

Low-Power Quantized Convolutional Neural Network for Early Breast Cancer Detection in Remote Communities

Kawthar Dellel* ^{† **}, Emanuel Trabes* ^{‡‡ †}, Hana Ben Fredj^{§, ** ††}
Carlos Valderrama* [¶] and Hassene Faiedh^{|| ** ††}

*Service d'électronique et de Microélectronique,
University of Mons, Belgium

**Laboratory of Electronics and Microelectronics,
Faculty of Science of Monastir, University of Monastir

^{††} Higher Institute of Applied Sciences & Technology,
University of Sousse, Sousse, Tunisia

^{‡‡} Department of Electronics,

Universidad Nacional de San Luis, Argentina

[†] KAOUTHER.DELLEL@student.umons.ac.be, [‡] emanuel.trabes@umons.ac.be,

[§] ben.fredj.hanaa@gmail.com, [¶] carlos.valderrama@umons.ac.be,

^{||} hassene.faiiedh@gmail.com

Abstract—Breast cancer remains a leading cause of mortality in low-resource communities due to late diagnosis and limited access to specialized healthcare facilities. Leveraging recent advancements in deep learning, this study presents a low-power, cost-effective solution for early breast cancer detection tailored to resource-constrained environments. We developed a quantized convolutional neural network (CNN) using Brevitas and deployed it on a PYNQ-Z1 development board via the FINN framework. The CNN efficiently classifies ultrasonic topography images into three categories: normal, benign, or malignant. Achieving up to 91.2% accuracy on the BUSI dataset, our results highlight the effectiveness of 4-bit quantization, offering a viable trade-off between computational efficiency and accuracy. This makes it viable for real-time medical image classification in underserved communities, potentially facilitating early diagnosis and timely referrals, and ultimately contributing to improved healthcare outcomes.

Index Terms—FPGA, breast cancer detection, ultrasound, remote healthcare, FINN

I. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer-related deaths among women worldwide [1]. While significant advances have been made in early detection and treatment in high-income countries, low-resource settings, particularly in remote and impoverished regions, continue to face high mortality rates due to late diagnosis [2]. The main challenges in these areas include limited

access to healthcare facilities, shortage of specialized medical professionals, and inadequate diagnostic equipment [3].

Early detection of breast cancer significantly increases the chances of successful treatment and survival. Mammography is the standard screening tool in developed countries; however, its high cost, need for specialized equipment, and requirement of trained radiologists make it impractical in low-resource settings [4]. Ultrasound imaging emerges as a viable alternative due to its affordability, portability, and safety, as it does not involve ionizing radiation [5]. Despite its advantages, the interpretation of ultrasound images requires expert radiologists, who are scarce in remote communities [6].

To address the shortage of medical experts, several solutions have been proposed. The utilization of telecommunications technology to transmit medical data to specialists in urban centers for diagnosis [7] is one such proposal. While promising, this solution depends on reliable internet connectivity and electricity, which are often unreliable in remote areas [8]. Other alternative is the deployment of mobile units equipped with diagnostic tools and staffed by medical professionals to visit remote communities periodically [9]. This approach is logistically complex and cannot provide continuous access to care. Other option is training local individuals to perform basic

health screenings [10]. However, interpreting complex diagnostic images remains beyond the scope of basic training. Recently, implementing AI algorithms to assist in the interpretation of medical images [11] has surged as a viable alternative. AI has the potential to provide expert-level analysis without the need for a specialist on-site.

While AI diagnostics present a promising avenue, deploying AI solutions in low-resource settings faces several challenges. AI models, particularly deep learning algorithms, typically require significant computational power, which is not feasible in areas with limited electricity and hardware resources [12]. High-performance computing devices are expensive and not accessible to impoverished communities [13]. Power-hungry devices are impractical where electricity supply is intermittent or non-existent [15].

To overcome these challenges, we propose a low-power, cost-effective AI system for early breast cancer detection tailored for remote and underserved communities. The key components of our solution are:

- 1) The utilization of ultrasound images. These devices are relatively affordable, battery-operated, and safe for repeated use [16].
- 2) Developing a lightweight CNN model trained to classify ultrasound images into normal, benign, or malignant categories.
- 3) Implementing the quantized CNN on a PYNQ-Z1 development board. This hardware is energy-efficient, affordable, and capable of performing the necessary computations locally.

