has less inference time on the CPU than all transfer learning models to detect BC using BUS images.

F. COMPARISON WITH EXISTING METHODS

Moreover, we have compared the classification performance of the proposed LWDCNN model with BusNet and MSADL methods inference on CPU framework to detect malignant tumors using the BUS images from database 2. The comparison results are depicted in Table 9. It is noted that both BusNet and MSADL models have obtained accuracy

TABLE 11. Comparison with existing methods to detect malignant tumors using BUS images from different databases.

Methods	Accuracy(%)	Images used
ResNetV2-based transfer learning [39]	78.11	769
Adaptive histogram equalization and VGG19 [40]	85.00	831
6-dense layer based deep learning model [41]	92.50	1051
Grayscale and Doppler features with Logistic regression [42]	84.93	140
Proposed LWDCNN model	80.76	657

and recall values of less than 80% for detecting malignant tumors using BUS images. The proposed LWDCNN model has produced accuracy and recall values of 80.76% and 80.86% for classifying benign and malignant tumors using an imbalanced BUS image database. However, the precision value of the MSADL method is higher than the proposed LWDCNN model. The BusNet and MSADL models have more than 1, 20, 000 parameters. The time complexity values of BusNet and MSADL methods are high due to more layers and parameters than the proposed LWDCNN model [2]. The latency or simulation time of the LWDCNN model is less than that of the BusNet and MSADL models. Hence, the proposed LWDCNN is a lightweight deep-learning model suitable for detecting malignant tumors using BUS images on edge computing devices. Furthermore, the accuracy and resource utilization of the proposed LWDCNN model are compared with an existing deep CNN method proposed by Guddati et al. [9] to detect malignant tumors using BUS images on PYNQ-Z2 FPGA. The deep CNN model was introduced in [9] to detect pneumonia and tuberculosis diseases using chest X-ray images. In this work, we have evaluated the classification performance of the deep CNN model for classifying benign and malignant tumors using BUS images. The comparison results are depicted in Table 10. The classification performance of deep CNN and LWDCNN is evaluated using BUS images from database 1. It is observed that the accuracy of the proposed LWDCNN model is higher than that of the deep CNN for detecting malignant tumors using BUS images. The deep CNN model consumes more resources (LUT, LUTRAM, FF, DSP, and BRAM) on PYNQ-Z2 FPGA than the proposed LWDCNN model to detect malignant tumors using BUS images. The deep CNN model has two dense layers (8 and 16 neurons) and 8 filters in the second convolution layer compared to the proposed LWDCNN architecture shown in Fig. 1. Hence, it uses more resources on FPGA than the proposed LWDCNN model. The latency of the proposed LWDCNN model inference on PYNQ-Z2 FPGA is less than that of CPU inference for detecting malignant tumors using BUS images from database 1.

Furthermore, we have also compared the performance of the proposed LWDCNN model with various existing methods to classify benign and malignant classes using BUS images from different databases, and these comparison results are depicted in Table 11. The proposed LWDCNN model has shown higher accuracy in detecting malignant tumors than the ResNetV2-based transfer learning technique [39]. The methods reported in [40], [41], and [42] have demonstrated higher classification accuracy than the LWDCNN model to detect malignant tumors using BUS images. The machine learning-based approach mentioned in [42] requires extracting features and selecting features for detecting malignant tumors using BUS images. The VGG19 model has 143.67 million parameters in the inference phase. Similarly, the 6-dense layer-based deep learning model has 4161 parameters. The proposed LWDCNN model has only 1837 parameters, less than other deep learning-based methods for detecting BC using BUS images.

G. LIMITATIONS AND FUTURE DIRECTIONS

In this work, only 647 BUS images are given in database 2, and we used these images to develop and evaluate our proposed LWDCNN model for detecting malignant tumors. The developed LWDCNN model can be tested using BUS images in the hospital with more patients to detect BC. The proposed LWDCNN-based automated system can assist radiologists and clinicians as a computer-aided framework for the automated detection of malignant tumors using BUS images. However, the security and service availability are very challenging for edge devices for real-time patient monitoring [43]. The security of the deep learning model parameters and patient data is crucial in automated diagnostic systems. The tampering of the model parameters can lead to the false diagnosis of the disease and disrupt the service continuity of the automated diagnostic systems. Hence, robust security protocols such as deep learning model encryption are required to help make model predictions secure for clinical studies. In the clinical environment, the LWDCNN model needs to successfully predict malignant tumors from BUS images from other sources or hospitals. Hence, the model parameters must be updated periodically to avoid service interruption. The federated learning can obtain the updated model without transferring the patient's data to the cloud [44]. The proposed LWDCNN model can be tested in the federated learning framework to obtain the optimized model for deploying and inference on edge devices to detect malignant tumors using BUS images from multiple hospitals. The service-oriented network (SON) helps the edge nodes adapt to changes in the environment. Such networks use stimulation or suppression chains to protect against malfunctions and intrusions. Hence, the SONs are essential and relevant in the clinical setting for service continuity of automated diagnostic systems. In [45], authors have leveraged FPGA technology with the implementation of a SON-based service in the clinical setting. The SON can be used to improve service continuity and robustness within the FPGA-based proposed LWDCNN system for the automated detection of malignant tumors using BUS images.

