AiEEG: Personalized Seizure Prediction Through Partially-Reconfigurable Deep Neural Networks

Daniel Uvaydov, Raffaele Guida, Pedram Johari, Francesco Restuccia, and Tommaso Melodia
Institute for the Wireless Internet of Things, Northeastern University, United States
Email: {uvaydov.d, guida.r, p.johari, frestuc, melodia}@northeastern.edu

***Abstract*—With more than 65M people affected by epilepsy worldwide, early prediction and response to seizure onsets have become more important than ever. Cutting-edge research in implantable medical devices (IMDs) has shown that deep neural networks (DNNs) applied to intracranial electroencephalogram (iEEG) data can predict seizures up to an hour before onset. However, offloading of iEEG data to the edge/cloud is highly prohibitive, due to the sheer size of the generated data. Existing work either focuses on the DNN training phase only, or does not consider the severe energy/space limitations of IMDs. Moreover, the technical aspects of patient personalization, which allows for patient-specific hyper-parameter tuning, still remain unaddressed. In this paper, we propose a platform called *AiEEG* for *in vivo* early seizure prediction, whose DNN hardware circuitry can be reconfigured remotely without surgery. We prototype *AiEEG* on a system on chip (SoC) platform and demonstrate its end-to-end capabilities in seizure prediction with a population of 30 epileptic patients, with iEEG signals coming from a real-world dataset. Extensive experimental results shows that (i) our embedded and personalized DNN has an area under the curve (AUC) averaging at 0.97 and as low as zero false positives per hour (FPH) for over half the patients, an improvement of about 3.5x with respect to a non-personalized prediction method and the best for a dataset of this size when compared to the state-of-the-art; (ii) our *AiEEG* platform consumes 4.2x less energy than a cloud-based approach, leading to a 4x battery lifetime improvement; (iii) we are able to remotely fine-tune the DNN through partial reconfiguration as needed in about 10s.**

***Index Terms*—Ultrasound, Wireless, IoT, Deep Learning, Seizure Prediction**

I. Introduction

Major progress in the field of implantable medical devices (IMDs) is transforming healthcare, with about 32M Americans currently using pacemakers, defibrillators, artificial joints, stents, and heart valves [1]. Moreover, recent advances are enabling the medical community with a new set of tools to create novel and *connected* IMDs. These include, among others, implantable brain activity monitors and neuro-stimulator devices that are used to combat pain or to treat diseases caused by brain disorders such as Parkinson's and epilepsy. With 150,000 Americans being diagnosed with epilepsy every year, and with 3.4 million people with epilepsy nationwide [2], treatment for the disorder is in increasing need [3], [4]. For this reason, numerous studies have investigated the possibility of predicting epileptic seizures up to a couple of hours before onset [3], which would enable patients to take precautions against self-inflicted body injury well in advance. Currently, seizure prediction algorithms utilize digital signal processing (DSP) techniques operating on electroencephalography (EEG) or intracranial electroencephalogram (iEEG) techniques. Proposed solutions can predict brain activity with an accuracy ranging from 85% up to 100% using a variety of methods including deep neural networks (DNNs) [5]–[9]. However, current DNN methods are too inefficient to be deployed on embedded systems – as yet, only one approved clinical trial for a real-time implementation of such systems exists [4]. Most notably, [4] requires continuous transmission of the entire iEEG data to an external device for processing. This poses a challenging problem from an energy efficiency perspective, especially if the IMD, purposed for long term use, is powered by batteries which substitution involves complex surgery.

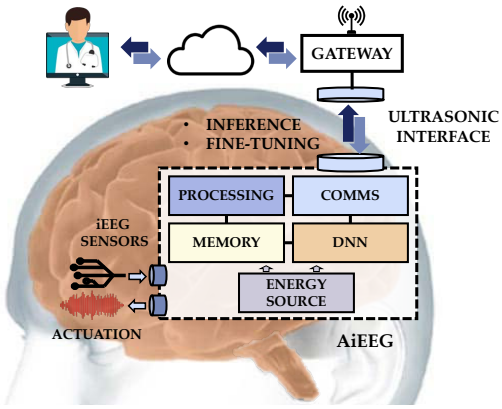

Fig. 1: *AiEEG* enables Internet of Things functionalities for DNN-based seizure prediction on implantable platform.

Furthermore, recent advances in IMD miniaturization [10]–[12], IMD sensing [13], [14] and IMD communications [15] have not considered the compelling need to perform DNN-based inference inside the platform itself. The specific issues behind DNN inference at IMDs are (i) the extremely small space that can be dedicated to the DNN circuitry; (ii) the need for personalization of the DNN with respect to the given patient. As EEG patterns vary from patient to patient, it is nearly impossible to develop universal predictors [16]. Thus, the need for algorithm personalization has become a consensus amongst experts in epilepsy treatment [17], [18]. This implies the need for a *reconfigurable* DNN where portions of the network – not only the weights – can be adapted to fit the individual patient's EEG signals.

To address these challenges, in this paper we present *AiEEG*, a prototype platform that is able to infer seizure onsets up to one hour before occurrence, communicate the

978-1-6654-1643-6/22/$31.00 ©2022 IEEE

results, and choose different convolutional neural networks (CNNs) to be remotely reconfigured through wireless connection. Figure 1 illustrates the high level architecture of *AiEEG*. *AiEEG* would be implanted along with the iEEG electrodes and communicates DNN-based classifications of iEEG signals with the outside world via ultrasonic communication. The proposed *AiEEG* platform bridges the existing gap between two apparently disjoint paradigms in epilepsy treatment/monitoring, namely, the state-of-the-art in IMDs and advances in DNN based seizure prediction. Furthermore, *AiEEG* allows for patient personalization via tunable classifier parameters and personalized CNN weights that may be easily reconfigured wirelessly. Although miniaturization of *AiEEG* is out of the scope of this paper, the computational and hardware constraints that come with small-form factors is taken into account in our design.

Summary of Novel Contributions

The *AiEEG* platform brings to the seizure prediction landscape what it has been missing so far – not only a mere implementation of hardware-based embedded deep learning, but a fully implemented prototype that (i) has connectivity capabilities tailored for intra-body communication and networking; (ii) enables *in vivo* AI-enabled seizure prediction and responsive neuro-stimulation; (iii) is wirelessly reconfigurable to allow for post-surgery algorithm tuning. To summarize:

- We design, train, and prune a personalized (*i.e.*, per-patient) CNN for early seizure prediction and localization from a 30 patient human iEEG dataset (Section III);
- We boost the classification performance by utilizing a multi-dimensional voting mechanism (Section III) and show that the area under the curve (AUC) of the hardware-based CNN averages 0.97, and false positives per hour (FPH) as low as zero for 17 patients out of 30, leading to an improvement of about 3.5x with respect to a prediction method that is not patient-specific, the best for a dataset of this size when compared to prior work [19]–[22] (Section V);
- We derive a mathematical formulation to aid with the design of the *AiEEG* system that accounts for all the latency, memory, and transfer data rates between the *AiEEG* subcomponents (Section IV). Based on this model, we prototype *AiEEG* on a system on chip (SoC) device that includes an FPGA and a microprocessor with partial reconfiguration and adaptive pacing capabilities. Experimental results show that our platform runs with 7.8x less latency than a cloud-based approach and consumes 4.2x less energy (Section V).

The rest of the paper is organized as follows: Section II goes over the state-of-the-art in seizure prediction. Section III describes our seizure prediction algorithm in detail. Section IV overviews the system design of *AiEEG* and highlights the hardware constraints. Section V discusses *AiEEG* experimental evaluation metrics, setup and results. Finally, we conclude the paper in Section VI.

II. Related Work

Seizure prediction has been extensively investigated in the past decade [5]–[9], [20]. While some of these works have good classification performance, the real-time constraints of IMDs make these solutions not applicable to real-world systems. For example, [7] and [20] utilize Long Short-Term Memory (LSTM) networks to predict seizures with near perfect performance. However, the real-time implications of their algorithm are not considered and while [20] does, they do so minimally by only mentioning the number of parameters that their network occupies. Furthermore, conversely from the present paper where the testing dataset is 30 patients, existing work mostly validates on less than 15 patients.

