

## RESEARCH ARTICLE

# FPUS23: An Ultrasound Fetus Phantom Dataset With Deep Neural Network Evaluations for Fetus Orientations, Fetal Planes, and Anatomical Features

BHARATH SRINIVAS PRABAKARAN<sup>1</sup>, PAUL HAMELMANN<sup>2</sup>,  
ERIK OSTROWSKI<sup>1</sup>, (Graduate Student Member, IEEE),  
AND MUHAMMAD SHAFIQUE<sup>3</sup>, (Senior Member, IEEE)

<sup>1</sup>Institute of Computer Engineering, Technische Universität Wien (TU Wien), 1040 Vienna, Austria

<sup>2</sup>Philips Research, 5656 AE Eindhoven, The Netherlands

<sup>3</sup>Division of Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Corresponding author: Bharath Srinivas Prabakaran (bharath.prabakaran@tuwien.ac.at)

This work was supported in part by the Moore4Medical Project funded by the ECSEL Joint Undertaking under Grant H2020-ECSEL-2019-IA-876190; and in part by the Doctoral College Resilient Embedded Systems, which is run jointly by the Faculty of Informatics, TU Wien, and UAS Technikum Wien. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

**ABSTRACT** Ultrasound imaging is one of the most prominent technologies to evaluate the growth, progression, and overall health of a fetus during its gestation. However, the interpretation of the data obtained from such studies is best left to expert physicians and technicians who are trained and well-versed in analyzing such images. To improve the clinical workflow and potentially develop an at-home ultrasound-based fetal monitoring platform, we present a novel fetus phantom ultrasound dataset, FPUS23, which can be used to identify (1) the correct diagnostic planes for estimating fetal biometric values, (2) fetus orientation, (3) their anatomical features, and (4) bounding boxes of the fetus phantom anatomies at 23 weeks gestation. The entire dataset is composed of 15, 728 images, which are used to train four different Deep Neural Network models, built upon a ResNet34 backbone, for detecting aforementioned fetus features and use-cases. We have also evaluated the models trained using our FPUS23 dataset, to show that the information learned by these models can be used to substantially increase the accuracy on real-world ultrasound fetus datasets. We make the FPUS23 dataset and the pre-trained models publicly accessible at <https://github.com/bharathprabakaran/FPUS23>, which will further facilitate future research on fetal ultrasound imaging.

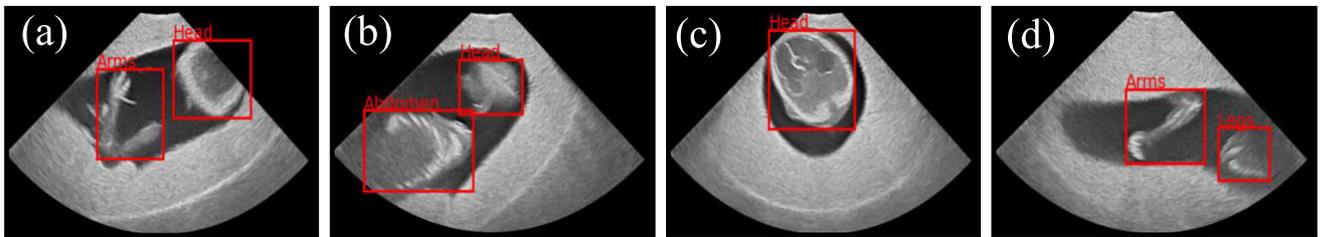
**INDEX TERMS** Fetus, phantom, ultrasound, dataset, artificial intelligence, features, deep neural networks, radiology, obstetrics.

## I. INTRODUCTION

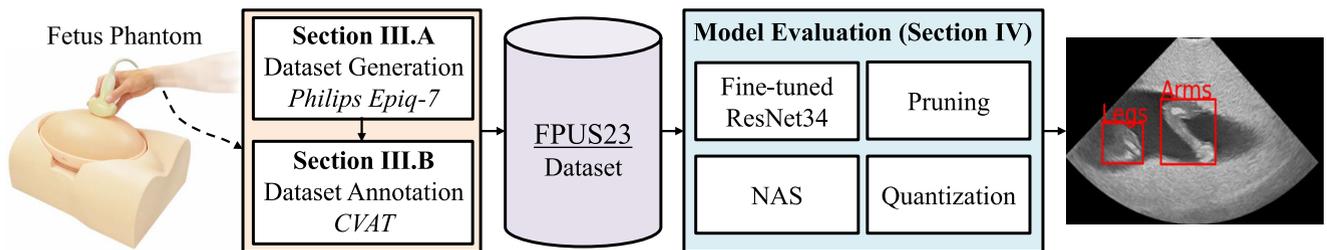
Ultrasound imaging techniques are used to create an image of organs and tissues inside the human body without the use of radiation, such as X-Rays, or expensive equipment, like Magnetic Resonance Imaging (MRI). Ultrasound technologies are used in day-to-day healthcare clinics to efficiently diagnose diseases like COVID-19 [1], [2] and detect

The associate editor coordinating the review of this manuscript and approving it for publication was Gyorgy Eigner<sup>1</sup>.

tumors [3]. Ultrasound is also widely used to monitor the development of an unborn human fetus to obtain information regarding its development and overall health. An ultrasound examination is performed at various stages of the fetus' gestation to confirm the pregnancy and determine its location, condition, size, growth, orientation, gestational age, identify potential birth defects and complications, and many other factors relevant to the healthy development and delivery of the fetus. However, the data obtained from such examinations are difficult to understand and require the expertise or



**FIGURE 1.** Identifying anatomical fetal features, such as limbs, which can enable the extraction of biometric parameters that determine the growth of the fetus. Figures (a) - (d) illustrate the ability of an object detection model in detecting various fetus anatomies, across different views, with little fine-tuning on our FPUS23 dataset.