V. CONCLUSION

A lightweight LWDCNN model has been proposed in this paper to detect malignant tumors using BUS images. The LWDCNN model has been trained on the cloud-based framework to obtain the optimized weight and bias parameters. The register-transfer level (RTL) and post-implementation of the optimized LWDCNN model have been generated using VIVADO through the HLS4ML framework. The inference of the LWDCNN model has been performed on a PYNQ-Z2 FPGA device to categorize benign and malignant tumors using BUS images. Furthermore, the post-training INT8-based quantization has been utilized for the proposed LWDCNN to obtain the reduced precision-based representation of the model parameters. The LWDCNN model has been deployed on FPGA and Android-based edge devices to detect malignant tumors using BUS images. The proposed LWDCNN has yielded accuracy values of more than 90% in the inference phases on Android and FPGA-based edge devices. The inference time of the proposed LWDCNN model is 6.44 seconds on an FPGA-based edge device, which is lower than that on CPU, Android, and cloud-based frameworks for detecting malignant tumors using BUS images. The federated learning-based automated detection of malignant tumors using BUS images from different sources or hospitals is essential in intelligent healthcare systems for better management of BC. The proposed lightweight LWDCNN model can be tested in the federated learning scenario for classifying benign and malignant tumors using BUS images.

REFERENCES

[1] K. E. Reeder-Hayes and B. O. Anderson, "Breast cancer disparities at home and abroad: A review of the challenges and opportunities for system-level change," *Clin. Cancer Res.*, vol. 23, no. 11, pp. 2655–2664, Jun. 2017.

[2] A. Gade, D. K. Dash, T. M. Kumari, S. K. Ghosh, R. K. Tripathy, and R. B. Pachori, "Multiscale analysis domain interpretable deep neural network for detection of breast cancer using thermogram images," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[3] J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 236–251, Jan. 2009.

[4] F.-L. Wang, F. Chen, H. Yin, N. Xu, X.-X. Wu, J.-J. Ma, S. Gao, J.-H. Tang, and C. Lu, "Effects of age, breast density and volume on breast cancer diagnosis: A retrospective comparison of sensitivity of mammography and ultrasonography in China's rural areas," *Asian Pacific J. Cancer Prevention*, vol. 14, no. 4, pp. 2277–2282, Apr. 2013.

[5] Y. Li, H. Gu, H. Wang, P. Qin, and J. Wang, "BUSnet: A deep learning model of breast tumor lesion detection for ultrasound images," *Frontiers Oncol.*, vol. 12, Mar. 2022, Art. no. 848271.

[6] R. Guo, G. Lu, B. Qin, and B. Fei, "Ultrasound imaging technologies for breast cancer detection and management: A review," *Ultrasound Med. Biol.*, vol. 44, no. 1, pp. 37–70, Jan. 2018.

[7] M. N. Bhuiyan, M. M. Rahman, M. M. Billah, and D. Saha, "Internet of Things (IoT): A review of its enabling technologies in healthcare applications, standards protocols, security, and market opportunities," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10474–10498, Jul. 2021.

[8] G. Sivapalan, K. K. Nundy, S. Dev, B. Cardiff, and D. John, "ANNet: A lightweight neural network for ECG anomaly detection in IoT edge sensors," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 1, pp. 24–35, Feb. 2022.

[9] P. Guddati, S. Dash, and R. K. Tripathy, "FPGA implementation of the proposed DCNN model for detection of tuberculosis and pneumonia using CXR images," *IEEE Embedded Syst. Lett.*, early access, Feb. 27, 2024, doi: 10.1109/LES.2024.3370833.

[10] Y.-C. Lee, D.-H. Ko, M.-H. Son, S.-H. Yang, and J.-Y. Um, "Arterial distension monitoring scheme using FPGA-based inference machine in ultrasound scanner circuit system," *IEEE Trans. Biomed. Circuits Syst.*, vol. 18, no. 3, pp. 702–713, Jun. 2024.

[11] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, no. 1, pp. 299–317, Jan. 2010.

[12] W. K. Moon, Y.-W. Lee, H.-H. Ke, S. H. Lee, C.-S. Huang, and R.-F. Chang, "Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 190, Jul. 2020, Art. no. 105361.