In the past decade, a limited number of seizure prediction prototypes and devices have been implemented [4], [23]–[25]. These devices, however, are either not able to maintain high performance, are not IoT compatible, or lack energy efficiency by wirelessly transmitting raw iEEG signals. For example, [4] requires a large implanted battery as well as a cable that extends from inside the brain to the chest–which by itself can be challenging, discomforting and even impractical. Throughout the clinical trial in [4], subjects had to maintain their implant and external devices including daily recharging and data card replacement. The approach in [25] simply implements a deep learning (DL) algorithm on an field-programmable gate array (FPGA), without providing a communication paradigm. The work in [26] offloads iEEG signals to the cloud. While cloud-based offloading is a viable option, as we show later in Section V, the transfer process necessarily impacts the latency (7.8x longer in our experiments), which is an issue in health applications where the response time is critical. Besides the computational and networking aspects, cloud-based systems almost completely neglect the energy efficiency side. For example, proposed DL algorithms for healthcare applications have shown high levels of accuracy ($> 90\%$) but require a $2.50\,\mathrm{GHz}$ CPU with $16\,\mathrm{GB}$ of RAM [27]. This, in particular becomes *quintessential* in IMD technology, because IMDs often needs non-trivial surgery for battery replacement.

Moreover, life-time exposure to wireless signals (of any kind) may cause undesirable tissue damages [15]. Due to the severe path loss introduced by the human tissue, *AiEEG* refrains from using RF-based communications and uses novel ultrasound-based communication to increase the energy efficiency and assure biological safety measures [15], [28]. Recently, in [12] an ultrasonic communication based IMD is proposed that is real-world tested, but it is not DL compatible and has not demonstrated its capabilities in a specific application. In this paper, we implement efficient DL algorithms in an implantable platform. To the best of our knowledge, *no past work has addressed wireless reconfiguration of IMDs for seizure prediction*. Wireless reconfiguration allows for constant upkeep of the prediction algorithm in the IMD with the introduction of new innovations as well as patient specific parameter tuning without any further surgery. To accommodate such a need, in the design of the *AiEEG* platform we develop a partially reconfigurable hardware that is described later in Section IV. Moreover, as elaborated in Sections III and V, our DL networks are not only built with hardware overhead in mind but are pruned to further minimize the computational

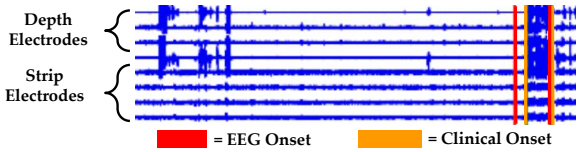

Fig. 2: 15 minute snippet of patient iEEG signals while experiencing a partial seizure

overhead while maintaining high performance, which has also never been done for this application.

III. DL-based Seizure Prediction

The erratic nature and almost non-deterministic behaviour of seizure occurrence makes modeling seizures extremely challenging. Up until the past few decades the onset of seizures were not even thought to be preceded by any interpretable information in the brain signals, making modeling a precursor to an onset very difficult [29], [30]. *AiEEG* aims to take advantage of the emerging evidence of a pre-seizure state to predict seizures before they happen via precursors that are not easily modeled. To demonstrate our capabilities for seizure prediction we employ a personalized CNN for 30 patients that we then prune for hardware efficiency. This means that the weights of the utilized CNN are trained on only their data and their data only. To further improve classification performance we employ a voting mechanism to amalgamate many consecutive classifications into fewer more robust ones.

A. Background and Problem Definition

Seizures are a unique, rapid, and rhythmic firing of neurons that cause different symptoms depending on location in the brain. The states of epilepsy fall into three categories: non-seizure (*interictal*), pre-seizure (*preictal*), and seizure (*ictal*). Classifying the pre-seizure state is key to seizure prediction. The very existence of a pre-seizure state is made more prevalent in recent studies [26], [29], [30]. This is generally challenging as the difference between pre-seizure states and non-seizure states are not easily visualized.

Due to their higher accuracy, using intracranial electroencephalography (iEEG) recordings is generally more desirable than external EEG in seizure prediction algorithms [22], [31]. iEEG uses electrodes implanted directly on the exposed surface of the brain to record electrical activity from the cerebral cortex, hence it features a higher spatial resolution and higher signal-to-noise ratio than EEG.

Seizures themselves fall into two main categories, general and partial (focal) [32]. General seizures occur throughout most of the brain, while partial seizures are localized to a specific area of the brain. Depending on the type of seizure (general or partial) and placement of the electrodes, some channels will not experience the drastic changes that other channels detect. This can be seen in Figure 2, where we see a 15 minute snippet of a patient's iEEG signals while experiencing a partial seizure. Here each horizontal line represents a different EEG channel, corresponding to a specific brain location, with the top half representing signals from inside the brain via depth electrodes and the bottom representing signals from the surface of the brain through strip electrodes. The red and orange intervals depict EEG seizure onset as labeled by a medical professional and clinical/symptomatic seizure onset respectively. As it can be seen in Figure 2, some channels have larger spike fluctuations than others; therefore, it is critical to take the varying experiences at different channels into account when deciding how to feed data to the CNN.

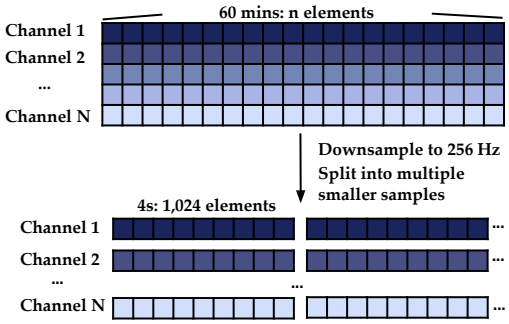

Fig. 3: CNN Data Preprocessing.

B. Dataset

We use iEEG data obtained from the The European Epilepsy Database, *EPILEPSIAE*[1]. This is the most extensive and largest publicly available iEEG dataset in existence, recording hundreds of seizures [33]. While there exists other publicly available free iEEG datasets [34], none of them are as extensive as *EPILEPSIAE*, offering fewer patients and/or less descriptive data. The database includes raw iEEG data from 30 human patients, sampled from a range of 256Hz to 2500Hz with durations from 110 hours to 440 hours. The number of iEEG channels for each patient also varies from 30 to 124 iEEG electrodes. The raw iEEG data is broken into multiple hour long samples, each with their own metadata. We define *pre-seizure* samples as any iEEG samples that come in the hour preceding a seizure; data registered before this time or before 60 minutes are *non-seizure* samples. Preictal periods have actually been shown to vary amongst different patients and even different seizures within the same patient in [35]; nevertheless, they average at less than an hour in the aforementioned paper. For this reason we chose a 60 minute pre-seizure period. The whole dataset is split by: 80% for training, 10% for validation, and 10% for testing. Our network is validated utilizing a validation set as opposed to other techniques such as k-fold cross validation as our dataset is extremely large and doing so would prove inefficient.

To prepare the data for training and testing, we first down-sample the data utilizing an FIR antialiasing lowpass filter to bring the sampling frequency to 256 Hz. Down-sampling the raw signals to the same frequency ensures that the DL network gets a consistent form of data at the cost of lower

[1]The database is sponsored by the European Union and is approved by The Ethics Committee of the University of Freiburg with the file number 131/08. In addition, all patients signed an informed consent.

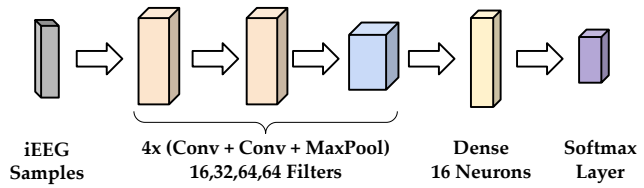

Fig. 4: Baseline CNN for all patients.

resolution. Then, the channels of the 60 minute samples are separated and each broken into smaller 4 second samples to be fed to the CNN. Smaller input samples to the CNN decrease the amount of BRAM memory used in the FPGA, which is beneficial as we use the BRAM to store the temporary input and parameters of the CNN. The pre-processing is showcased in Figure 3. The input to the CNN is a 1-dimensional array of size 1024 elements (each represented by 32 bits), representing a 4 second sample from only one iEEG channel. We only feed one iEEG channel at a time due to the varying presence that a seizure can have on different channels of an iEEG. This also allows for the model to be easily adaptable to iEEGs of varying number of channels. Furthermore, predicting on a per-channel basis allows us to localize the seizure which has not been given much emphasis in the literature.