**FIGURE 2.** Overview of our methodology for generating and annotating the FPUS23 dataset.

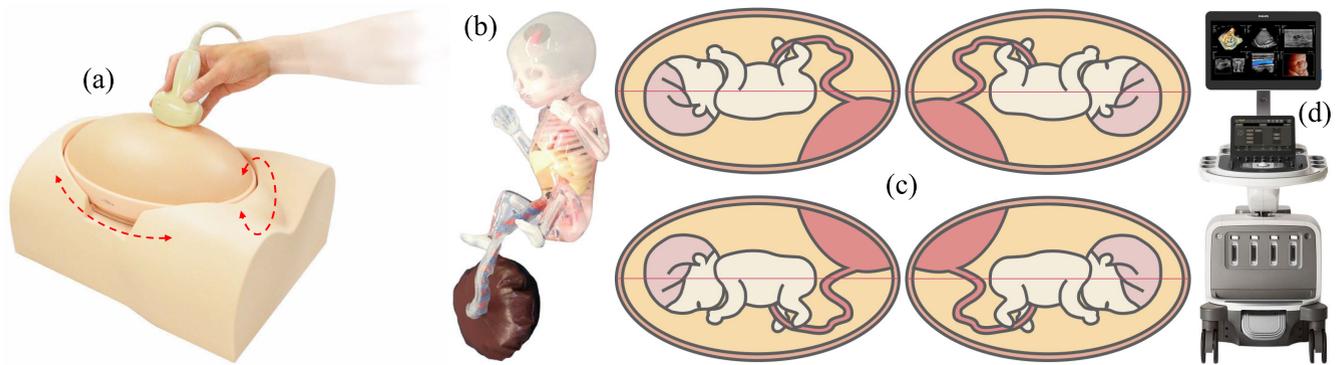
training of sonographers or physicians to accurately interpret the data. For instance, as depicted in an ultrasound image of a fetus at 23 weeks gestation (fig. 1), identifying a fetus can be quite easy. However, identifying the orientation of the fetus and evaluating key biometric parameters, like the abdominal circumference or femur length, which are used to ascertain the gestational age of the fetus, requires the level of expertise that is currently offered only by trained sonographers and physicians. This “*experience*” can be learned and embedded within the deep learning models, which can be deployed in clinical use-cases as assistants to aid healthcare professionals in data interpretation. These models can also be used to build an *at-home portable ultrasound-based fetal monitoring platform* that can enable the user to understand the data by collating and interpreting the vital information.

However, the development of such a deep learning model for analyzing fetal ultrasound data requires investigation of the following key research challenges: (1) ultrasound fetus data fall under the category of healthcare information that is heavily protected by regulatory requirements regarding their generation, storage, usage, etc. to ensure patient privacy; (2) due to these regulations, there are very few openly accessible fetal ultrasound datasets, which can be used to develop such clinical assistants and at-home monitoring platforms; (3) even the state-of-the-art datasets that are accessible are not properly annotated with the relevant information in order to be able to train DNN models, which can be used to infer relevant fetus anatomy information; and (4) the existing datasets are not large enough to enable the DNN model to learn the required features efficiently, despite the use of existing transfer learning approaches.

To address these research challenges, we build the FPUS23 dataset (see fig. 2) by: (1) using a fetus phantom at 23 weeks gestation, instead of actual human fetuses, thereby circumventing the regulations associated with healthcare data; (2) not generating or using any healthcare data in our dataset – FPUS23 will be openly accessible to further facilitate future research and advancements in this domain; (3) properly labeling and annotating the dataset with the help of scientists with experience in fetal ultrasound imaging to generate a dataset that can be used to identify (i) diagnostic planes for extracting fetal biometric parameters, (ii) fetus orientation, (iii) fetus anatomies, and (iv) bounding boxes of the fetus anatomies; (4) building a dataset with 15,728 ultrasound image samples that can be used to learn the required information; (5) extensive evaluation of our datasets with appropriate transfer learning approaches, including model compression techniques, as illustrated in Section IV.

## II. RELATED WORK

There is an abundance of ultrasound datasets for various use-cases, which can be used to generate DNN-based models for classification and segmentation. For instance, the breast ultrasound image dataset presented by Al-Dhabyani et al. [6], which is composed of normal, benign, and malignant images that can be used to train to a model to act as a classifier. Similarly, the POCUS dataset, presented by Born et al. [1], and the COVIDX-US dataset, by Ebadi et al. [2], are openly accessible for building DNN-based clinical assistants that can aid in the analytics and diagnosis of COVID-19. Leclerc et al. [7] presented a cardiac ultrasound electrocardiography dataset containing image



**FIGURE 3.** Overview of (a) the phantom abdomen in the mother body torso and its possible rotations; (b) the fetus phantom placed in the abdomen; (c) the four possible fetus orientations; (d) the Philips Epiq-7 ultrasound system (adapted from [4] and [5]).

sequences with two and four-chamber views of the heart of 500 patients. Likewise, there are a wide number of ultrasound datasets for diagnosing and analyzing several internal body organs.

#### A. FETAL ULTRASOUND

Deep learning has also been explored for fetal ultrasound imaging, albeit not as widely or comprehensively. Reference [8] proposed a multi-scale self-attention generator that can be used to automatically generate ultrasound images from various segmentation masks, which can then be used for fetal brain segmentation and analysis. Reference [9] proposed the use of deep learning to automatically analyze the fetal heart by encoding and translating Spatio-temporal information in order to classify amongst three different fetal heart planes. Reference [10] proposed a CNN-based classifier that can be used to detect cardiac abnormalities in fetal ultrasound images. Reference [11] have proposed the use of CNNs to detect the six standard fetal brain planes on a proprietary dataset containing 30,000 2D fetal ultrasound images gathered between 16 and 34 weeks gestation, to achieve 91% accuracy. Reference [12] presented two DNN models that are used to detect the ideal frame for the fetal head, followed by its segmentation, which is used to measure the head circumference, a key biometric parameter. Reference [13] proposed an improved multi-task learning network that improves the segmentation capabilities of the model when compared to [12]. Although ultrasound examinations of a fetus are very common, there are very few openly accessible datasets for researchers to build DNN-based models that can aid in the analytics of fetal ultrasound images. Reference [14] presented a dataset of fetal ultrasound images with annotations regarding the head circumference, as shown in [12] and [13]. Reference [15] presented a fetal ultrasound dataset with over 12,400 images from 1,792 patients, which were categorized into six classes containing the anatomical planes. Most of the other works in this category primarily work on proprietary datasets, which are not accessible for analysis and evaluation.

