A Look at Signal Analysis in Resource-constrained Medical-sensor Applications

Naveen Verma, Zhuo Wang, Jintao Zhang
Department of Electrical Engineering
Princeton University
Princeton, NJ USA

***Abstract*— Given the increasing emphasis on data-driven medicine, this overview paper explores how healthcare functions can be extended to on-patient sensor platforms. While the functions that will be of interest in this context are not precisely known, the approaches within data-driven medicine are becoming clear. This guides us to the algorithms that will be used for diagnosis, monitoring, and therapy. In particular, there is an emphasis on machine-learning algorithms, which enable the creation of robust models for signal analysis from data. Focusing on such algorithms, this paper looks first at digital platforms for signal analysis that provide adequate programmability while addressing computational energy. Then, it considers approaches to signal acquisition that focus on acquiring the specific information needed within algorithms for signal analysis, with the aim that such focus relaxes both the specifications for mixed-signal interfaces and the subsequent computations required. The ideas presented reference hardware prototype demonstrations.**

***Keywords*—biomedical electronics, machine learning, medical signal processing.**

I. Introduction

Data-driven medicine is transforming the way that healthcare is being approached on virtually all levels. A key driver of data-driven medicine is the scale of the healthcare enterprise today. This scale poses both challenges, in ensuring high-quality and cost-effective delivery to a large base of patients, and opportunities, in the knowledge that can be extracted for diagnosing diseases and prescribing treatments when patient data is viewed over large aggregations. Often, the implications of such scale are thought of in the context of centralized analytics. However, the challenges and opportunities have clear extensions to individualized patient healthcare, where patients are generally highly distributed. Extending data-driven medicine to individual patients raises the need for corresponding platforms *on* the patients. Thus, the microelectronics industry, which is well positioned to address scale, can potentially play a critical role.

This paper looks at one particularly important realm within data-driven medicine, oriented around the acquisition of medical data itself. Recent years have seen tremendous advances in sensing technologies, which have raised the possibility of acquiring highly-informative signals. However, the functions of interest in systems today progress beyond simple signal acquisition, to robust *signal analysis*. For intsance, this can enable extensive patient monitoring for diagnosis, persistent decision support towards treatment/care plans, and intelligent actuation of therapies. Whether the signal analysis is ultimately performed locally on the patient or away from the patient will generally depend on the bandwidth, latency, and energy of communication with respect to the functions of interest. Thus, generall, to address various application scenarios, on-patient platforms capable of providing baseline signal-analysis capabilities are of interest. The primary challenge brought on by on-patient platforms is resource constraints (energy, size, etc.). The following sections start by outlining the high-level signal-analysis challenges and then go on to examine platform approaches for digital signal analysis and mixed-signal acquisition under such constraints.

Support is provided by SRC, NSF (CCF-1253670), AFOSR (FA9550-14-1-0293), as well as Center for Future Architectures Research (CFAR) and Systems on Nanoscale Information fabriCs (SONIC), two of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

II. Analyzing Physiological Signals

The analysis of physiological signals faces numerous challenges. Indeed, before analysis can even begin, there is the problem of robust acquisition. This is affected by sensing non-idealities, such as stray coupling and motion artifacts on electrodes, as well as readout non-idealities, such as noise and nonlinearities in the analog frontend. These represent significant challenges, which have warranted active research. While promising solutions have emerged, these often require expending resources, making signal acquisition a significant share of the system energy. We return to opportunities for addressing signal acquisition from the perspective of the signal-analysis algorithms, but first we start by outlining the major challenges faced assuming robust acquisition. Broadly, these are as follows:

1. The signals available through sensors typically express the physiological processes of interest in the presence of various other processes. Accurate detection thus requires high-order modeling of targeted processes. However, the processes are often highly complex, without strong analytical representations. As an example, Fig. 1 illustrates patient EEG data during a seizure; before the seizure onset, normal physiologic activity (spindle EEG) having close electrographic resemblance to seizures is observed. Thus, distinguishing this requires robust models of the electrographic expressions.