By processing the diagnostic images locally, our system eliminates the need for constant internet connectivity and reduces dependency on external power sources. The preliminary implementation of our quantized CNN achieves an accuracy of 91.2% in classifying breast ultrasound images, demonstrating the potential of this approach in aiding early detection and prompting timely medical referrals.

II. RELATED WORK

Previous studies have explored FPGA implementations for breast cancer detection using mammogram images and other data that may not be readily accessible in remote or impoverished areas. For instance, the authors in [20] introduced a Digital Mammogram Diagnostic Convolutional Neural Network (DMD-CNN) designed to classify mammogram images according to the BI-RADS scoring system. Their model achieved an accuracy of 98.2% and was deployed on a PYNQ-based Artix 7 FPGA, processing nearly 91 images per second with a power consumption of only 3.12

W. Despite these impressive results, the reliance on mammograms—which require X-ray imaging—makes deployment in impoverished locations challenging due to the need for specialized equipment.

Similarly, in [21], a linear Support Vector Machine (SVM) classifier was implemented on an FPGA for breast cancer detection using the Wisconsin Breast Cancer Dataset (WBCD). The implementation achieved a classification accuracy of 91.08% and utilized approximately 49% of the device’s resources, staying within a 25% band of total device capacity after post-implementation. However, the training data was derived from fine-needle aspiration studies, procedures that are difficult to perform in remote areas.

Another study presented in [22] developed a custom convolutional neural network (CNN) for classifying mammogram images as malignant or non-malignant, achieving high accuracy across multiple datasets—92.1% on MIAS, 96.8% on INbreast, and 98.2% on DDSM. This model was miniaturized for deployment on a PYNQ-Z2 FPGA board, demonstrating practical applicability in clinical settings with a reported accuracy of 99.38

In contrast, our work focuses on affordable and easily deployable sensors such as ultrasound-based devices. By advancing this approach with a focus on ultrasound data, we provide a novel contribution that addresses both the accuracy and deployment challenges in medical imaging.

III. METHODOLOGY

This section outlines the workflow adopted for the training, quantization, and deployment of our model on PYNQ-Z1 board. The complete process is depicted in Figure 2, which illustrates the key stages involved.

A. Dataset Preparation

The dataset used for training the model comprises ultrasound images categorized into three classes: Normal, Benign, and Malignant. The dataset was sourced from the publicly available Breast Ultrasound Images (BUSI) dataset [14]. This dataset is challenging. Firstly, because intrinsic characteristics of ultrasound images (when comparing to more easily usable x-ray imaging). Secondly, because of the low quantity of samples. The data was split into training, validation, and testing sets in the ratio of 80%, 10%, and 10%, respectively. The resulting training set is composed of 624 images, and the test and validations sets of 78 images. To avoid overfitting as much as possible and help generalization, we applied data augmentation techniques, namely: random horizontal and vertical

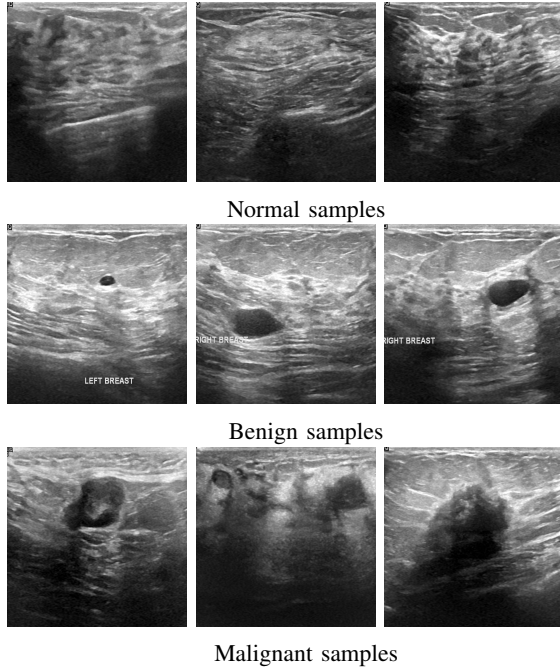


Fig. 1: Samples from BUSI dataset

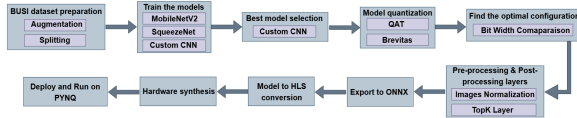


Fig. 2: Model deployment workflow

flips, random rotation and scale changes and random brightness and contrast changes. We finally resizing images to 255×255 pixels.