C. CNN

Training: We chose to use a CNN as our classifier due to its consistently high performance in the literature [9]. Furthermore and more importantly CNNs can offer higher ease of implementation in hardware depending on the activation functions used, whereas other classifiers require higher computational complexity. A combination of multiple classifiers in hardware can be considered for future work. Each patient is trained on their own CNN to allow for personalization. Patient specific parameters or algorithm personalization for seizure prediction has generally shown performance improvements [17], [18]. We build our model upon CNN-based seizure prediction, as it has shown to be very promising in the medical field [36], [37], and furthermore CNNs are easily implementable in hardware.

All patients are first individually trained on a larger network shown in Figure 4. This CNN serves as the baseline for each patient, and follows a VGG style network with multiple convolutions before a max pooling layer [38]. Furthermore, we attempted to train a much larger network, double the size of the baseline network, on all the patients data to be transfer learned for individual training after. However, once transferred to each individual, the patients experienced non-logistic classification performance, much worse than our results described in Section V-C, therefore we only train individually.

The network only has two output classes, pre-seizure and non-seizure, without actually classifying the seizure itself. This is because iEEG readings of a seizure contain higher energy spikes than both pre-seizure and non-seizure readings, and are fairly distinguishable [39]. Therefore, calculating the energy of iEEG readings periodically is enough to detect the seizure as it is happening.

Pruning: In order to meet memory and latency constraints (as explained later in Section IV), the CNN must be pruned before implemented onto the FPGA. There are other methods to be considered for minimizing CNN hardware overhead such as weight quantization; however, weight quantization alone only reduces memory, while pruning can reduce both memory and number of computations. For this reason we decided to focus on pruning our network which can be used in conjunction with weight quantization. However, pruning a CNN can make it more difficult to implement in hardware, usually requiring a specialized sparse library, that needs to map neurons to each other in a very specific way. For this reason we have decided to prune whole filters and their corresponding feature maps as is done in [40]. The resulting pruned network does not have any unique connections and looks like any other CNN, making hardware implementation simpler. We prune each patient's baseline CNN for 15 rounds, retraining the network for a few epochs after each round. The filters with the smallest l_1-norm are the ones that are pruned. More specifically, we choose the bottom 10% of filters in each layer independent of other layers of the CNN to be pruned in each pruning round. We have attempted to prune for more than 15 rounds, but we saw high performance degradation after 15. This will of course vary with how much the network is decreased by in each iteration. We slightly differ from [40] as we prune multiple layers simultaneously in each pruning round before training rather than pruning just single layers before training as we experience comparable results and faster training time. This results in each patient having 16 networks (including the baseline), which are compared later in Section V.

Prediction Boosting: As classification performance can vary from patient to patient due to a multitude of factors (quality of data, quantity of data, pre-existing conditions, etc.), increasing classifier performance through other means can help alleviate the shortcomings of the CNN to provide high performance for all patients. A low-overhead approach to strengthening overall classifier performance is to aggregate multiple individual classifications before coming to a conclusive decision, a form of ensemble learning [41]. This increases the robustness of the classifier to variance and furthermore does not necessarily require highly complex computations.

We use multiple channel classifications at the same point in time to boost the classification accuracy by way of majority vote without increasing memory consumption. This can be done at the receiver end (gateway) after the transmitter has sent out the individual channel classifications outside the body. The receiver will have access to all the individual classifications for each channel; hence, is able to not only take a vote but also have a full picture on the seizure occurrence and localize the seizure. This concept is showcased in Figure 5 (top), which shows a 15 channel iEEG in a 4 second sample interval with all the CNN predictions for each channel. To even further improve classification, a majority vote is taken across time

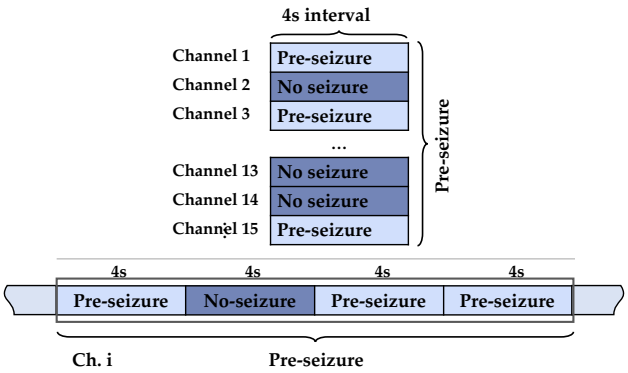

Fig. 5: CNN seizure prediction: (top) CNN predictions of a 15 channel iEEG in a 4 s sampling period; (bottom) CNN predictions of a single channel over 16 s window.

as well as space in a moving window, as shown in Figure 5 (bottom), which illustrates a 4-sample time window where each sample is 4 s. Concurrently, varying the moving window size for each patient allows another level of personalization, as patients can have larger or smaller windows depending on their classification performance.

It is worth noting that for channel voting, voting comes at the cost of latency because a final decision cannot be made until all channel classifications are finished. Unlike channel voting, increasing the size of the time voting window does not necessarily need to be accompanied by a linear increase in latency if classifications are stored in a FIFO buffer. The FIFO style buffer has a capacity of $n_c \times n_t$ classifications (or equivalently $n_c \times n_t$ bits if each binary classification is represented as one bit) where n_c is the number of total iEEG channels and n_t is the number of samples in time per channel in the window. When a new set of n_c classifications arrive, a new channel vote can happen immediately. At this time, a new time vote can also happen over the n_t newest classifications (with the addition of the newest set of n_c). This would require only waiting for every n_c classifications before a new time and channel vote can be made rather than waiting to refill the whole buffer or wait for $n_c \times n_t$ classifications. Furthermore, because each classification can be represented with only one bit, the buffer will not require a large amount of hardware space. Although there is no substantial increased hardware overhead from increasing the time window size, there will be an increased delay in the detection of a pre-seizure state. This is due to the nature of a sliding time window as it transitions from a non-seizure state to a pre-seizure state. The window will raise an alert only once over half the window is filled with classifications representing the new state (because of majority vote). For example, a 30 minute time voting window will only raise an alert that a pre-seizure state has been entered 15 minutes after entering the pre-seizure state (or 45 minutes before the seizure onset).

IV. System Architecture and Design

As illustrated in Figure 1, our proposed architecture consists of multiple components, some designed to be implantable and others external, that constitute a sensing, processing, and stimulating *closed-loop* system.

We envision that *AiEEG* will be implanted subcutaneously and partially sitting in the bone of the skull, while short wires of few centimeters are needed to connect *AiEEG* to an iEEG sensor array and to actuators (*i.e.*, leads for real-time preemptive and responsive neurostimulation) that are located in the brain region to stimulate.

The iEEG sensor records the activity of the brain and its readings are processed in the DL module. Through a wireless ultrasonic link, the CNN classification results are sent to the external gateway that pinpoints the origin of a seizure. Specific pacing settings can be sent back to the implant wirelessly based on the information inferred from the CNN results. Interfaces to actuators are designed to connect the leads to the pulse width modulation (PWM) signals generated by the FPGA to pace the brain.

At its heart, the implant features a processing and a communication unit. The processing unit (CPU) provides the computational resources to process messages, access the files containing new CNNs, and to execute the partial reconfiguration protocol (see Section IV-B for more details). The communication unit, completely developed on the FPGA instead of the CPU to reduce power consumption and avoid non-deterministic processing delays, hosts the ultrasonic physical layer, and the interfaces to other modules and to the transducer. Given its 7mm thickness, the ultrasonic transducer needs to be implanted subcutaneously and partially sitting in the bone, which means that the the implanted and the external transducer are separated by only few millimeters of tissue –mostly skin. For this reason, the power for data communication can be kept within the FDA limits ($720mW/cm^2$ [15]), and specifically, below 10 mW that, as demonstrated by previous literature [12], is enough for communications over 10 cm of tissues, which is about 10x the length of the *AiEEG* ultrasonic link. Therefore, not only does the proposed architecture eliminate extensive subcutaneous wiring, but it also removes the need for through-body radio-frequency links, hence it offers a safer and more energy efficient solution. The platform includes a memory unit composed of a general read/write memory, RAM, as well as FIFO queues to store software and data. Finally, an energy unit provides the operating power to all the active components.