### III. FPUS23: THE FETAL ULTRASOUND DATASET

#### A. DATA COLLECTION

The required ultrasound fetal data was generated and collected using the “US-7 SPACE FAN-ST” fetus phantom [5], which has been typically used to train sonographers to assess the development and condition of the fetus. Figs. 3(a) and 3(b) illustrate an overview of the oval-shaped phantom abdomen, which mimics the uterus containing a fetus at 23 weeks gestation, and the life-size fetus demonstration model that is placed inside the phantom abdomen, respectively. We propose to use a 23-week old phantom as a mid-pregnancy scan is typically performed around this time to check for fetus anomalies. The phantom abdomen can be rotated into four different positions to change the orientation and presentation (cephalic or breech) of the fetus phantom (see fig. 3(c)). The fetus model includes full skeletal structure and key organic features that can be observed and used to train the sonographer to assess the fetus’ anatomy (like head, arms legs, abdomen) and internal body organs (like brain, skull, spine, cardiac chambers, stomach, kidney, blood vasculature, etc.). The biometric parameters of the fetus can also be measured/learned using an ultrasound of the fetus phantom at the appropriate positions or the correct *diagnostic planes*. Besides the aforementioned features, the quantity of amniotic fluid, any potential abnormalities, location of the placenta, fetal posture, etc. can also be learned with the help of this model.

We use the X6-1 xMATRIX array transducer [16], which is interfaced with the Philips Epiq-7 system [4] to collect and process the data to generate the final ultrasound image (see fig. 3(d)). The Anatomically Intelligent Ultrasound (AIUS) imaging technology deploys advanced organ modeling and imaging techniques to generate a two-dimensional image of the fetus phantom using the default settings for the “OB Fetal Echo” imaging option. The imaging depth was set to 12cm and captured at a 23Hz frame rate. Sufficient ultrasound gel is applied on the phantom abdomen to ensure acoustic coupling with the probe, thereby reducing acoustic impedance, and enabling clear imaging. We executed two protocols to collect the images used in the *FPUS23* dataset:

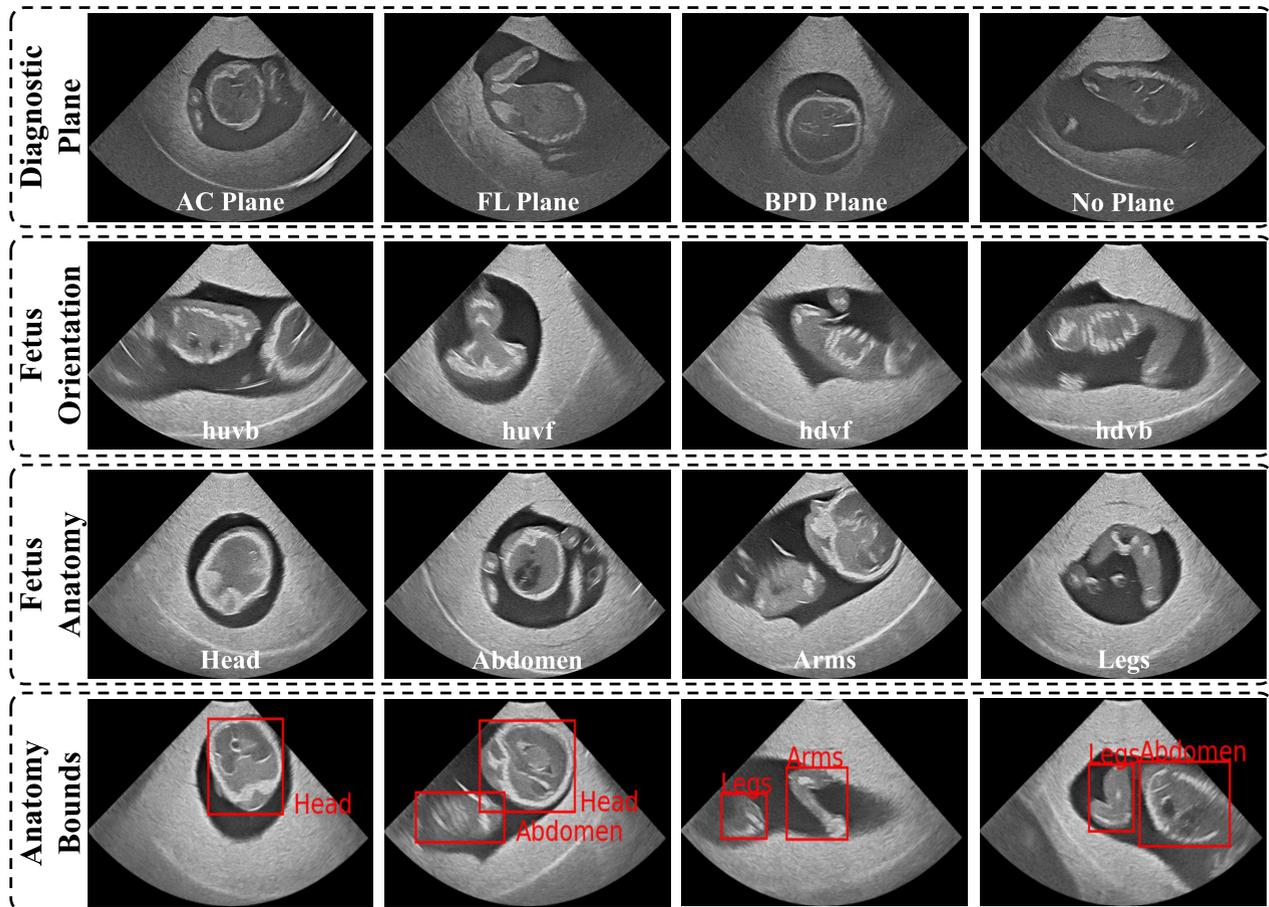


FIGURE 4. Super-labels of the FPUS23 dataset and corresponding samples in each class.