Figure 1 depicts some samples from the BUSI dataset.

B. Network Topology Selection

We analyzed three different network topologies specifically aimed to embedded devices: MobileNetV2 [23], SqueezeNet [24], and a custom CNN designed specifically for our use case. MobileNetV2 is known for its efficiency in mobile devices due to the use of depthwise separable convolutions. SqueezeNet aims to achieve AlexNet-level accuracy with 50 times fewer parameters. The custom CNN is a simplified network with fewer layers and parameters, tailored for efficient deployment on FPGA hardware.

1) *MobileNetV2*: We utilized the implementation provided in the Torchvision package. Since our dataset comprises grayscale images, we modified the first convolutional layer to accept single-channel input. We also adjusted the last layer to output three classes

corresponding to our application. The network weights were initialized using the pre-trained weights provided by Torchvision, and no further modifications were made to the network architecture.

2) *SqueezeNet*: Similarly, we employed the default implementation of SqueezeNet from the Torchvision package and modified its first layer to accept single-channel images. The final layer was adjusted to output three classes, and the network weights were initialized with the pre-trained weights available in Torchvision.

3) *Custom CNN*: We designed a custom CNN consisting of six convolutional layers and two fully connected (dense) layers. Each convolutional layer uses the ReLU activation function, with batch normalization applied between the convolutional and ReLU layers. MaxPooling is added at the end of each convolutional block, except for the last one. The number of output channels for the convolutional layers are 8, 16, 32, 32, 32, and 32, respectively. The first dense layer takes as input a tensor of $32 \times 4 \times 4$ size, and outputs one of 128 size. The second dense layer takes as input the previous layer output, and outputs a tensor of size 64. A final linear layer takes the 64 outputs from the last ReLU activation and outputs the three classes. Since no pre-trained weights were available for this network, it was trained from scratch.

4) *Training Settings*: At this stage, quantization was not applied to identify the baseline achievable accuracy for each network. The Adam optimizer was used with a learning rate of 0.0001. The optimization criterion was the CrossEntropyLoss function, commonly used in multi-class classification tasks. During training, we saved the model weights that yielded the highest test accuracy. Each model was trained for 200 epochs, with a batch size of 16.

5) *Network Comparison*: MobileNetV2 achieved an accuracy of 88.46%, where accuracy is defined as the percentage of correctly classified inputs. The confusion matrix for MobileNetV2 is shown in Table I.

TABLE I: Confusion Matrix for MobileNetV2

True Class	Normal	Benign	Malignant
Normal	76.9	15.3	7.6
Benign	4.5	93.1	2.2
Malignant	0.0	14.2	85.7

For the Normal category, the model correctly classified 76.9% of the samples, with 15.3% misclassified as benign and 7.6%. In the Benign category, 93.1% of samples were correctly classified, while 4.5% were misclassified as normal and 2.2% as malignant. The Malignant category had a correct classification rate

of 85.7%, with 14.2% missclassified into the Benign category.

SqueezeNet achieved an overall accuracy of 87.17%. The resulting confusion matrix can be seen in Table II.

TABLE II: Confusion Matrix for SqueezeNet

True Class	Normal	Benign	Malignant
Normal	92.3	7.69	0.0
Benign	2.2	86.36	11.36
Malignant	0.0	14.28	85.71

It was able to correctly classify normal images with a 93% accuracy, having 7.6% wrongly classified as benign. For the benign category, 86.3% accuracy was achieved. For the Malignant category, the model was able to achieve a 85.71% accuracy, with some missclassification into the benign category.

For the custom CNN, the achieved accuracy was of 89%. The confusion matrix is presented in Table VI.

TABLE III: Confusion Matrix for Custom CNN

True Class	Normal	Benign	Malignant
Normal	76.9	15.3	76.9
Benign	0.0	93.1	6.8
Malignant	0.0	19.0	80.95

In the Normal category, the custom CNN correctly classified 76.6% of the test cases, with 15.3% misclassified as Benign. For the Benign category, 93.1% of samples were correctly identified, while 6.8% were misclassified as Malignant. The Malignant category had a correct classification rate of 80.09%, with 19.0% of cases misclassified as Benign.