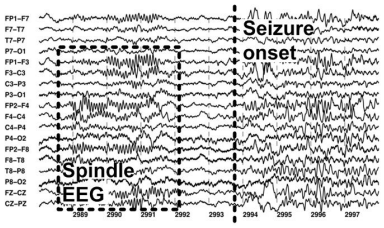

Fig. 1. EEG during seizure onset, compared with spindle.

2. In addition to requiring robust, high-order models for signal analysis, the expressions of targeted physiological states are typically variable from patient to patient. Thus, some mechanism must exist whereby the models can be efficiently customized on a patient-by-patient basis.
3. The physiological processes of interest within a patient are often dynamic. In particular, changes can be prominent following actuate pathophysiologic events, which might represent important periods for patient monitoring. Consequently, some mechanism must exist whereby the patient-customized models can also be efficiently adapted in response to the changes.

Considering these challenges, data-driven medicine, and in particular its extension via on-patient platforms presents important capabilities. In large part, these are actuated by advances in *machine-learning algorithms*, which enable powerful methods both for constructing high-order models from data and for applying the models to derive inferences (i.e., decisions based on the data). The goal is to enable such algorithms within on-patient platforms so that functions based on the inferences can be performed outside the resource-rich hospital environment, potentially over chronic time scales. With respect to the challenges above, the functions include the following: (1) accurate detection of targeted physiological states in the absence of strong analytical models of the signal expressions; (2) acquisition and analysis of patient-specific data to refine from generic to customized models; and (3) on-going acquisition and analysis of data to actuate changes to the models in response to evolving physiology.

III. On-Patient Platforms for Signal Analysis

Focusing on machine-learning algorithms we start by considering digital platforms for signal analysis and then look at mixed-signal interfaces, for acquiring the specific information needed for analysis.

A. Digital Platforms for Signal Analysis

A primary objective for digital platforms is to achieve low energy while providing a high degree of programmability, both to address a broad range of medical applications and to address various use cases and parameters within an application. An opportunity that arises in machine-learning algorithms is that the computations requiring the greatest programmability are often separate from those requiring the greatest energy efficiency [1]. A reason for this is that, while machine-learning algorithms enable the construction of high-order models based on data, such models often do not present compact parametric representations but are employed within specific computational kernels. This can make the approach of heterogeneous

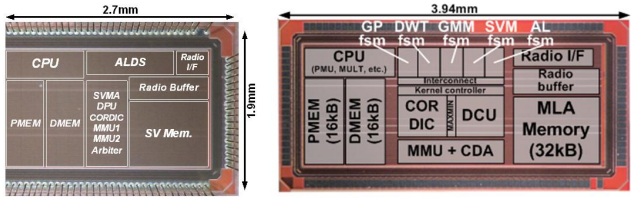

Fig. 2. Accelerator-based custom microprocessors [1,2].

accelerator-based platforms appealing. Herein, computations requiring a high degree programmability can be delegated to a general-purpose CPU; broadly, this includes feature-extraction computations, which depend on the signals in the applications and may evolve as appropriate biomarkers within the signals are discovered (often in a patient specific manner). On the other hand, inference computations, which are required to apply the high-order models through specific kernels, can be delegated to accelerators.

While acceleration enables substantial energy reduction, two challenges must be addressed: (1) even the inference computations themselves present important design optionsm, which must be exposed to the application level; and (2) acceleration primarily addresses computational energy, leaving the energy of memory accessing emerging as a bottleneck. With regards to application-level configurability, two studies, resulting in the custom microprocessors shown in Fig. 2, have investigated how this can be achieved within the microarchitectures of the integrated accelerators [1,2]. In [1], the focus is on energy-scalable classification. A support-vector machine (SVM) accelerator is employed. Applications level studies show that the SVM kernel function required, which strongly impacts computational energy, depends on the distribution of application-level data. Thus, on a low level, configurability in the choice of kernel and configurability in the computational structuring for the kernel is provided, while on a high level, structured ways of combining kernels to implement various meta-algorithms is provided. In [2], the focus is on enabling a broader range of inference algorithms and functions beyond SVM classification/regression (e.g., Gaussian mixture modeling, hidden-Markov modeling, support vector machine, etc.). To achieve this, the microarchitecture employs the 'layers' shown in Fig. 3. From design analysis and application measurements, it is found that the most critical layer is the 'Kernel Controllers', which play the role of accessing data from memory, and structuring them (into matrices and vectors) to optimally feed the 'Arithmetic Engines.'