- (1) **Protocol-I:** The probe is placed on the phantom abdomen surface and navigated to the *correct diagnostic planes* that can be used for the measurement of the three primary biometric parameters of the fetus, namely the transverse plane, which is used to obtain the brain's Biparietal Diameter (BPD), abdominal standard plane, which is used to estimate the Abdominal Circumference (AC), and the femur standard plane, which is used to estimate the fetus' Femur Length (FL). The correct diagnostic planes were identified using the clinical protocols discussed by Salomon et al. [17] and Bethune et al. [18]. To further enrich the dataset, after the acquisition of several frames at the correct diagnostic plane, we tilt, rotate, or traverse the ultrasound probe in random directions to collect more information.
- (2) **Protocol-II:** The focus of this protocol is to obtain images capturing the anatomies of the fetus phantom in the generated images. We do this by navigating the probe to obtain the head, abdomen, arms, and legs, individually or combined, in the picture and move the probe in different directions to obtain a heterogeneous set of images capturing the fetal anatomies. Furthermore, the phantom abdomen was also rotated

and placed in the four possible orientations [head up (hu) or down (hd), view front (vf) or back (vb)], when collecting the ultrasound data, to potentially mimic the real-life behavior of fetus orientation and presentation (see fig. 3(c)). Additionally, the probe orientation was also changed between horizontal and vertical, with respect to the abdomen, when the data was collected to enhance the dataset with more information.

#### B. ANNOTATION

The data streams obtained by the Philips Epiq-7 ultrasound system are converted to PNG image sequences, of dimension  $664 \times 388$ , using custom in-house software for easier labeling, annotating, and processing. The stored PNG files are annotated using a customized version of the Computer Vision Annotation Tool – CVAT [19], which was primarily opted for its ease-of-use and wide-range features. Each acquired ultrasound frame was subsequently annotated by scientists with experience in fetal ultrasound imaging.

The sequences obtained using *Protocol-I* are labeled as a correct diagnostic plane for one of the three biometric parameters (BPD, AC, FL) or as a non-diagnostic plane. Since the number of data samples obtained for each of the

**TABLE 1.** Breakdown of the number of labeled data samples in each class of our FPUS23 dataset; Of the 15, 728 images, not all contain relevant information for labeling – images obtained during probe movements may not accurately depict anatomies.

Diagnostic Plane	Total			
	5, 265			
	AC Plane	BPD Plane	FL Plane	No. Plane
	1, 386	1, 280	1, 281	1, 318
Fetus Orientation	Total			
	15, 728			
	hdvb	hdvf	huvb	huvf
	3, 757	3, 235	3, 980	4, 756
Fetus Anatomy	Total			
	9, 317			
	Head	Arms	Legs	Abdomen
	3, 003	1, 629	2, 159	2, 526
Anatomy Bounds	Total			
	9, 455			
	Head	Arms	Legs	Abdomen
	4, 370	4, 853	4, 572	6, 435

diagnostic planes is quite smaller than the non-diagnostic plane output class, the samples were augmented in each of the other three output classes to ensure equal representation of data across all classes. The data obtained using *Protocol-II* is, first, labeled with fetus orientations, namely huvf, huvb, hdvf, or hdvb, as discussed earlier, based on the position of the phantom abdomen when the scans are made. Next, the images are tagged with the anatomies present in the image, such as heads, arms, legs, and abdomen. The images are subsequently exhaustively annotated with boxes representing the respective anatomies to determine their bounds and potentially estimate biometric parameters later, such as femur length, at the correct diagnostic plane. Images that do not contain any vital and/or relevant information regarding the fetus are not labeled or annotated. These finalized labels and annotations, for each valid image in the dataset, are extracted as an XML file, which can be used to train various deep learning models as required.

Table 1 depicts an overview of our FPUS23 dataset and the number of input samples present in each class for each of the four different super-labels: (1) Diagnostic Plane, (2) Fetus Orientation, (3) Fetus Anatomy, and (4) Anatomy Bounds, using box annotation. The dataset is split in the ratio of 8 : 1 : 1, with respect to training, validation, and testing, respectively, for the first three cases. We split the anatomy bounds data and use 80% of it for training and 20% for validating the model. Fig. 4 depicts a few sample images and their corresponding annotations for the four different super-labels of our FPUS23 dataset.

## IV. EXPERIMENTS AND RESULTS

### A. EXPERIMENTAL SETUP

The experimental evaluations illustrated in this section, which depict the efficacy of the DNN models trained using our FPUS23 dataset, are primarily completed on a CentOS 7.9

**TABLE 2.** Preliminary quality evaluations (accuracy and F1-score) and hardware requirements (Flops and memory) of the modified ResNet34 model trained using the FPUS23 dataset.

	Accuracy	F1-Score	No. of Flops	Memory (MB)
Diagnostic Plane	98.29%	.9833		
Fetus Orientation	99.80%	.9936	3.67G	157.57
Fetus Anatomy	99.46%	.9989		
Anatomy Bounds	mAP	mAR	64.4G	357.20
	98.60%	84.40%		

Operating System running on an Intel Core i7-8700 CPU with 16GB RAM and 2 Nvidia GeForce GTX 1080 Ti GPUs. Our scripts were executed with the following software versions: CUDA 11.5, Pytorch 3.7.4.3, torchvision 0.11.1, and Pytorch-lightning 1.5.1. We use a ResNet34 [20] DNN model, pre-trained using the ImageNet [21], and retrain it with our dataset for 15 epochs using the cross-entropy loss function. The initial and final layers of the ResNet34 architecture were adapted to accommodate the custom input data dimensions and output classes, respectively. We use a modified Faster-RCNN [22] with our ResNet34 backbone to build the model used for determining the anatomy bounds of the fetus in our dataset. The learning rate for all models was set to 0.001 using the Adam Optimizer with a step size of 20 and  $\gamma = 0.1$ .