[13] Y. Gu et al., "Deep learning based on ultrasound images assists breast lesion diagnosis in China: A multicenter diagnostic study," *Insights Imag.*, vol. 13, no. 1, p. 124, Dec. 2022.

[14] M. Byra, "Breast mass classification with transfer learning based on scaling of deep representations," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102828.

[15] W.-C. Shia and D.-R. Chen, "Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine," *Computerized Med. Imag. Graph.*, vol. 87, Jan. 2021, Art. no. 101829.

[16] A. Raza, N. Ullah, J. A. Khan, M. Assam, A. Guzzo, and H. Aljuaid, "DeepBreastCancerNet: A novel deep learning model for breast cancer detection using ultrasound images," *Appl. Sci.*, vol. 13, no. 4, p. 2082, Feb. 2023.

[17] S.-Y. Lu, S.-H. Wang, and Y.-D. Zhang, "BCDNet: An optimized deep network for ultrasound breast cancer detection," *IRBM*, vol. 44, no. 4, Aug. 2023, Art. no. 100774.

[18] Y. Mo, C. Han, Y. Liu, M. Liu, Z. Shi, J. Lin, B. Zhao, C. Huang, B. Qiu, Y. Cui, L. Wu, X. Pan, Z. Xu, X. Huang, Z. Li, Z. Liu, Y. Wang, and C. Liang, "HoVer-trans: Anatomy-aware HoVer-transformer for ROI-free breast cancer diagnosis in ultrasound images," *IEEE Trans. Med. Imag.*, vol. 42, no. 6, pp. 1696–1706, Jun. 2023.

[19] S. Balasubramaniam, Y. Velmurugan, D. Jaganathan, and S. Dhanasekaran, "A modified LeNet CNN for breast cancer diagnosis in ultrasound images," *Diagnostics*, vol. 13, no. 17, p. 2746, Aug. 2023.

[20] K. S. Rao, P. V. Terlapu, D. Jayaram, K. K. Raju, G. K. Kumar, R. Pemula, M. V. Gopalachari, and S. Rakesh, "Intelligent ultrasound imaging for enhanced breast cancer diagnosis: Ensemble transfer learning strategies," *IEEE Access*, vol. 12, pp. 22243–22263, 2024.

[21] M. G. Lanjewar, K. G. Panchbhai, and L. B. Patle, "Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images," *Comput. Biol. Med.*, vol. 169, Feb. 2024, Art. no. 107914.

[22] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.

[23] V. A. Sairam. (Nov. 2022). *Ultrasound Breast Images for Breast Cancer*. Accessed: Aug. 15, 2023. [Online]. Available: https://www.kaggle.com/datasets/vuppalaadithyasairam/ultrasound-breast-images-for-breast-cancer?resource=download