We can conclude that the achieved accuracy for the custom CNN is comparable to the best performing model, MobileNetV2. Furthermore, the sub-90% accuracy across all models highlights the inherent challenges of ultrasound imaging, where high noise levels and indistinct edges make feature recognition difficult, limiting model performance.

Nevertheless, we can observe that for our custom CNN model, the malignant class was either classified as malignant or benign, but never as normal. This means that if the models outputs a malignant category, the correct category could be either malignant (with 80% probability) or benign (with 19 % accuracy), but it will be very unlikely that the correct category was normal (0% probability with our test dataset). On the other hand, if the model outputs a normal category, it is very unlikely that the model misclassified it from a benign or malignant category (0 % probability

for our test data). This means that the models has a 100% accuracy to differentiate between normal or benign-malignant categories, result which could be very useful for our aim of early diagnosis.

C. Network Quantization

Quantization refers to the process of reducing the number of bits used to represent model weights and activations, thereby decreasing the computational complexity and memory footprint. We utilized Brevitas [25] for network quantization. The PyTorch layers Conv2D, ReLU, and Linear were replaced with their quantized counterparts: QuantConv2D, QuantReLU, and QuantLinear. Bias terms were not used, and the weights and activations were quantized. The bit width for each type of layer was left as a parameter, we explore several bit widths in the results. The quantized model was initialized from the weights of the unquantized version and retrained for 200 epochs to fine-tune the performance.

IV. RESULTS

Despite its difficulties, we were able to reach an accuracy of 90% in the classification of normal, benign and malignant classes, and a 100% accuracy in the classification of normal and benign-malignant classes (in our test dataset). This results highlights the usefulness of our model for early diagnosis. We used the PYNQ-Z1 board to test our model. It has a Xilinx Zynq-7000 (XC7Z020) SoC. The Programable logic (PL) has 13,300 Logic Slices, 630 KB BRAM, 220 DSP Slices. The Programable System part has dual core ARM A9 CPU, with 512GB of RAM memory. The main frequency use for the PL design was 100MHz.

TABLE IV: Comparison with existing works

Work	Dataset	Data type	Model	Board	Accuracy
[20]	BI-RADS scoring	Mammogram	DMD-CNN	Artix 7	98.2%
[21]	WBCD	Fine-Needle Aspiration (FNA)	Linear SVM	FPGA	91.08%
[22]	MIAS, INbreast, DDSM datasets	Mammogram	Custom CNN	PYNQ-Z2	92.1% (MIAS), 96.8% (INbreast), 98.2% (DDSM)
Our model	BUSI Dataset	Ultrasound	Custom CNN	PYNQ-Z1	91.2%

A. Bit width comparison

First, we analyze the impact of the selected bit width to the accuracy. In Table V, it is shown the accuracy of the model relative to the bit with of the weights and the activation layers.

TABLE V: Accuracy of Custom CNN by weight and activation width.

Weights - Activations	8 bits	4 bits	2 bits
8 bits	83.3	88.4	84.6
4 bits	84.6	89.7	88.4
2 bits	85.8	91.0	78.2

It can be seen a deterioration in the accuracy when reducing the bit width of both the weights and the activations to 8 bits. It can also be seen an improvement in the accuracy with further bit width reduction, with a peak in 2-4 bits. We observed that reducing the bit width of the weights to 2 bits, and activations to 4 bits, improved accuracy significantly. This counterintuitive result is likely due to the regularization effect of quantization. Lower precision can force the model to focus on more significant patterns in the data, reducing over-fitting and improving generalization, especially for datasets like BUSI, where features are more abstract and noise prominent. This phenomena is likely due to the relative low amount of training data, which gives rise to considerable over-fitting in high bit quantization. Also, its been reported in the literature that lower bit width introduces quantization related noise, that can also prevent overfitting and improve the performance of the model [26] [27]. It can also be observed that 2-2 bit widths gives the worst accuracy, which is to be expected when using such an extreme level of quantization. The accuracy drop at 2-bit activations (78.2%) illustrates the limitations of extreme quatization. With 2-bit quantization, the model’s representational capacity diminishes, making it difficult to capture the fine details required for medical imaging tasks.