We use the traditional metrics of accuracy, precision, and recall to determine the efficacy of the deep learning models illustrated in this section. Accuracy is the ratio of the total number of inputs accurately predicted with respect to the total number of predictions, which is a primary evaluation metric for classification systems. For the anatomy bounds, we use the mean Average Precision (*mAP*; IoU = [.50:.05:.95]), which is the ability of the model to not label a negative sample as positive, and mean Average Recall (*mAR*; IoU = [.50:.05:.95]), which denotes the ability of the model to identify all positive instances of each class:

$$mAP = \frac{1}{N} \frac{\sum_{i=1}^N p_{i,i}}{\sum_{i=0}^N \sum_{j=1}^N p_{i,i} + p_{i,j}} \quad (1)$$

$$mAR = \frac{1}{N} \frac{\sum_{i=1}^N p_{i,i}}{\sum_{i=0}^N \sum_{j=1}^N p_{i,i} + p_{j,i}}, \quad i, j = 1, \dots, N \quad (2)$$

where  $N$  denotes the total number of output classes,  $p_{i,i}$  the number of pixels classified as class  $i$  and labeled as class  $i$ , and  $p_{i,j}$ ,  $p_{j,i}$  are the number of pixels classified as class  $i$  and labeled as class  $j$  and vice-versa. We also evaluate the model's F1-score, which is the harmonic mean of the model's precision and recall.

### B. BASELINE MODEL

To illustrate the effectiveness of the FPUS23 dataset, we retrain a modified version of the ResNet34 architecture, which we consider as the baseline, to illustrate the capability

TABLE 3. Exhaustive evaluations of the compressed baseline ResNet34 model.

	Pruning							
	0%		30%		50%		70%	
	Quantization							
	FP32	INT8	FP32	INT8	FP32	INT8	FP32	INT8
No. of Flops	3.67G	-	2.58G	-	1.85G	-	1.12G	-
Memory (MB)	157.57	39.39	125.82	31.45	104.65	26.16	83.48	20.87
	Diagnostic Planes							
Accuracy	98.29%	97.34%	98.10%	98.29%	98.10%	97.91%	97.72%	<b>99.25%</b>
F1-Score	.9833	.9759	.9818	.9832	.9813	.9799	.9799	<b>.9834</b>
	Fetus Orientation							
Accuracy	99.80%	99.73%	99.49%	<b>99.93%</b>	99.55%	99.87%	99.68%	<b>99.93%</b>
F1-Score	.9936	.9992	.9949	<b>.9994</b>	.9956	.9988	.9968	<b>.9994</b>
	Fetus Anatomy							
Accuracy	99.46%	99.46%	99.57%	<b>99.89%</b>	99.78%	<b>99.89%</b>	99.67%	99.78%
F1-Score	<b>.9989</b>	.9749	.9952	.9987	.9978	.9987	.9968	.9978
	Anatomy Bounds - Faster-RCNN with ResNet34 backbone							
mAP	98.60%	98.60%	98.00%	98.00%	<b>98.80%</b>	97.80%	98.40%	98.00%
mAR	84.40%	81.20%	<b>87.90%</b>	78.50%	85.90%	78.60%	86.00%	79.80%
No. of Flops	64.40G	-	45.09G	-	32.21G	-	19.33G	-
Memory (MB)	357.20	239.02	250.04	155.67	178.60	100.11	107.16	44.55

of the network to learn relevant information regarding the classification of labels and detecting fetal anatomies. Accuracy and F1-score are the two metrics used to determine the quality of the model, whereas the number of floating-point operations (Flops) and memory (MB) are relevant to estimate the hardware and resource requirements of the baseline model and determine its deployability in edge devices. The results of these experiments are illustrated in Table 2. Achieving  $\sim 99\%$  accuracy in the classification of diagnostic planes, fetus orientation, and fetus anatomy gives us the understanding that the models are able to learn the features quite well. Similarly, the modified Faster-RCNN model, embedded with our ResNet34 backbone, is able to detect fetal anatomies at significantly high precision. Note, the significantly high quality of the models can also be attributed to their over-parameterization, which implies that smaller networks, achieving similar output accuracy, can be obtained with Neural Architecture Search (NAS) [23] and model compression techniques [24] (see Section IV-C).

Moreover, the use of an inverted probe during an ultrasound exam leads to the generation of an inverted image. To design a robust network model that can extract features and information from the inverted image to ensure correct classification, relevant image samples need to be collected, with the probe inverted, and included in the dataset during the training stage. However, this information can also be added to the model by flipping the collected images along the y-axis, which mimics probe-inverted images, and adding them to the original dataset before training.

### C. MODEL COMPRESSION

To further reduce the model's hardware requirements, we use compression techniques like pruning and quantization. We have implemented the technique proposed by Han et al. [25], which proposes to eliminate the smallest  $x\%$  of total weights and associated connections from the network, followed by a network retraining stage, wherein the model relearns the information on the reduced set of available parameters, to potentially achieve similar output quality as the original model. We analyze the quality and hardware requirements of the models that are 30%, 50%, and 70% pruned. The pruned networks are subsequently quantized to 8-bit integer (INT8) precision, using the quantization-aware training strategy presented by Khudia et al. [26], to further reduce the model's size and improve its computational performance; INT8 computations are several orders of magnitude faster than 32-bit floating-point (FP32) operations [27]. Similar to the regularizing effect illustrated in [28], compression of the baseline models led to potential scenarios where the compressed models outperform the original. This regularizing effect is especially prominent when the ResNet34 classifiers are compressed, due to their heavy over-parameterization. Table 3 illustrates the quality evaluations and the hardware requirements of these models when trained on FPUS23.

Plenty of research works on automated NAS have demonstrated their efficacy in reducing the number of parameters and computations for achieving similar quality results compared to over-parameterized architectures that are typically designed by hand. Towards this, we first investigate the

**TABLE 4.** Evaluations of the DNN model obtained through exhaustive neural architecture search; As expected, a smaller model achieves the same or similar output quality as the baseline while reducing the hardware requirements by up to 17×.