The external gateway is a removable device, contained, for instance, in a patch that is attached to the skin. It receives the classification results from the implant through the ultrasonic link, executes the majority voting, and sends commands to the implant to adjust its pacing settings, such as pulse duration or pacing frequency. This process only requires minimal wireless transmission, that is a few bits representative of the classification results of the embedded DL algorithm. The gateway also includes a traditional radio-frequency (Bluetooth) communication chip. When the gateway receives a partial reconfiguration command and relative file from its Bluetooth interface, it forwards them to *AiEEG* implant –through ultrasonic communication– that initiates the partial reconfiguration process.

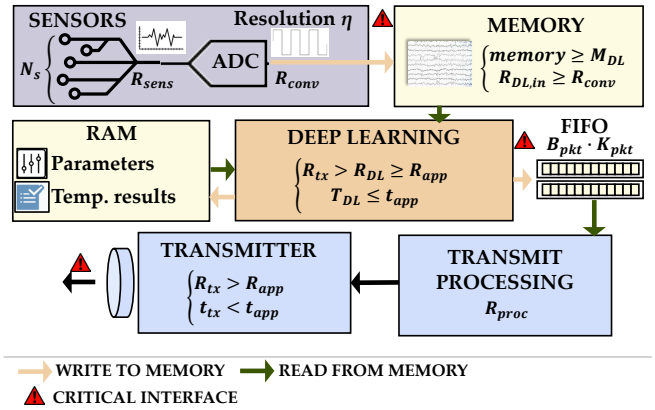

Fig. 6: System model showing *AiEEG*'s components interaction with each other and with the buffers. The critical interfaces from a design perspective are marked.

It is worthy of note that to be compliant with the FDA regulations, the implantable parts of AiEEG system have to be encapsulated in a hermetically isolated casing of bio-compatible material, such as titanium. While the design of the casing and choice of material is beyond the scope of our study, we have considered the computational and hardware limitations that are imposed by the small-form factor of the implant in our design.

Key challenges in the design of the AiEEG system were: (i) decreasing the memory allocation and computational resources required by the CNN on hardware; (ii) reducing the execution latency of the DL algorithm guaranteeing that the *real-time* condition is respected; and (iii) enabling a low overhead partial hardware reconfiguration to allow for non-invasive improvement of the prediction algorithm. For this reason, we introduce a system model of the interactions between components and report a mathematical formulation to systematically account for latencies and data rates.

A. Timing and Memory Constraints

Figure 6 shows the block diagram of the *AiEEG* system highlighting the critical interfaces between its components. The two most stringent design constraints are the minimum data rate that has to be supported by the ultrasonic communication interface, and the amount of data that the CNN needs to process in a time unit, as these identify (i) the maximum tolerable processing delay of the DL module and (ii) the smallest required memory sizes. Noticeably, memory buffers are required at the interface between the sensor unit and the processing unit and at the interface between the processing and the communication units.

Let us suppose that an application needs to be designed on top of *AiEEG* and that the results of the DL processing are encoded into B_{app} bits (application bits) that are requested each t_{app} seconds, hence the minimum bit rate required by the application is $R_{app} = B_{app}/t_{app}$ (in bit/s).

Condition I. The communication unit introduces a short processing overhead, t_{proc} [s], for each bit. $t_{tx} = 1/R_{tx}$ is the actual 1-bit transmit time in [s], where R_{tx} in $[bit/s]$ is the transmission rate. The transmitter needs to send data at an average rate R_{tx} equal or larger than R_{app}.

Condition II. The DL module computes a single classification in t_{DL} [s] and the result is encoded with B_{DL} bits. Thus, the DL module produces (and transfers to the communication unit) information bits at an average rate of $R_{DL} = B_{DL}/t_{DL}$ $[bit/s]$. The B_{app} bits must be generated by the DL unit in time to be ready for transmission, that is in an interval (T_{DL}) smaller or equal to t_{app}.

To avoid memory overflows, R_{DL} has to be smaller than R_{tx}. Hence, the condition on the output data rate is given by:

$$R_{tx} > R_{DL} = \frac{B_{DL}}{t_{DL}} = \frac{B_{app}}{T_{DL}} \geqslant R_{app}. \tag{1}$$

The data exchange between the sensing unit (iEEG sensor grid) and the processing unit requires a buffer to temporarily store the sensed data. An iEEG sensor grid is composed of N_s sensors, sending each $r_{sens} = 1/t_{sens}$ voltage values to the ADC per unit of time. The cumulative rate of ADC input values -before digitization- per unit of time is $R_{sens} = N_s \cdot r_{sens}$. The ADC converts the analog input signals into digital samples with a resolution of η bits per sample. The cumulative conversion rate (R_{conv} in $[bit/s]$) of the ADC can be given as

$$R_{conv} = \left(N_s \cdot \left(r_{sens} + \frac{1}{t_{conv}} \right) \right) \cdot \eta, \tag{2}$$

where t_{conv} is the ADC conversion latency for a single sample.

Condition III. The DL algorithm takes in input M_{DL} bits and must process them before the sensing unit terminates its conversion. Thus, the minimum size of the buffer between the sensing unit and the DL module is M_{DL} bits while, at the same time, the number of bits that the DL module can read per second ($R_{DL,in} = M_{DL}/T_{DL}$) has to be

$$R_{DL,in} \geqslant R_{conv}. \tag{3}$$

Condition IV. The transmitter module in the communication interface transfers data in bursts of K_{pkt} packets of B_{pkt} bits each. A FIFO is needed at the interface between the communication and the DL modules to momentarily store the bits produced by the DL classification before enough bits are produced to fill the payloads of the packets. Thus, the minimum size of the FIFO can be set to $B_{pkt} \cdot K_{pkt}$ bits.

B. Partial Reconfiguration

Given the infancy stage of AI applied to epilepsy prediction, we cannot disregard the fact that new deep learning algorithms will be proposed in the future, therefore, we provided AiEEG with the capability to easily update its hardware wirelessly. We achieve this goal through FPGA partial reconfiguration (PR), that consists in updating only one or few subsections (modules) of the FPGA *fabric* and their internal routing, while the fixed non reconfigurable resources of the hardware keep running during the PR operation. The advantage of the PR is threefold: (i) it allows for changing the CNN network on

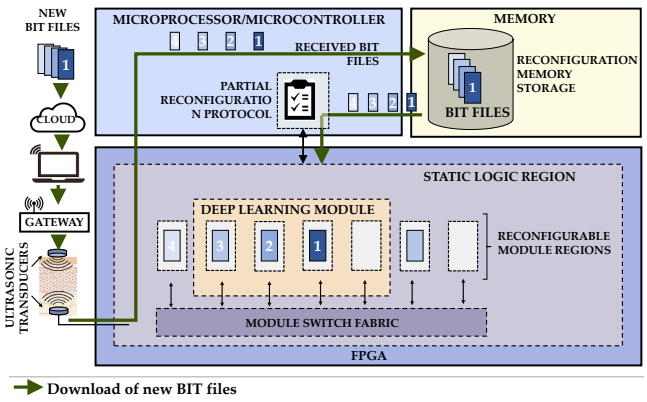

Fig. 7: *AiEEG* partial reconfiguration (PR) logic. New BIT files containing personalized CNNs can be downloaded to the implant wirelessly.

the device wirelessly, without needing to physically access the device and removes the need and discomfort of surgery; (ii) it permits to update the deep learning algorithm whenever a new one is designed, or new EEG data from the patient suggest that a different processing is required; (iii) PR has been shown to provide embedded systems with higher performance, and better energy and computational efficiency [42].

Keeping the device alive, or interrupting its operations only for short intervals, at most in the order of few seconds (it takes *AiEEG* about than 10 s to partially reconfigure its FPGA), is essential while monitoring patients suffering from chronic neurological diseases. Long interruptions to update the hardware of the implanted device, might lead to missing important anomalies in the monitored physiological signal that could indicate major events, such as a seizure, happening in the future.