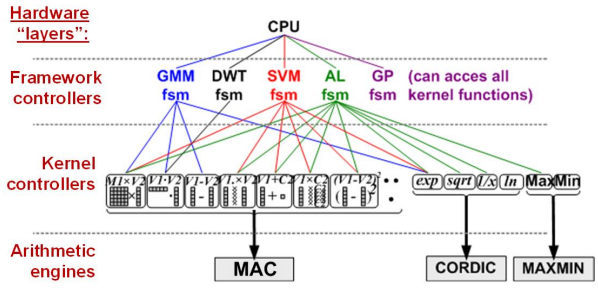

Fig. 3. Visualization of accelerator microarchitecture [2].

With regards to memory-accessing energy, two studies have investigated explicit signal and model encodings to

reduce the amount of memory accessing and also the size of the local cache needed, both of which reduce energy [3,2]. In [3], compressive random projections are employed, which have the property of preserving inner products between vectors. Since inner products serve as a similarity metric for pattern-recognition in a number of inference algorithms, this enables computation entirely on the compressed representations. Initially, such representations increase the memory required due to loss of regularity in the stored co-efficients that results from random projection; however, even at modest compression factors, the overall energy (for both computation and memory accessing) can be substantially reduced. In [2], an explicit compression/decompression accelerator is employed at the local cache. By using low-complexity, lossy encodings, the energy and latency of on-line compression/decompression can be minimized. This is a viable option since machine-learning algorithms for inference have been shown to present substantial tolerance to low-order bit errors. In [2], ADPCM encoding is used, and the compression/decompression accelerator consumes just 8.4pJ per access while achieving 4× compression, compared to over 30pJ per access from the 32kB SRAM (whose energy would be even higher if the size were increased). As an illustration, Table 1 compares the performance of an EEG-based seizure detector both with and without the encoding, showing minimal impact on application-level performance.

Table 1. Seizure detector performance with compression [2].

	w/o Compression		w/ Compression	
	True Pos	True Neg	True Pos	True Neg
Patient 1	96.1%	98.1%	94.1%	98.9%
Patient 2	93.8%	99.7%	93.8%	99.9%
Patient 3	91.7%	98.7%	90.4%	99.4%

Demonstrations. The accelerator-based microprocessors have been demonstrated is a range of medical sensor applications, employing various inference algorithms involving models of various complexity. As an illustration, Table 2 shows the measured power of six applications, along with the savings achieved compared to an implementation using a general-purpose CPU alone [2]. While computations delegated to accelerators are expected to result in substantial savings, *overall* power savings in the range of 3-500× come as a result of exploiting the high-level algorithmic structure, separating the need for programmability from energy efficiency.

Table 2. Measured power of six applications on the accelerator-based microprocessor [2]

	Application	Total Power	Savings w/ accel.
1	ECG-based arrhythmia detector (morphology features)	25.8 µW	497×
2	ECG-based arrhythmia detector (wavelet features)	12.0 µW	71×
3	ECG-based patient-adaptive arrhythmia detector	30.6 µW	419×
4	EEG-based seizure detector	93.6 µW	43×
5	EEG-based sleep-stage detector	3.1 mW	3×
6	Spike-/LFP-based motor intention decoder	6.7 µW	87×