	Pruning							
	0%		30%		50%		70%	
	FP32	INT8	FP32	Quantization INT8	FP32	INT8	FP32	INT8
No. of Flops	1.42G	-	1.00G	-	0.73G	-	0.45G	-
Memory (MB)	41.26	10.31	37.03	9.25	34.21	8.55	31.38	7.84
Diagnostic Planes								
Accuracy	99.24%	98.29%	<b>99.62%</b>	98.67%	99.24%	99.43%	99.24%	97.72%
F1-Score	.9930	.9821	<b>.9965</b>	.9877	.9930	.9947	.9930	.9789
Fetus Orientation								
Accuracy	99.57%	99.61%	<b>99.83%</b>	98.60%	99.53%	99.80%	99.70%	99.80%
F1-Score	.9980	.9962	.9976	.9861	.9935	<b>.9982</b>	.9957	<b>.9982</b>
Fetus Anatomy								
Accuracy	<b>99.89%</b>	99.78%	<b>99.89%</b>	<b>99.89%</b>	<b>99.89%</b>	99.67%	99.78%	99.46%
F1-Score	<b>.9987</b>	.9975	<b>.9987</b>	<b>.9987</b>	<b>.9987</b>	.9969	.9975	.9934
Anatomy Bounds - Faster-RCNN with ResNet10 backbone								
mAP	<b>97.90%</b>	<b>97.90%</b>	97.30%	97.50%	97.20%	97.60%	97.20%	97.30%
mAR	<b>81.30%</b>	78.90%	75.80%	75.30%	76.20%	75.70%	76.90%	75.00%
No. of Flops	28.18G	-	19.74G	-	14.11G	-	8.46G	-
Memory (MB)	290.84	194.61	203.59	126.74	145.42	81.51	87.25	36.27

effectiveness of two state-of-the-art automated NAS approaches presented by Fang et al. [29] and Wang et al. [30] in reducing the number of Flops and memory while retaining the output quality. Both these approaches yield a residual DNN with 10 intermediate feature extraction layers to achieve an output quality similar to that of the baseline. However, while exhaustively generating and exploring smaller networks from scratch, we generated a residual DNN with 8 layers, instead of the 10 proposed by [29] and [30], which offers similar output quality while requiring a fewer number of parameters. We use this ResNet8 architecture to build models for classifying the diagnostic planes, fetus orientation, and fetus anatomy. However, we have observed that the quality of the Faster-RCNN model, built using a ResNet8, is substantially lower as opposed to the model using a ResNet10 backbone. Therefore, the Faster-RCNN model is built using a ResNet10 backbone instead. From the results, it is quite evident that the ResNet8 model achieves a similar output quality as the original baseline models, even when both of them are compressed using pruning and/or quantization. For instance, a fetus anatomy classifier built using the ResNet8 architecture, which has been 30% pruned and INT8 quantized, achieves an output quality greater than the baseline ResNet34 model, while requiring less than 10MB in memory, making it ideal for deployment on resource-constrained processing platforms. The Faster-RCNN built using the ResNet10 backbone also achieves similar quality to the baseline model while substantially reducing the hardware requirements. Table 4 illustrates the quality evaluations and

the hardware requirements of the NAS models when trained on the *FPUS23* dataset.

#### D. EVALUATION ON STATE-OF-THE-ART REAL-WORLD DATASET

To further demonstrate the applicability of our dataset in real-world fetal ultrasound use-cases, we fine-tune the models trained using *FPUS23* on the training set and evaluate them on the test set presented by [15]. We train two models: first, we train the baseline model on the *FPUS23* dataset, before retraining on the real-world dataset, and in the second, we train the model directly on the real-world dataset. Model-1 achieves 91.92% accuracy in detecting the anatomical planes after training for just 1 epoch, whereas Model-2 achieves the same accuracy only after training for more than 16 epochs. Therefore, the models can learn relevant features regarding fetal ultrasounds from the *FPUS23* dataset before being deployed for other fetal ultrasound datasets with very little fine-tuning. To re-emphasize this, we perform a small analysis on the networks under consideration. We consider the weights of the *FPUS23*-trained model before and after fine-tuning on the real-world dataset. Each of the weights in these two sets are subtracted, squared, and aggregated together (like sum of squared errors) to obtain a single value, which denotes the amount of fine-tuning undergone by the model. We do the same for the ImageNet-trained model and compare the two. The *FPUS23*-trained model aggregates a value of  $2.01 \times 10^{-7}$ , whereas the ImageNet-trained model aggregates to  $3.02 \times 10^{-6}$ , which is more than 15× larger.

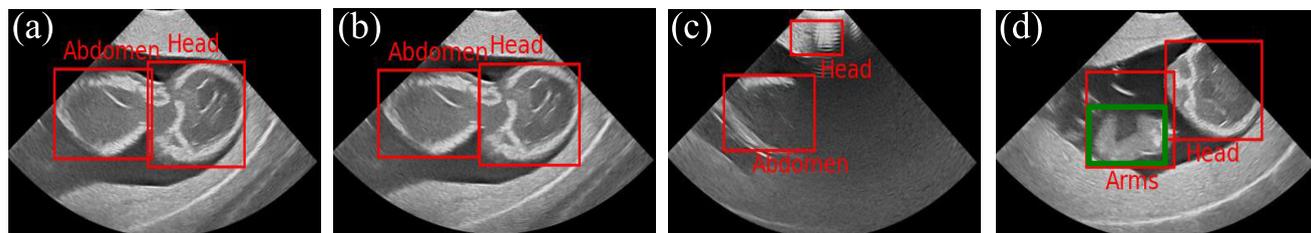


FIGURE 5. Evaluation of Faster-RCNN trained on *FPUS23*; comparison between (a) ground truth and (b) prediction; (c) detecting anatomies incorrectly during probe navigation; (d) misestimating dimensions of fetus anatomy.

TABLE 5. Evaluation of the models trained using ImageNet and *FPUS23* on the real-world fetal ultrasound dataset [15] without any fine-tuning.

Metric	ImageNet-trained	FPUS23-trained
Number of Errors	562	319
Total Number of Images	656	656
Accuracy Percentage	14.33	51.37

TABLE 6. Class-wise prediction breakdown of the ImageNet-trained model when fine-tuned for the real-world dataset [15] after 5 epochs.

		Ground Truth			
		Abdomen	Brain	Femur	Thorax
Prediction	Abdomen	52	11	12	1
	Brain	0	72	2	0
	Femur	7	2	134	0
	Thorax	13	9	5	336
Accuracy		72.22	76.60	87.58	99.70

This implies that the former model requires less fine-tuning in comparison to the latter. Section IV-E presents a comprehensive analysis of the knowledge retained and transferred by models trained on the *FPUS23* dataset using ablation studies, followed by a discussion of the anatomy detection results for the trained DNN model.