Figure 7 shows a conceptual block diagram of *AiEEG* PR logic and, as illustrated, an initial plan separates the static (non-reconfigurable) portions of the FPGA from the reconfigurable ones. The static plan, or static logic region, serves as a fixed context to place and route new reconfigurable modules and it is used for all subsequent configurations. Then, during the initial plan design, specific hardware resources are reserved for PR and to contain future *bitstream files (BIT)*.

To configure the FPGA for PR, we developed multiple reconfigurable modules, each contained in a separated BIT file. When a PR module needs to be updated, the relative bitstream file is downloaded from the gateway onto the dedicated reconfigurable portion in the FPGA. A read/write memory on the implant stores one or more BIT files, each containing an updated version of an existing CNN or a completely new neural network. The processor, which has access to the memory, then uploads the BIT files to the FPGA using a specific PR protocol defined in [43].

PR bitstream files occupy few kilobits of memory, therefore they can be transferred over the ultrasonic interface in less than 10 s, while the process to replace a BIT file on the implant, once it has been received, only takes 10 ms. This reconfiguration time is very low when considering that seizures generally happen hours apart (in the non-cluster case), reducing the possibility of a seizure occurring in the 10 seconds it takes to reconfigure wirelessly. Considering that AiEEG power supplied to the communication unit is 20 mW during reception, the whole PR process requires 200mJ.

V. Experimental Results

The extensive experimental evaluation presented in this section serves two fundamental purposes. We need to demonstrate (i) whether the deep learning techniques presented in the previous sections can be effective and successfully integrated inside a resource-challenged embedded implantable system, and (ii) whether a hardware-based DL can provide better energy and latency performance with respect to a cloud-based offloading of the DL classification. In fact, when DL classification does not happen in real-time, but it is instead executed offline in the cloud, as for example in [26], machines have resources that are far beyond what a tiny IMD can offer. This means that a task can be executed using more computational power, in a shorter amount of time –in some cases– and with disregard for the energy consumption. The drawback of this approach is that the implant needs to send large amounts of raw data to the cloud. Thus, we need to obtain a complete break down of the energy and latency spent in different tasks (communication, computation etc.), and establish (a) if the on-board processing is energetically more efficient than cloud offloading, (b) if the processing latency of the embedded device allows for DL computation on board, and (c) if a cloud-based approach provides a better trade-off between the energy needed for communication and the energy that the implant would save given the offloaded processing.

In the next subsections, we first present the performance metrics used to evaluate our classifier performance in Section V-A. Then the *AiEEG* prototype in Section V-B. We then report the performance of the trained CNN for individual patients and the achieved improvements from the prediction boosting in Sections V-C. Finally, in Section V-D, we benchmark the FPGA based approach of *AiEEG* against a cloud-based solution and further present a system-wide demonstration of *AiEEG* to evaluate its latency, power, and energy consumption.

A. Performance Metrics

Two metrics that are highly indicative of a classifier's performance for seizure prediction and widely used in the seizure prediction research domain are sensitivity and specificity [44], [45]. Sensitivity is defined as the ability for a classifier to predict true positives (pre-seizure state), while specificity is the ability to predict true negatives (non-seizure state). These two metrics alone however only tell a part of the story, time is also a factor in seizure prediction. For this reason we report sensitivity as a function of time, and for specificity, the *Time In Warning* (TIW). More specifically, TIW is defined as the percentage of time that a patient spends incorrectly waiting due to a false positive or false alarm for a seizure to come,

ideally spending no time at all (or 0% TIW). TIW is calculated as,

$$TIW = PredictionHorizon * \frac{FalsePositives}{TotalTestPeriod}, \quad (4)$$

where in our case the prediction horizon is 1 hour and the total test period (total time of all test samples classified) and total number of false positives varies per patient. The fraction on the right is the FPH. Furthermore, we clip TIWs at 100% as anything over essentially just means the patient is always falsely waiting for a seizure. In addition, for a better holistic view of our classifier, we also utilize the area under the receiver operating characteristic curves (AUC-ROC or AUC), which are widely used in previous works [3].

B. Prototype Implementation

We implemented *AiEEG* on a Zynq UltraScale+ system-on-chip (SoC) on top of a ZCU102 evaluation board.This board features an FPGA that can be fabricated in format as small as 31×31 mm. For perspective, NeuroPace, an FDA approved brain implantable for epilepsy treatment is 28×60 mm at its smallest and 42×60 mm at its largest, therefore this FPGA is well within the range of admissible sizes.

The CNN was trained and tested on a local computer. The weights and architecture of the CNN were then transferred to the FPGA. A key point here is that the CNN was not trained on the FPGA, but only used on the FPGA for predicting new outputs from new inputs once already trained offline. The weights and architecture of the CNN were first coded in C and then synthesized using High Level Synthesis (HLS) tools.

We prototyped the implant according to the model in Section IV and in consideration of the conditions in Section IV-A. We dimension the system for the *worst case scenario*, that is with respect to the patient's iEEG that has the largest number of electrodes (124).

- *AiEEG*'s bitrate over the ultrasonic link is $R_{tx} = 150$kbit/s with a BER of 10^{-6}. Thus, according to **Condition I**, 150 kbit/s is the maximum application rate R_{app} that can be satisfied. The processing delay introduced by the communication unit before transmission is $t_{proc} = 75$ µs per byte. The CNN carries out a classification over all the channels in $t_{DL} = 26\,\text{ms} \cdot 124$. The CNN processing result is encoded into $B_{DL} = 124$ bits, as only 1 bit per channel is enough to represent a binary outcome (pre-seizure/non-seizure).
- To classify a 4 s sample the DL takes in input $1024 \cdot 32$ bits per channel and a total of $M_{DL} = 4.1$ Mbits across all the channels. **Condition II** ($R_{DL} < R_{tx}$) is seamlessly respected.
- As discussed in Section III-B, the fastest sampling rate per channel is 2500 Hz and each sample is digitized with a resolution of $\eta = 32$ bits. Therefore, in the most conservative case, the ADC unit needs to sustain a rate of $124 \cdot 2500$ Sa/s, and the conversion frequency of the ADC unit is $R_{conv} = 9.92$ Mbit/s, reduced to 1.02 Mbit/s after downsampling at 256Hz, which is the input rate of the CNN. **Condition III** is also met, as the DL module reads bits at rate $R_{DL,in} = 1.27$ Mbit/s, which is > 1.02 Mbit/s.
- Finally, the FIFO size needs to be at least $B_{pkt} \cdot K_{pkt} = 124$ bits to avoid overflows (**Condition IV**).

The external gateway was implemented on a Zynq-7000 system-on-chip (SoC) on top of a Zedboard evaluation board and it features an ultrasonic communication interface, similar to the one of the implant system. Internet access, is provided to the gateway either through Bluetooth, or by a host computer using a wired connection, *e.g.* Ethernet.

In this paper, we present a prototype version of *AiEEG*. However, in our vision, three major development and test phases which include R&D and manufacturing, pre-clinical testing, and regulatory clinical trials, will be necessary before *AiEEG* can reach a commercial stage. *AiEEG* devices will initially be miniaturized and manufactured as class III medical implants. The first step towards miniaturization is the design of a single printed circuit board (PCB) containing the FPGA, a microcontroller, a memory slot and the ultrasonic transducer that, in the current prototype, are provided by large evaluation boards as can be seen from Figure 11. Reduced form factor will introduce new technical challenges mostly to design the ultrasonic front-end, the memory circuitry, and to reserve enough space to allocate the energy storage in a space and energy limited environment. Miniaturized devices will be packaged and submitted to pre-clinical testing. After the completion of validation and verification testing (*in-vitro* and *in-vivo* animal study), aiming at mitigating risk among other compliance testing, US FDA regulatory approval will be sought by completing a first in man trial with a small sample of tents of patients. This preliminary trial is required to demonstrate technical feasibility, effectiveness, and safety. FDA approval will be obtained after a successful controlled pivotal study involving a larger patient set (hundreds of implantations). Fault tolerance is a crucial aspect for medical systems and determining factor to obtain regulatory approval. Therefore, one of the ways to increase the robustness of the system against hardware/firmware failures is replication. Replication could be achieved by reserving other portions of the FPGA for emergency backup purposes and using them in case of failure of the main logic. Hardware replication, *i.e.* equipping *AiEEG* with a secondary backup FPGA to be used in case of failure, is another option but an increase in the size of the PCB has to be considered. Adopting this solution also requires to create extra connections between the memory, the microcontroller and the auxiliary FPGA. The whole development and approval process of *AiEEG* can take approximately between 7 and 10 years.