To highlight the algorithmic possibilities raised by such on-patient platforms, we describe in greater detail an application that exploits continuous data acquisition for customization of the patient model by employing an active-learning framework. Since a potentially large amount of data is being acquired, the goal of active learning is to identify a reduced subset that is optimal for customizing the model. Fig. 4 illustrates a patient-adaptive ECG-based cardiac-arrhythmia detector. The detector is seeded with a population-level patient generic model. In addition to arrhythmia detection based on the current model, instances of data are assessed to form a batch consisting of patient-specific data for model training. A key aspect of the detector is the choice of metrics used to asseses the data. For active learning with marginal-distance classifiers (e.g., SVM) used, marginal-distance metrics are commonly employed; however, previous work has shown that diversity metrics are also beneficial for achieving rapid convergence of the model. The microprocessor in [1] thus employs a dedicated accelerator that enables various metrics to be computed in the background and then combined with configurable weighting in order to assess data for inclusion in the batch. Fig. 4 shows the performance of the arrhythmia detector for a particular patient over the course of the model customization process (starting from a low-performance patient-generic model).

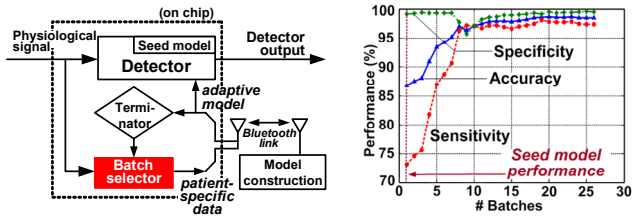

Fig. 4. Patient-adaptive cardiac-arrhythmia detector.

B. *Mixed-signal Interfaces for Signal Acquisition*

As mentioned at the outset, robust signal acquisition poses an important challenge in medical-sensor applications. This arises due to the need for low noise and low nonlinearity to ensure high-quality digital representation for subsequent stages of signal analysis. Previous work has primarily approached this as a *circuit* problem, focusing on signal readout for high SNDR. In this paper, we consider signal readout to preserve the specific information required for particular signal-analysis *algorithms* of interest. While such a perspective remains the subject of on-going research, we describe two possible approaches (in both cases, the final objective is to perform simple inferences, e.g., classification): (1) permit errors in the signal-acquisition stage, and evaluate the extent of information loss incurred through the *subsequent* stages of signal analysis; and (2) move substantial computations *within* the signal-acquisition stage and permit errors, mitigating sensitivity to information loss in subsequent stages of signal-analysis.

The first approach follows from the idea of Data-driven Hardware Resilience (DDHR) [4]. DDHR leverages data-driven training of a classification model by explicitly using training data affected by errors, which arise due to circuit non-idealities. DDHR views such errors as effectively causing perturbations in the data distributions of the classes; thus training to the new distributions results in an 'error-aware model'. Research in DDHR has shown that as long as the

error-affected data maintains mutual information with its class membership, the error-aware model can preserve overall performance even in the presence of severe errors. Though initial application of DDHR focused on faults in digital computational circuits, recent results have shown extension to analog signal-acquisition circuits, as described below.

One challenge with the approach above is that, generally, the information loss incurred in a particular signal-analysis stage depends on the distribution of inputs *and* the precise operations being performed. Thus, even if errors in the signal-acquisition stage preserve information, the resulting data distributions may potentially exhibit increased sensitivity to information loss through the subsequent signal-analysis stages. In [5], a matrix-multiplying ADC (MMADC) is presented that attempts to combine the majority of computations required for feature extraction and classification *within* the A-D conversion process. While the robustness this affords to information loss requires further investigation, energy benefits can be derived from the algorithmic formulation proposed. Namely, in applications employing linear feature-extraction computations (e.g., PCA, DWT, FIR filtering, etc.) classification can be performed without explicitly having to compute the feature vector. This is achieved by formulating the classifier as multiple linear classifiers, which together adequately fit to the training-data distributions (using a machine-learning approach known as boosting). With primarily linear compuations involved, feature extraction and classification can be combined into a single matrix transformation, followed by a small number of thresholding and addition operations. Such a formulation has the potential to substantially reduce the total number of operations, and computation within A-D conversion (via the MMADC) has the potential to further reduce energy.