[24] J. Karhade, S. Dash, S. K. Ghosh, D. K. Dash, and R. K. Tripathy, "Time–frequency-domain deep learning framework for the automated detection of heart valve disorders using PCG signals," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[30] T. Aarrestad, V. Loncar, N. Ghielmetti, M. Pierini, S. Summers, J. Ngadiuba, C. Petersson, H. Linander, Y. Iiyama, G. Di Guglielmo, J. Duarte, P. Harris, D. Rankin, S. Jindariani, K. Pedro, N. Tran, M. Liu, E. Kreinar, Z. Wu, and D. Hoang, "Fast convolutional neural networks on FPGAs with hls4ml," *Mach. Learn., Sci. Technol.*, vol. 2, no. 4, Dec. 2021, Art. no. 045015.
[31] F. Fahim et al., "Hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices," 2021, *arXiv:2103.05579*.
[32] F. Yu, L. Cui, P. Wang, C. Han, R. Huang, and X. Huang, "EasiEdge: A novel global deep neural networks pruning method for efficient edge computing," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1259–1271, Feb. 2021.
[33] S. Abdelhedi, M. Baklouti, R. Bourguiba, and J. Mouine, "Vivado HLS-based implementation of a fall detection decision core on an FPGA platform," in *Proc. 11th Int. Design Test Symp. (IDT)*, Dec. 2016, pp. 115–120.
[34] P. Warden and D. Situnayake, *Tinyml: Machine Learning With Tensorflow Lite on Arduino and Ultra-low-power Microcontrollers*. Sebastopol, CA, USA: O'Reilly Media, 2019.
[35] V. S. Parupudi, A. K. Panda, and R. K. Tripathy, "A smartphone-enabled deep learning approach for myocardial infarction detection using ECG traces for IoT-based healthcare applications," *IEEE Sensors Lett.*, vol. 7, no. 11, pp. 1–4, Nov. 2023.
[36] M. A. Raheem, S. K. N. S. Tabassum, and S. A. Anzer, "A deep learning approach for the automatic analysis and prediction of breast cancer for histopathological images using a webapp," *Int. J. Eng. Res. Technol.*, vol. 10, no. 6, pp. 996–1001, 2021.
[37] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021.
[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, Nov. 2008.
[39] Q. Wang, H. Chen, G. Luo, B. Li, H. Shang, H. Shao, S. Sun, Z. Wang, K. Wang, and W. Cheng, "Performance of novel deep learning network with the incorporation of the automatic segmentation network for diagnosis of breast cancer in automated breast ultrasound," *Eur. Radiol.*, vol. 32, no. 10, pp. 7163–7172, Apr. 2022.
[40] A. Boulenger, Y. Luo, C. Zhang, C. Zhao, Y. Gao, M. Xiao, Q. Zhu, and J. Tang, "Deep learning-based system for automatic prediction of triple-negative breast cancer from ultrasound images," *Med. Biol. Eng. Comput.*, vol. 61, no. 2, pp. 567–578, Feb. 2023.
[41] W.-J. Shen, H.-X. Zhou, Y. He, and W. Xing, "Predicting female breast cancer by artificial intelligence: Combining clinical information and BI-RADS ultrasound descriptors," *WFUMB Ultrasound Open*, vol. 1, no. 2, Dec. 2023, Art. no. 100013.
[42] T. Wu, L. R. Sultan, J. Tian, T. W. Cary, and C. M. Sehgal, "Machine learning for diagnostic ultrasound of triple-negative breast cancer," *Breast Cancer Res. Treat.*, vol. 173, no. 2, pp. 365–373, Jan. 2019.
[43] S. Wang, J. Xu, N. Zhang, and Y. Liu, "A survey on service migration in mobile edge computing," *IEEE Access*, vol. 6, pp. 23511–23528, 2018.
[44] S. Trindade, L. F. Bittencourt, and N. L. S. D. Fonseca, "Resource management at the network edge for federated learning," *Digit. Commun. Netw.*, vol. 10, no. 3, pp. 765–782, Jun. 2024.
[45] V. Conti, C. Militello, L. Rundo, and S. Vitabile, "A novel bio-inspired approach for high-performance management in service-oriented networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 4, pp. 1709–1722, Oct. 2021.

ADITYA VINOD is currently pursuing the bachelor's degree in electronics and communication engineering with the Birla Institute of Technology and Sciences Pilani, Hyderabad, India. His research interests include machine learning, deep learning, and embedded systems.

PRABHAV GUDDATI is currently pursuing the bachelor's degree in electronics and communication engineering with the Birla Institute of Technology and Sciences Pilani, Hyderabad, India. His research interests include machine learning, deep learning, and embedded systems.

AMIT KUMAR PANDA (Senior Member, IEEE) received the M.Sc. degree in electronics from Berhampur University, Berhampur, India, in 2004, the M.Tech. degree in electronic design and technology from Tezpur Central University, Tezpur, India, in 2009, and the Ph.D. degree in electrical engineering from IIT Patna, Patna, India, in 2020. He was an Assistant Professor with the Electronics and Communication Engineering Department, Guru Ghasidas Vishwavidyalaya, Bilaspur, India, from 2009 to 2012. He was an Assistant Professor with the Centre for Nanotechnology, Central University of Jharkhand, Ranchi, India, from 2012 to 2013. He has been an Assistant Professor with the Department of Electrical and Electronics Engineering, BITS Pilani, Hyderabad Campus, Hyderabad, India, since 2020. His research interests include VLSI architectural design, FPGA-based system design, FPGA-based hardware accelerator, VLSI cryptography, and hardware security for IoT systems.

RAJESH KUMAR TRIPATHY (Senior Member, IEEE) received the B.Tech. degree in electronics and telecommunication engineering from the Biju Pattnaik University of Technology, Rourkela, the M.Tech. degree in biomedical engineering from the National Institute of Technology, Rourkela, and the Ph.D. degree in electronics and electrical engineering from the Indian Institute of Technology, Guwahati, India. He was an Assistant Professor with the Faculty of Engineering and Technology (FET), Siksha 'O' Anusandhan, from March 2017 to June 2018. Since July 2018, he has been an Assistant Professor with the Department of Electrical and Electronics Engineering (EEE), Birla Institute of Technology and Science (BITS), Pilani, Hyderabad Campus. He has published research papers in reputed international journals and conferences. His current research interests include machine learning, deep learning, biomedical signal processing, sensor data processing, medical image processing, embedded artificial intelligence, and the Internet of Things (IoT) for healthcare. He has served as a reviewer for more than 15 scientific journals and a technical program committee (TPC) member for various national and international conferences. He is an Associate Editor of IEEE Access and *Frontier in Physiology*.