The resulting confusion matrix of the 2-4 bit quantization can be observed in table VI

TABLE VI: 2 bit weights, 4 bit activation

True Class	Normal	Benign	Malignant
Normal	92.3	7.69	0.0
Benign	0.0	97.72	2.2
Malignant	0.0	23.8	76.19

B. Model preparation

To run inferences using the current quantized model, two further data processing stages are required. The first one is an image pre-processing stage, where the 8-bit images are converted to the floating point datatype, in the range of 0 to 1.0. The second one is a post-processing of the output of the model. The model

outputs a 3 value vector, representing the weight the model gives to the image being in one of the 3 categories. To output the most likely category, the index of the element in the vector with the highest values must be given.

Both these pre and post-processing steps could be run on the CPU. This would leave these two stages without hardware acceleration. Furthermore, the data movement between the CPU and the hardware accelerator is increased. For the pre-processing, an image of float-datatype has to be transferred to the accelerator, instead of an 8-bit image. And for the post-processing, 3 values have to be recovered from the accelerator instead of just 1, for the most likely class. This increase in the amount of data movement will deteriorate the performance of the system.

To avoid these issues, before hardware synthesis, both these stages were added to the model. The pre-processing was implemented as a simple division by 255 of the input image. For the post-processing stage, the FINN library provides the topK layer, which outputs the index of the element of a vector with the highest value. This layer was added into the model as the latest layer.

C. Execution preparation

To run the model on the board, we prepared a short python script. The 78 validation images were loaded from the ssd card one at a time. We used the OpenCV library for both read the image file and resize the images to 255×255 pixels. We read the true labels from the dataset, and compare them with the output from the model. We compute the mean accuracy and the mean processing time.

D. Hardware utilization and power consumption

We present result for the 2 most performing bit width configurations: 4 bits for weights, 4 bits for activations, and 2 bits for weights, 4 bits for activations.

E. 4-4 bit configuration

The design reached a runtime of 432.05783 ms, which gives a throughput 2.31 images per second. In table VII, it can be seen the resources utilized.

The results for power consumption were 1.719 W total, 1.576 W dynamic, 1.257 W PS, 0.256 W PL.

F. 2-4 bit configuration

In this case, the design reached a runtime of 317.66 ms, gibing a throughput 3.14 frames per second. In table VIII, the hardware resources are detailed.

The hardware usage is very similar to the 4-4 bit configuration. This is likely caused by the design

TABLE VII: Resources used for the 4-4 bit configuration, 3 fps target

Resource	Utilization	%
LUT	22320	41.95
LUTRAM	2601	14.95
FF	28044	26.36
BRAM	13	9.29
DSP	12	5.45

TABLE VIII: Resources used for the 2-4 bit configuration, 3 fps target

Resource	Utilization	%
LUT	22077	41.5
LUTRAM	2602	14.95
FF	27908	26.23
BRAM	13	9.29
DSP	10	3.13

characteristics of LUTs. These resources have 6 inputs each, which would give similar resource utilization if only 2 or 4 inputs were to be used. Similarly, the results for power consumption were almost identical to the 4-4 configuration, using 1.718 W total, 1.574 W dynamic, 1.257 W PS, 0.255 W PL.

V. CONCLUSION

In conclusion, this study demonstrates the feasibility and effectiveness of utilizing a quantized convolutional neural network (CNN) for early breast cancer detection in resource-constrained environments. By achieving a classification accuracy of 91.2% on the BUSI dataset, the proposed model not only addresses the pressing need for accessible healthcare solutions in remote communities but also showcases the potential of artificial intelligence in transforming medical diagnostics. The research highlights the significance of optimizing the bit width in CNNs, revealing that quantization can enhance model performance while maintaining low power consumption. The integration of pre-processing and post-processing stages into the hardware design further contributes to the efficiency of the system, making it suitable for deployment on low-power platforms like the PYNQ-Z1 development board. This work paves the way for future advancements in AI-driven healthcare technologies, emphasizing the importance of developing cost-effective and low-power solutions that can operate independently of constant internet connectivity. By empowering local healthcare providers with advanced diagnostic tools, we can improve early detection rates and ultimately enhance patient outcomes in underserved communities.

Future research will focus on expanding the dataset, refining the model and exploring additional applications for quantized CNNs in various medical imaging domains to further validate and extend the impact of this innovative approach.