C. Seizure Prediction and Prediction Boosting

After 15 rounds of pruning, the fully pruned network for each patient remained either the highest performing network or only slightly lower than the baseline (about 4% lower in the worst case scenario for a single patient). We use the receiver operating characteristic (ROC) curves to judge the overall classification performance of our predictor, amongst

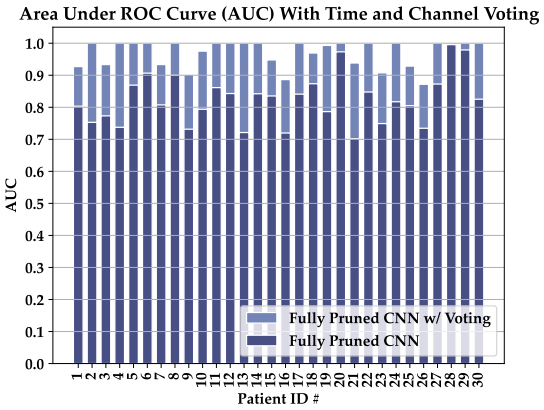

Fig. 8: Prediction improvements for each patient from channel and time voting employed on the smallest pruned CNN.

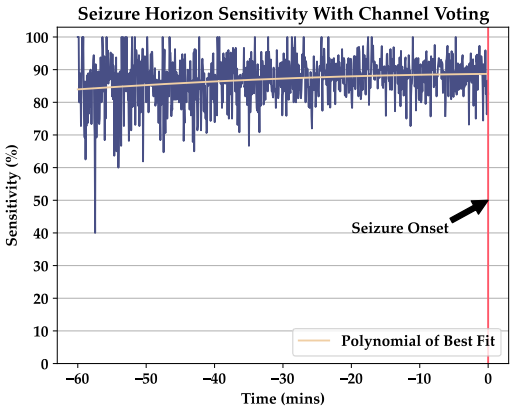

Fig. 9: Average sensitivity as a function of time in the seizure horizon using a fully pruned CNN w/ channel voting.

other metrics discussed later. More specifically, we report the area under the ROC curve (AUC) in Figure 8, where each bar is a different patient. This figure illustrates both the AUC for the fully pruned network alone as well as the AUC for pruned network with voting. Without vote boosting, the *mean* AUC is 0.82. This *mean* is higher than the highest performing AUCs on substantially sized (but not nearly as large as our dataset) datasets in two seizure prediction competitions [21], [22], both of which had an AUC of 0.76 on further unseen data.

With voting we can clearly see a substantial performance improvement from combining multiple classifications in space and time, increasing the AUC average form 0.82 to 0.97. For Figure 8 we use a 120 sample time window, which is a total of about 8 minutes worth of data. We choose this voting window as the majority of patients start to experience significant improvement at this time window size, most likely due to the large sample size.

It is worth noting that the majority of improvements occur with the addition of time voting. This is due to our flexibility in increasing the voting window size. The time window can be made larger or smaller depending on the patient, allowing for fine tuning of the CNN's performance even after *AiEEG* is implanted. This is made possible with partial reconfiguration. The channel voting, however, is bottle-necked by the number of available iEEG channels, which varies from one patient to another.

For space or channel voting, we use all the iEEG channels available for each patient. The more iEEG channels the better the performance, but the higher the latency to make a classification for a single point in time. Figure 9 plots the average sensitivity as a function of time in the seizure horizon or pre-seizure period (one hour before the seizure onset) with just channel voting. We can see that our sensitivity decreases only slightly towards the beginning of the pre-seizure period, showing that we can make predictions far in advance to the onset. This also allows us to take a vote in time without concerns of contaminating the voting window with low performing points in time.

Finally, we present the Time In Warning (TIW) for each patient in Figure 10 as a specificity metric, which describes the percentage of time a patient is waiting for a seizure that will never come due to false positives. This value ideally needs to be 0 and is calculated as done in [44] using FPH. Each patients interval voting window size is varied depending on the CNNs performance. By relying on *AiEEG* patient personalization capabilities, we increased the window size for those that have shown worse specificity in order to optimize the system performance for each patient. By personalizing time windows for each patient we are able to reduce the average FPH by 3.5 times. We notice that while few of the patients experience TIW higher than 50%, most experience little to no TIW. Our framework averages a TIW of 29% over 30 patients. For comparison, [19] and [20] have an average TIW of 27% and 0.4% respectively but are only averaged over 10 and 8 patients respectively, where in the latter the patients whose EEG data deemed to be incomplete were omitted from the dataset. To put that in perspective we have 17 patients with 0% TIW or 0 FPH.

D. AiEEG End-to-End Performance Evaluation

To exhaustively measure the latency and energy consumption of each and every component, we set up a testbed shown

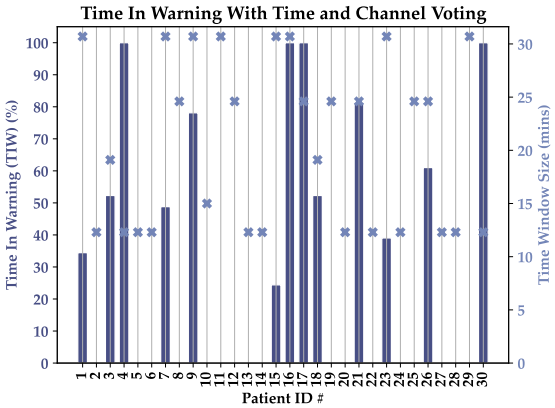

Fig. 10: Time In Warning (TIW) for each patient using the fully pruned CNN with interval and channel voting.

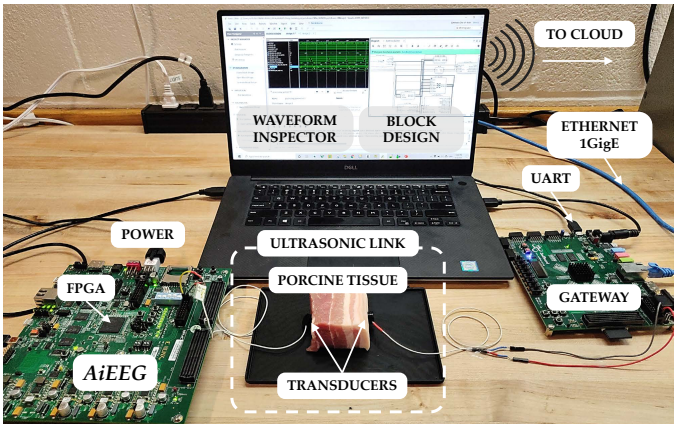

Fig. 11: *AiEEG* test set-up.

in Figure 11. The first board –labelled as "*AiEEG*"– hosts the *AiEEG* prototype, while the other board on the right side (which includes an ultrasonic module and a 1GigE Ethernet interface) serves as the gateway and it is connected to a host that exchanges data with the cloud. An ultrasonic transducer is connected to each of the two boards to send and receive data through the ultrasonic link, mimicked by a piece of porcine tissue (6 cm). We used a piece of fresh pork belly made out of different tissue layers, including skin, to model the human scalp, fat and muscle. The external transducer was attached on the skin side of the meat. *AiEEG* is designed to be implanted subcutaneously, thus the extra thickness of the porcine meat guarantees the communication performance of the ultrasonic link in more superficial applications.

The *AiEEG* platform is initially programmed with a CNN and the iEEG data loaded to the BRAM are processed and classified. The classification results are sent to the gateway and from here to the host.