### E. ABLATION STUDIES

With no fine-tuning on the state-of-the-art dataset [15], the *FPUS23*-trained model is able to make predictions with relatively higher accuracy when compared to an ImageNet-trained model, as shown by the results in Table 5.

Next, we provide a class-wise prediction breakdown of the models when fine-tuned on the state-of-the-art dataset. After 15 epochs, the ImageNet-trained model converges to the same accuracy as the *FPUS23*-trained model, which requires just 1 epoch for fine-tuning. The results of these experiments are presented in Tables 6 and 7; the ImageNet-trained model is fine-tuned for 5 epochs and converges to the same accuracy-level as the *FPUS23*-trained model after 15 epochs.

We have also trained the two models using a 40% subset of the real-world training data to achieve the same accuracy as the baseline (92%) when using the *FPUS23*-trained model. Whereas the ImageNet trained model is unable to achieve the same accuracy and falls short at 88% accuracy.

Fig. 5 provides a sample of the anatomy detection results of the Faster-RCNN model with the ResNet34 backbone trained

TABLE 7. Class-wise prediction breakdown of the *FPUS23*-trained model when fine-tuned for the real-world dataset [15] after just 1 epoch.

		Ground Truth			
		Abdomen	Brain	Femur	Thorax
Prediction	Abdomen	39	0	5	0
	Brain	3	91	7	6
	Femur	21	1	137	0
	Thorax	9	2	4	331
Accuracy		54.17	96.81	89.54	98.22

using the *FPUS23* dataset. As illustrated by the ground truth and prediction, in figs. 5(a) and 5(b), respectively, the precision achieved by the model is quite high, and can detect anatomies quite accurately and precisely in most instances, as illustrated by the overall mAP and mAR values of the model. However, the model also misclassifies certain abstract artifacts during probe navigation as fetus anatomies, as illustrated by fig. 5(c). Fig. 5(d) illustrates an instance wherein the model over- or under-estimates the bounds of the fetus anatomy.

### V. CONCLUSION AND FUTURE WORK

In this paper, we present the *FPUS23*, which is an ultrasound dataset of a fetus phantom at 23 weeks gestation. The data streams are collected and annotated by scientists with relevant fetal ultrasound experience to obtain information regarding the (1) diagnostic plane, (2) fetus orientation, (3) fetus anatomy, and (4) their bounds, using box annotations. The generated dataset is used to train a variety of deep learning models to illustrate the model’s ability to extract vital information, which can be used to accurately distinguish among the classes in different categories and detect the fetus anatomy bounds. Furthermore, to evaluate their deployability in portable resource-constrained devices, we evaluated the capability of a smaller DNN compressed using pruning and quantization to illustrate that smaller DNNs are equally competent at extracting relevant information from the dataset and are capable of execution on resource-constrained devices and embedded platforms. The *FPUS23* dataset is open-source and the trained models are accessible online. In our future work, we plan to include annotated data of fetus phantoms at different gestation durations to offer a more comprehensive fetal ultrasound dataset.

## REFERENCES

- [1] J. Born, G. Brändle, M. Cossio, M. Disdier, J. Goulet, J. Roulin, and N. Wiedemann, "POCOVID-Net: Automatic detection of COVID-19 from a new lung ultrasound imaging dataset (POCUS)," 2020, *arXiv:2004.12084*.
- [2] A. Ebadi, P. Xi, A. MacLean, S. Tremblay, S. Kohli, and A. Wong, "COVIDx-U.S.—An open-access benchmark dataset of ultrasound imaging data for AI-driven COVID-19 analytics," 2021, *arXiv:2103.10003*.
- [3] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, and T. Wang, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, Jun. 2016.
- [4] Philips. *EPIQ Elite*. Accessed: May 10, 2023. [Online]. Available: <https://www.usa.philips.com/healthcare/product/HC795098/epiq-elite-a-new-class-of-premium-ultrasound-has-arrived>
- [5] K. Kagaku. (2020). *Fetus Ultrasound Examination Phantom, SPACE FAN-ST*. [Online]. Available: [https://www.kyotokagaku.com/en/products\\_data/us-7\\_en/](https://www.kyotokagaku.com/en/products_data/us-7_en/)
- [6] W. Al-Dhabyani, M. Goma, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.
- [7] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2198–2210, Sep. 2019.
- [8] J. M. J. Valanarasu, R. Yasarla, P. Wang, I. Hacihaliloglu, and V. M. Patel, "Learning to segment brain anatomy from 2D ultrasound with less data," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1221–1234, Oct. 2020.
- [9] A. Patra, W. Huang, and J. A. Noble, "Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video," in *3rd Int. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support Workshop, 7th Int. Workshop, Conjunct. (DLMIA/ML-CDS/MICCAI)*. Québec City, QC, Canada: Springer, Sep. 2017, pp. 276–284.
- [10] M. Komatsu, A. Sakai, R. Komatsu, R. Matsuoka, S. Yasutomi, K. Shozu, A. Dozen, H. Machino, H. Hidaka, T. Arakaki, K. Asada, S. Kaneko, A. Sekizawa, and R. Hamamoto, "Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning," *Appl. Sci.*, vol. 11, no. 1, p. 371, Jan. 2021.
- [11] R. Qu, G. Xu, C. Ding, W. Jia, and M. Sun, "Deep learning-based methodology for recognition of fetal brain standard scan planes in 2D ultrasound images," *IEEE Access*, vol. 8, pp. 44443–44451, 2020.
- [12] T. L. A. van den Heuvel, H. Petros, S. Santini, C. L. de Korte, and B. van Ginneken, "Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries," *Ultrasound Med. Biol.*, vol. 45, no. 3, pp. 773–785, Mar. 2019.
- [13] Z. Sobhaninia, S. Rafiei, A. Emami, N. Karimi, K. Najarian, S. Samavi, and S. M. R. Soroushmehr, "Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 6545–6548.
- [14] T. L. A. van den Heuvel, D. de Bruijn, C. L. de Korte, and B. V. Ginneken, "Automated measurement of fetal head circumference using 2D ultrasound images," *PLoS ONE*, vol. 13, no. 8, Aug. 2018, Art. no. e0200412.
- [15] X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Jun. 2020.
- [16] Philips. *X6-1: xMATRIX Array Transducer With Purewave Crystal Technology*. Accessed: May 10, 2023. [Online]. Available: <https://www.usa.philips.com/healthcare/product/HC989605409281/x6-1>
- [17] L. J. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. Johnsen, K. Kalache, K.-Y. Leung, G. Malingier, H. Munoz, F. Prefumo, A. Toi, and W. Lee, "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan," *Ultrasound Obstetrics Gynecol.*, vol. 37, no. 1, pp. 116–126, Jan. 2011.
- [18] M. Bethune, E. Alibrahim, B. Davies, and E. Yong, "A pictorial guide for the second trimester ultrasound," *Australas. J. Ultrasound Med.*, vol. 16, no. 3, pp. 98–113, Aug. 2013.
- [19] B. Sekachev et al., "OpenCV/CVAT: V1.1.0," Tech. Rep., Aug. 2020. [Online]. Available: <https://zenodo.org/record/4009388/export/hx#X6rwxmgzZPY>, doi: 10.5281/zenodo.4009388.
- [20] 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.
- [21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.
- [23] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.
- [24] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," 2017, *arXiv:1710.09282*.
- [25] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.
- [26] D. Khudja, J. Huang, P. Basu, S. Deng, H. Liu, J. Park, and M. Smelyanskiy, "FBGEMM: Enabling high-performance low-precision deep learning inference," 2021, *arXiv:2101.05615*.
- [27] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.
- [28] A. Marchisio, M. A. Hanif, M. Martina, and M. Shafique, "PruNet: Class-blind pruning method for deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.
- [29] J. Fang, Y. Sun, Q. Zhang, Y. Li, W. Liu, and X. Wang, "Densely connected search space for more flexible neural architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10625–10634.
- [30] T. Wang, K. Wang, H. Cai, J. Lin, Z. Liu, H. Wang, Y. Lin, and S. Han, "APQ: Joint search for network architecture, pruning and quantization policy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2075–2084.