REFERENCES

- [1] Bray, F., et al. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), 394-424.
- [2] Jemal, A., et al. (2012). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2), 69-90.
- [3] Sankaranarayanan, R., et al. (2011). Cancer survival in Africa, Asia, and Central America: a population-based study. *The Lancet Oncology*, 11(2), 165-173.
- [4] Smith, R. A., et al. (2017). Cancer screening in the United States, 2017: A review of current American Cancer Society guidelines and current issues in cancer screening. *CA: A Cancer Journal for Clinicians*, 67(2), 100-121.
- [5] Stavros, A. T., et al. (1995). Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology*, 196(1), 123-134.
- [6] Shulman, L. N., et al. (2010). Breast cancer in developing countries: opportunities for improved survival. *Journal of Oncology*, 2010, 1-6.
- [7] Wootton, R. (2008). Telemedicine support for the developing world. *Journal of Telemedicine and Telecare*, 14(3), 109-114.
- [8] Scott, R. E., & Mars, M. (2015). Telehealth in the developing world: current status and future prospects. *Smart Homecare Technology and TeleHealth*, 3, 25-37.
- [9] Zurosky, E., & Romansky, A. (2015). Mobile health clinics: Increasing access to care in central Massachusetts. *Family Medicine and Community Health*, 3(3), 43-45.
- [10] Perry, H., et al. (2014). Comprehensive review of the evidence regarding the effectiveness of community-based primary health care in improving maternal, neonatal and child health: 1. rationale, methods and database description. *Journal of Global Health*, 4(1), 010401.
- [11] Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- [12] Sze, V., et al. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12), 2295-2329.
- [13] Chen, Y. H., et al. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits*, 52(1), 127-138.
- [14] W. Al-Dhabyani, M. Goma, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.
- [15] Gupta, U., et al. (2020). Chasing Carbon: The Elusive Environmental Footprint of Computing. *arXiv preprint arXiv:2011.02839*.
- [16] Kim, K., et al. (2011). Portable ultrasound in resource-limited settings: a systematic review. *World Journal of Surgery*, 35(9), 1970-1979.
- [17] Y. Umuroglu et al., "FINN: A Framework for Fast, Scalable Binarized Neural Network Inference," Dec. 01, 2016, arXiv: arXiv:1612.07119.
- [18] M. Blott, T. Preusser, N. Fraser, G. Gambardella, K. O'Brien, and Y. Umuroglu, "FINN-R: An End-to-End Deep-Learning Framework for Fast Exploration of Quantized Neural Networks," Sep. 12, 2018, arXiv: arXiv:1809.04570.

- [19] P. Harrison, R. Hasan, and K. Park, "State-of-the-Art of Breast Cancer Diagnosis in Medical Images via Convolutional Neural Networks (CNNs)," *J Healthc Inform Res*, vol. 7, no. 4, pp. 387–432, Dec. 2023,
- [20] H. H. Maria, R. Kayalvizhi, S. Malarvizhi, R. Venkatraman, S. Patil, and A. S. Kumar, "Real-time deployment of BI-RADS breast cancer classifier using deep-learning and FPGA techniques," *J Real-Time Image Proc*, vol. 20, no. 4, p. 80, Jul. 2023, doi: 10.1007/s11554-023-01335-2.
- [21] H. S. Laxmisagar and M. C. Hanumantharaju, "FPGA implementation of breast cancer detection using SVM linear classifier," *Multimed Tools Appl*, vol. 82, no. 26, pp. 41105–41128, Nov. 2023,
- [22] K. R. H. M. H, M. S, R. Venkatraman, and S. Patil, "Hardware deployment of deep learning model for classification of breast carcinoma from digital mammogram images," *Med Biol Eng Comput*, vol. 61, no. 11, pp. 2843–2857, Nov. 2023,
- [23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.
- [24] Forrest N. Iandola and Song Han and Matthew W. Moskewicz and Khalid Ashraf and William J. Dally and Kurt Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, arXiv:1602.07360, 2016
- [25] Alessandro Pappalardo, Xilinx/brevitas, 2023, Zenodo, 10.5281/zenodo.3333552
- [26] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2017. Quantized neural networks: training neural networks with low precision weights and activations. *J. Mach. Learn. Res.* 18, 1 (January 2017), 6869–6898.
- [27] Penghang Yin and Jiancheng Lyu and Shuai Zhang and Stanley J. Osher and Yingyong Qi and Jack Xin, Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets, International Conference on Learning Representations, 2019