(1) CNN Latency. To compute the energy consumption to classify a total of 124 samples (one for each iEEG channel in the most conservative case), we first need to compute the latency to run the CNN on *AiEEG*'s FPGA. By using a timer on the FPGA, we find the latency of classifying one input sample on the FPGA (one time interval on one channel) as 26 ms.

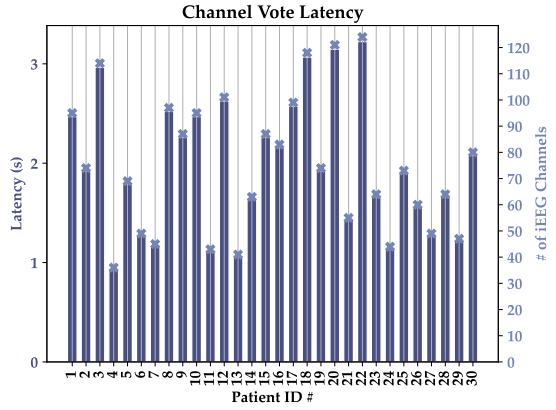

Fig. 12: Latency for classifying all channels for a different patients with different number of iEEG channels.

Depending on the number of iEEG channels each patient has, the latency for calculating a channel vote or classifying all the channels in a single time sample will vary. In Figure 12 we see the latency for performing classifications on all channels at one instance in time for different patients. Patients with more iEEG channels experience higher latency. In our application to predict the seizure up to an hour in advance, a few seconds of latency are negligible. If the classification, instead, is performed in the cloud for faster computing, the process would be bottle-necked by the transmission latency of sending raw multi-channel iEEGs from the wireless ultrasonic interface resulting in an end-to-end delay 7.8x longer when compared to on-baord processing as we report later in (3).

(2) Ultrasonic Link Latency. The next step is to measure the ultrasonic data rate –to estimate the time needed to transfer the CNN input data to the cloud. We measured a bitrate of 150 kbit/s with a BER of 10^{-6} for the ultrasonic connection and an average rate of 500 Mbit/s on the 1GigE Ethernet link. The average uplink time to exchange data with the cloud, for packets smaller than 2 Mbytes, is 232 ms. Both the Ethernet and the connection with the cloud are based on TCP transport protocol and use a socket to establish a connection. Each socket set-up time measured approximately 96 ms.

(3) End-to-End Performance. The latency information alone is not sufficient to have a complete insight of the task performance, therefore we measured the power consumption of each basic operation executed inside the SoC using the Vivado tool which gives a circuit-level breakdown of the power consumption on both the CPU and the FPGA. In Figure 13 we report the power, latency, and energy consumption in the case of classification carried out directly on the *AiEEG* system and in the case of cloud offloading. The histograms show the distribution of the power between DL, external communications (Ethernet and to cloud), ultrasonic link and other processing performed on the implant.

In cloud-based offloading, a large percentage of the implant power is spent for the ultrasonic data transmission. Furthermore, this solution takes almost 30 s for data communication, 91% of which are needed to transfer data from the implant to the the gateway over the ultrasonic link. This clearly shows the importance of having an embedded system like *AiEEG* that is capable of performing crucial tasks such as DL-based seizure

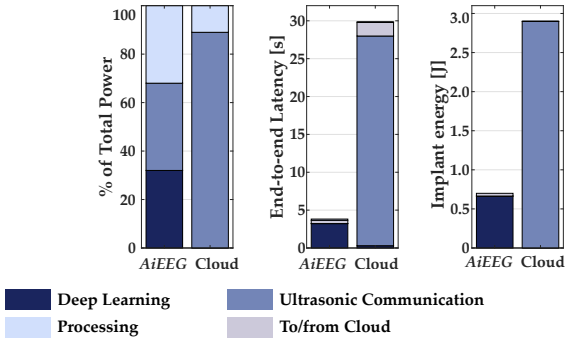

Fig. 13: Power, latency, and energy consumption of *AiEEG* vs computation offloading to cloud.

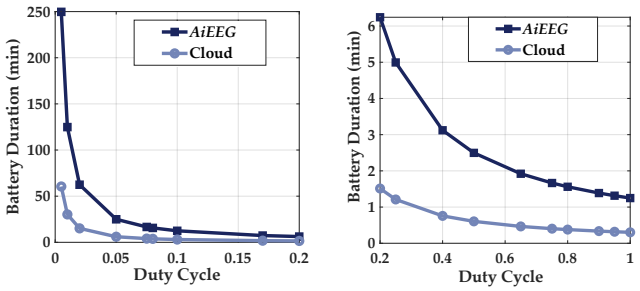

Fig. 14: Rechargeable battery duration for varying duty cycle of the CNN execution frequency on *AiEEG* vs cloud offloading.

prediction at the implant level.

As it can be seen in Figure 13, *AiEEG* has an even distribution of the power (0.56 W total) among DL, ultrasonic communication and other processing. Indeed, since the data that need to be transmitted outside are only the CNN classification results (*i.e.*, 16 bytes), the time spent during ultrasonic communication is smaller than the 1% of the total end-to-end delay. The latency histogram relative to *AiEEG* shows that 424 ms of the total 3.82 s are spent in sending data to the cloud. Also, there is a minimum delay of 232 ms to upload a packet of any size smaller than 2 MB to the cloud. This fixed delay is between the host and the remote database, it does not depend on the *AiEEG* system and it does not affect the power (and the energy) consumption of the implant.

(4) *AiEEG* Lifetime. To have a practical understanding of the lifetime of the platform, we consider the case in which the CNN is executed at periodic intervals. We compare the energy consumption of *AiEEG* with cloud offloading using a 12 mAh rechargeable battery (PowerStream GM300910 [46]) as a reference to power the implant. We define a duty cycle as the fraction of the time to execute the CNN (ON time) on the FPGA. Figure 14 shows the duration of the battery for different values of the duty cycle assuming energy consumption values measured above. This figure shows that the battery duration is about 4x longer when the DL classification is performed on the board than the cloud-based solution.

VI. Conclusion

We presented *AiEEG*, an embedded ultrasonically networked platform with a deep learning core for AI-enabled closed-loop responsive neuro-stimulators with *in vivo* reconfigurability. The system implements a convolutional neural network (CNN) to classify iEEG signals for the early prediction of epileptic seizures. *AiEEG*, furthermore, is wirelessly reconfigurable to allow for patient-specific hyper-parameter tuning and upgrading after implanted on a patient with minimal interruption. We proposed a practical implementation based on hardware, including a field programmable gate array (FPGA) as the core. In addition, we reported experimental results to show the feasibility of *AiEEG*, specifically of the implementation of the CNN on an embedded system that includes a communication unit. This allowed us to drastically save on energy as we only transmit CNN classifications rather than raw iEEG signals, as needed for cloud offloading. We demonstrated the transferring of the CNN output through animal tissues to a receiving unit and reported an average AUC of 0.97 and with 7.8x less latency and 4.2x less energy than cloud based methods.

References

[1] Jeanne Lenzer, "Can Your Hip Replacement Kill You?" https://tinyurl.com/yyn63aug, 2018.

[2] M. M. Zack and R. Kobau, "National and state estimates of the numbers of adults and children with active epilepsy—united states, 2015," *MMWR. Morbidity and mortality weekly report*, vol. 66, no. 31, p. 821, 2017.

[3] L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter, and H. P. Zaveri, "Seizure prediction—ready for a new era," *Nature Reviews Neurology*, vol. 14, no. 10, pp. 618–630, 2018.

[4] M. J. Cook, T. J. O'Brien, S. F. Berkovic, M. Murphy, A. Morokoff, G. Fabinyi, W. D'Souza, R. Yerra, J. Archer, L. Litewka *et al.*, "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study," *The Lancet Neurology*, vol. 12, no. 6, pp. 563–571, 2013.

[5] Y.-E. Lyu, X.-F. Xu, S. Dai, X.-B. Dong, S.-P. Shen, Y. Wang, and C. Liu, "Intracranial electrodes monitoring improves seizure control and complication outcomes for patients with temporal lobe epilepsy-a retrospective cohort study," *International Journal of Surgery*, vol. 51, pp. 174–179, 2018.

[6] D. Ahmedt-Aristizabal, C. Fookes, K. Nguyen, and S. Sridharan, "Deep classification of epileptic signals," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 332–335.