**BHARATH SRINIVAS PRABAKARAN** received the B.Eng. degree in electrical and electronics and the M.Sc. degree in biological sciences from the Birla Institute of Technology and Science (BITS), Pilani, India, in 2017. He is currently pursuing the Ph.D. degree with the Doctoral College Resilient Embedded Systems, Faculty of Informatics, TU Wien, Austria, under the supervision of Prof. Muhammad Shafique. In 2016, he was a Visiting Researcher with TU Dresden, Germany, for one year, where he completed his master's thesis. Since 2017, he has been a Project Assistant with TU Wien, for two years before starting work on his doctoral dissertation. His research interests include medical imaging, embedded machine learning, wearable architectures, healthcare systems, and energy-efficient technologies.



**PAUL HAMELMANN** received the M.Sc. degree in biomedical engineering from the University of Twente, Enschede, and the Ph.D. degree from the Biomedical Diagnostics Laboratory (BM/d), Electrical Engineering Faculty, Eindhoven University of Technology (TU/e), in 2020. He is currently a Data Scientist with Philips Research, Eindhoven. His current research interests include biomedical signal processing and (clinical) data science in the field of fetal and pregnancy monitoring.



**ERIK OSTROWSKI** (Graduate Student Member, IEEE) received the bachelor's and M.Sc. degrees in business mathematics from Universität zu Köln, Cologne, Germany, in 2020. He is currently pursuing the Ph.D. degree with the School of Informatics, TU Wien, Austria. His research interests include computer vision, semantic segmentation, weak supervision, healthcare systems, energy-efficient technologies, and AI and machine learning.



**MUHAMMAD SHAFIQUE** (Senior Member, IEEE) received the Ph.D. degree in computer science from the Karlsruhe Institute of Technology (KIT), Germany, in 2011.

He established and led a highly recognized research group at KIT for several years and conducted impactful collaborative research and development activities across the globe. In October 2016, he joined the Institute of Computer Engineering, Faculty of Informatics, Technische Universität Wien (TU Wien), Vienna, Austria, as a Full Professor of

computer architecture and robust, and energy-efficient technologies. Since September 2020, he has been with New York University (NYU), where he is currently a Full Professor and the Director of the eBrain Laboratory, NYU, Abu Dhabi, United Arab Emirates, and a Global Network Professor with the Tandon School of Engineering, NYU, New York City, USA. He is also a Co-PI/Investigator in multiple NYUAD Centers, including the Center of Artificial Intelligence and Robotics (CAIR), the Center of Cyber Security (CCS), Center for Interacting Urban Networks (CITIES), and the Center for Quantum and Topological Systems (CQTS). He holds one U.S. patent. He has (co)authored six books, more than ten book chapters, more than 350 papers in premier journals and conferences, and more than 100 archive articles. His research interests include AI and machine learning hardware and system-level design, brain-inspired computing, machine learning security and privacy, quantum machine learning, cognitive autonomous systems, wearable healthcare, energy-efficient systems, robust computing, hardware security, emerging technologies, FPGAs, MPSoCs, embedded systems, cross-layer analysis, modeling, design, and optimization of computing and memory systems. The researched technologies and tools are deployed in application use cases from the Internet of Things (IoT), smart cyber-physical systems (CPS), and ICT for development (ICT4D) domains. He is a Senior Member of the IEEE Signal Processing Society (SPS) and a member of ACM, SIGARCH, SIGDA, SIGBED, and HIPEAC. He received the 2015 ACM/SIGDA Outstanding New Faculty Award, the AI 2000 Chip Technology Most Influential Scholar Award, in 2020 and 2022, and the ASPIRE AARE Research Excellence Award, in 2021. He received six gold medals and several best paper awards and nominations at prestigious conferences. He has served as the PC chair, the general chair, the track chair, and a PC member for several prestigious IEEE/ACM conferences. He has given several keynotes, invited talks, and tutorials as well as organized many special sessions at premier venues.

• • •