Demonstrations. The DDHR concept has been demonstrated in an EEG-based seizure detector implemented in 32nm CMOS [6]. The seizure detector performs multiplication operations required for feature extraction in the analog domain using simple current-integrating circuits. Fig. 5 on the left shows the distribution of feature vectors computed using fixed-point digital multiplication following A-D conversion, and on the right shows the distribution of feature vectors computed by the IC in the analog domain. Non-idealities in the analog circuits substantially perturb the data distributions. Nonetheless, using data from a particular patient as an example, the baseline detector has sensitivity of 5/5, latency of 2.0 sec., and false alarms of 8. While the non-idealities degrade this to 5/5, 3.6 sec, and 443, the error-aware model restores performance to 5/5, 3.4 sec., and 4.

The MMADC has been implemented in 130nm CMOS, within a SAR ADC [5]. A major challenge is that multiplication increases the dynamic range required, which has a severe impact on analog energy. To address this, the MMADC implements multiplication in a floating-point, mixed-signal manner. A mantissa, with value between 1 to 2, is applied in the analog domain passively via attenuation in the SAR feedback path. An exponent (base 2), with value from -32 to 31, is applied in the digital domain simply through barrel shifting. Thus, very large valued multipliers can be applied, but the impact on the analog dynamic range is minimal, less than a factor of 2. A number of applications have been demonstrated

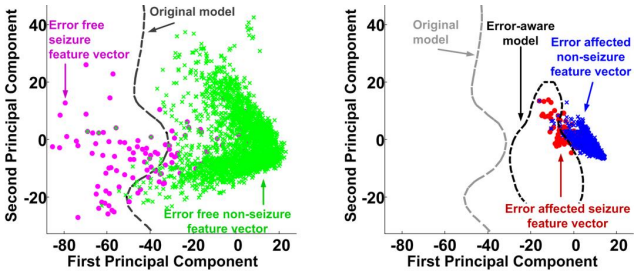

Fig. 5. Feature-vector distributions from seizure detector [6].

with the MMADC; most notably, an ECG-based cardiac-arrhythmia detector, is implemented by formulating feature-extraction and classification into a single matrix multiplication. Compared to a conventional system (ADC, digital feature extractor, and digital SVM classifier), the MMADC implementation reduces energy by over 9×, while achieving similar performance.

IV. Conclusions

A possible extension of data-driven medicine to on-patient platforms is the ability to perform simple inferences based on sensor data. Machine-learning algorithms are an important focus for such platforms, as they enable the creation of robust models for inference from data. Examining the structure, computations, and specific information utilized in such algorithms exposes opportunities to address system-design challenges. For microprocessors, the trade-off between computational energy and programmability can be addressed through accelerator-based architectures, wherein diverse feature-extraction computations are delegated to a general-purpose CPU and inference computations based on high-order data-driven models are delegated to accelerators. For signal acquisition circuits, traditional linearity and noise requirements can be relaxed by focusing on the acquisition of specific information as required in the algorithms. While generalized design methodologies have not been established, prototype demonstrations suggest promising areas for future research.

References

[1] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *J. of Solid-State Circuits*, vol. 48, no. 7, pp. 1625-1637, July 2013.

[2] K. H. Lee and N. Verma, "A low-power microprocessor for data-driven analysis of analytically-intractable physiological signals in advanced medical sensors," *VLSI Symp. Circuits*, June 2013, pp. C250-C251.

[3] M. Shoaib, K. H. Lee, N. K. Jha, and N. Verma, "A 0.6-107 µW energy-scalable processor for directly analyzing compressively-sensed EEG," *IEEE Trans. Circuits and Systems I,* vol. 61, no. 4, pp. 1105-1118, April 2014.

[4] Z. Wang, K. H. Lee, and N. Verma, "Overcoming computational errors in sensing platforms through embedded machine-learning kernels," *IEEE Trans. VLSI Systems,* Aug. 2014.

[5] J. Zhang, Z. Wang, and N. Verma, "A matrix-multiplying ADC implementing a machine-learning classifier directly with data conversion," *Tech. Dig. of Int'l Solid-State Circuits Conf.*, Feb. 2015, pp. 332-333.

[6] J. Zhang, L. Huang, Z. Wang, and N. Verma, "A seizure-detection IC employing machine learning to overcome data-conversion and analog-processing non-idealities," in press.