[7] R. Hussein, H. Palangi, R. Ward, and Z. J. Wang, "Epileptic seizure detection: a deep learning approach," *arXiv preprint arXiv:1803.09848*, 2018.

[8] M. Alhussein, G. Muhammad, M. S. Hossain, and S. U. Amin, "Cognitive iot-cloud integration for smart healthcare: Case study for epileptic seizure detection and monitoring," *Mobile Networks and Applications*, vol. 23, no. 6, pp. 1624–1635, 2018.

[9] T. N. Alotaiby, S. A. Alshebeili, T. Alshawi, I. Ahmad *et al.*, "Eeg seizure detection and prediction algorithms: a survey," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 183, 2014.

[10] S. A. Wirdatmadja, S. Balasubramaniam, Y. Koucheryavy, and J. M. Jornet, "Wireless optogenetic neural dust for deep brain stimulation," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016, pp. 1–6.

[11] J. Charthad, T. C. Chang, Z. Liu, A. Sawaby, M. J. Weber, S. Baker, F. Gore, S. A. Felt, and A. Arbabian, "A mm-sized wireless implantable device for electrical stimulation of peripheral nerves," *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 2, pp. 257–270, 2018.

[12] R. Guida, N. Dave, F. Restuccia, E. Demirors, and T. Melodia, "U-Verse: a miniaturized platform for end-to-end closed-loop implantable internet of medical things systems," in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 311–323.

[13] J. Shin, Z. Liu, W. Bai, Y. Liu, Y. Yan, Y. Xue, I. Kandela, M. Pezhouh, M. R. MacEwan, Y. Huang *et al.*, "Bioresorbable optical sensor systems for monitoring of intracranial pressure and temperature," *Science advances*, vol. 5, no. 7, p. eaaw1899, 2019.

[14] M. H. Yacoub and C. McLeod, "The expanding role of implantable devices to monitor heart failure and pulmonary hypertension," *Nature Reviews Cardiology*, p. 1, 2018.

[15] G. E. Santagati and T. Melodia, "An Implantable Low-Power Ultrasonic Platform for the Internet of Medical Things," in *Proc. of IEEE Conf. on Computer Communications (INFOCOM)*, Atlanta, USA, May 2017.

[16] M.-P. Hosseini, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh, "Optimized deep learning for eeg big data and seizure prediction bci via internet of things," *IEEE Transactions on Big Data*, vol. 3, no. 4, pp. 392–404, 2017.

[17] J. Birjandtalab, V. N. Jarmale, M. Nourani, and J. Harvey, "Impact of personalization on epileptic seizure prediction," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[18] S. B. Dumanis, J. A. French, C. Bernard, G. A. Worrell, and B. E. Fureman, "Seizure forecasting from idea to reality. outcomes of the my seizure gauge epilepsy innovation institute workshop," *Eneuro*, vol. 4, no. 6, 2017.

[19] I. Kiral-Kornek, S. Roy, E. Nurse, B. Mashford, P. Karoly, T. Carroll, D. Payne, S. Saha, S. Baldassano, T. O'Brien *et al.*, "Epileptic seizure prediction using big data and deep learning: toward a mobile system," *EBioMedicine*, vol. 27, pp. 103–111, 2018.
[20] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 5, pp. 804–813, 2019.
[21] B. H. Brinkmann, J. Wagenaar, D. Abbot, P. Adkins, S. C. Bosshard, M. Chen, Q. M. Tieng, J. He, F. Muñoz-Almaraz, P. Botella-Rocamora *et al.*, "Crowdsourcing reproducible seizure forecasting in human and canine epilepsy," *Brain*, vol. 139, no. 6, pp. 1713–1722, 2016.
[22] L. Kuhlmann, P. Karoly, D. R. Freestone, B. H. Brinkmann, A. Temko, A. Barachant, F. Li, G. Titericz Jr, B. W. Lang *et al.*, "Epilepsyecosystem. org: crowd-sourcing reproducible seizure prediction with long-term human intracranial eeg," *Brain*, vol. 141, no. 9, pp. 2619–2630, 2018.
[23] J. J. Howbert, E. E. Patterson, S. M. Stead, B. Brinkmann, V. Vasoli, D. Crepeau, C. H. Vite *et al.*, "Forecasting seizures in dogs with naturally occurring epilepsy," *PloS one*, vol. 9, no. 1, 2014.
[24] S. Hooper, E. Biegert, M. Levy, J. Pensock, L. van der Spoel *et al.*, "On developing an fpga based system for real time seizure prediction," in *51st Asilomar conference on signals, Systems, and Computers*. IEEE, 2017, pp. 103–107.
[25] S. Tamilarasi and J. Sundararajan, "Fpga based seizure detection and control for brain computer interface," *Cluster Computing*, vol. 22, no. 5, pp. 11 841–11 848, 2019.
[26] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.
[27] M. M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan *et al.*, "Deep learning approach for active classification of electrocardiogram signals," *Information Sciences*, vol. 345, pp. 340–354, 2016.
[28] G. E. Santagati and T. Melodia, "A software-defined ultrasonic networking framework for wearable devices," *IEEE/ACM Transactions on Networking*, vol. PP, no. 99, pp. 1–14, 2016.
[29] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz, "On the predictability of epileptic seizures," *Clinical neurophysiology*, vol. 116, no. 3, pp. 569–587, 2005.
[30] M. Sharma, R. B. Pachori, and U. R. Acharya, "A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension," *Pattern Recognition Letters*, vol. 94, pp. 172–179, 2017.
[31] N. D. Truong, A. D. Nguyen, L. Kuhlmann, M. R. Bonyadi, J. Yang, S. Ippolito, and O. Kavehei, "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Networks*, vol. 105, pp. 104–111, 2018.
[32] "Types of seizures." [Online]. Available: https://tinyurl.com/yxa4c3p3
[33] M. Ihle, H. Feldwisch-Drentrup, C. A. Teixeira, A. Witon, B. Schelter *et al.*, "Epilepsiae–a european epilepsy database," *Computer methods and programs in biomedicine*, vol. 106, no. 3, pp. 127–138, 2012.
[34] "American epilepsy society seizure prediction challenge." [Online]. Available: https://www.kaggle.com/c/seizure-prediction/overview
[35] M. Bandarabadi, J. Rasekhi, C. A. Teixeira, M. R. Karami, and A. Dourado, "On the proper selection of preictal period for seizure prediction," *Epilepsy & Behavior*, vol. 46, pp. 158–166, 2015.
[36] A. Yadollahpour and M. Jalilifar, "Seizure prediction methods: a review of the current predicting techniques," *Biomedical and Pharmacology Journal*, vol. 7, no. 1, pp. 153–162, 2015.
[37] S. Ramgopal, S. Thome-Souza, M. Jackson, N. E. Kadish, I. S. Fernández, J. Klehm, W. Bosl, C. Reinsberger *et al.*, "Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy," *Epilepsy & behavior*, vol. 37, pp. 291–307, 2014.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[39] S. Smith, "EEG in the diagnosis, classification, and management of patients with epilepsy," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 2, pp. ii2–ii7, 2005.
[40] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.
[41] H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet, "A survey on ensemble learning for data stream classification," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–36, 2017.
[42] C. A. de Albuquerque Silva, A. A. R. Diniz, A. D. D. Neto, and J. A. N. de Oliveira, "Use of partial reconfiguration for the implementation and embedding of the artificial neural network (ann) in fpga." in *PECCS*, 2014, pp. 142–150.
[43] Xilinx Inc., "Vivado Design Suite. Tutorial Partial Reconfiguration UG947 (v2017.1)," https://tinyurl.com/2bexbr9y, April 2017.
[44] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, "Seizure prediction: the long and winding road," *Brain*, vol. 130, no. 2, pp. 314–333, 2006.
[45] D. E. Snyder, J. Echauz, D. B. Grimes, and B. Litt, "The statistics of a practical seizure warning system," *Journal of neural engineering*, vol. 5, no. 4, p. 392, 2008.
[46] Guangzhou Markyn Battery Co., Ltd, "Polymer Lithium Ion Battery Specifications," https://tinyurl.com/y5dshlne, 2007.