Reduced Overhead Error Compensation for Energy Efficient Machine Learning Kernels

Sai Zhang, *Student Member, IEEE,* and Naresh R. Shanbhag, *Fellow, IEEE*
Dept. of Electr. & Comput. Eng., Univ. of Illinois at Urbana-Champaign, Urbana, IL, USA

Abstract*—Low overhead error-resiliency techniques such as RAZOR [1] and algorithmic noise-tolerance (ANT) [2] have proven effective in reducing energy consumption. ANT has been shown to be particularly effective for signal processing and machine learning kernels. In ANT, an explicit estimator block compensates for large magnitude errors in a main block. The estimator represents the overhead in ANT and can be as large as 30%. This paper presents a low overhead ANT technique referred to as ALG-ANT. In ALG-ANT, the estimator is embedded inside the main block via algorithmic reformulation and thus completely eliminates the overhead associated with ANT. However, ALG-ANT is algorithm-specific. This paper demonstrates the ALG-ANT concept in the context of a finite impulse response (FIR) filter kernel and a dot product kernel, both of which are commonly employed in signal processing and machine learning applications. The proposed ALG-ANT FIR filter and dot product kernels are applied to the feature extractor (FE) and SVM classification engine (CE) of an EEG seizure classification system. Simulation results in a commercial $45\,\mathrm{nm}$ CMOS process show that ALG-ANT can compensate for error rates of up to 0.41 (errors in FE only), and up to 0.19 (errors in FE and CE) and maintain the true positive rate $p_{tp} > 0.9$ and false positive rate $p_{fp} \leq 0.01$. This represents a greater than 3-orders-of-magnitude improvement in error tolerance over the conventional architecture. This error tolerance is employed to reduce energy via the use of voltage overscaling (VOS). ALG-ANT is able to achieve 44.3% energy savings when errors are in FE only, and up to 37.1% savings when errors are in both FE and CE.

I. Introduction

Error resiliency techniques have been proposed [3] to enhance energy efficiency by reducing design margins and compensating for the resultant errors. Large design margins arise from the need to provide robustness in the presence of process, voltage, and temperature variations [4], and represent an energy overhead as high as $3\times$-to-$4\times$ [5]. The key to the use of error resiliency for energy reduction is that such techniques need to be low overhead and yet effective in compensating for high error rates. Classical fault-tolerance techniques such as N-modular redundancy (NMR) rely on replication of the main computation block (main block) and as a result are ineffective for the purposes of energy reduction. Hence, low overhead error resiliency techniques such as RAZOR [6], [7], error-detection sequential (EDS) [8], confidence driven computing (CDC) [9] and statistical error compensation (SEC) techniques [3] have been proposed to enhance energy efficiency.

Error resiliency techniques have two sources of overhead: error detection and error correction. Techniques such as RAZOR [6], [7], EDS [8], and CDC [9] achieve lower error detection overhead compared with NMR. RAZOR employs a shadow latch for error detection with an error detection overhead of 1.2% (with 28.5% overhead in RAZOR flip flop), EDS [8] employs ideas similar to RAZOR but eliminates the metastability issue, and CDC [9] employs fine-grained temporal redundancy and confidence checking with an error detection overhead of $8\% - 46\%$. However, these techniques employ a roll back based error correction scheme, thus incur error correction overhead of $> 2\times$. Unlike the roll back scheme based techniques, SEC is a class of Shannon-inspired system level error compensation techniques that utilizes signal and error statistics. SEC [3] has been shown to compensate for error rates ranging from 0.21 to 0.89, with an combined error detection and correction overhead ranging from 5% to 30% resulting in energy savings ranging from 35% to 72%.

Algorithmic noise-tolerance (ANT) [10] is a specific SEC technique that has been shown to compensate for high error rates in signal processing and machine learning kernels. For example, the reduced precision replica ANT (RPR ANT) technique [11] was employed to compensate for error rates of 0.58 and 0.21 in a subthreshold ECG processor [12] and a Markov Random Field (MRF) stereo matching kernel [13], respectively, while delivering the required application-level performance. The total overhead in ANT ranges from 5% to 30% [12] due to the use of explicit estimator blocks in error compensation. This overhead, though small compared to other techniques, limits the achievable energy efficiency.

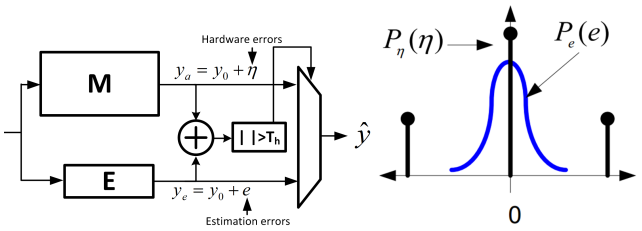
This paper develops a low overhead ANT technique by algorithmically embedding the estimator inside the main block thereby eliminating the estimator overhead completely and leaving only an adder and a mux as the overhead. A drawback of ALG-ANT is that it is algorithm-specific and thus needs to be customized for each algorithm. This paper demonstrates the ALG-ANT concept for two types of commonly employed compute kernels, i.e. an FIR filter kernel and a dot product kernel, and their application in a support vector machine (SVM) EEG seizure classification system consisting of a filter bank as the feature extractor (FE) and linear SVM as the classification engine (CE). Simulation results in a commercial $45\,\mathrm{nm}$ CMOS process show that ALG-ANT can compensate for error rates of up to 0.41 (errors in FE only), and up to 0.19 (errors in FE and CE) and maintain the true positive rate $p_{tp} > 0.9$ and false positive rate $p_{fp} \leq 0.01$. This represents a greater than 3-orders-of-magnitude improvement in error tolerance over the conventional architecture. ALG-ANT is able to achieve energy savings of 44.3% (errors in the FE only) and 37.1% (errors in both the FE and CE) when its error tolerance is employed to compensate for timing errors resulting from voltage overscaling (VOS) [11].

The rest of the paper is organized as follows. Section II describes the background of conventional ANT and the architecture of the SVM EEG seizure classification system. Section III describes the design of the ALG-ANT FIR filter kernel and dot product kernel. Section IV presents the simulation results when the proposed ALG-ANT kernels are applied as the FE and CE of the SVM EEG seizure classification system and compares it with an uncompensated and a retrained system [14]. Conclusions are presented in Section V.

II. Background

A. Conventional ANT

Conventional ANT incorporates a main block (**M**) and an estimator (**E**) (see Fig. 1(a)). In RPR ANT [11], the estimator is obtained by reducing the precision of the main block. The main block is subject to large magnitude errors η (e.g., timing errors due to critical path violations which typically occur in the most significant bits (MSBs)) while the estimator is subject to small magnitude errors e



(a) (b)

Fig. 1. Algorithmic noise-tolerance (ANT): a) conventional architecture, and b) the error statistics in the main (**M**) and estimator (**E**) blocks.

(see Fig. 1(b), e.g., due to quantization noise in the least significant bits (LSBs)), i.e.:

$$y_a = y_o + \eta \tag{1}$$
$$y_e = y_o + e \tag{2}$$

where y_o, y_a, y_e is the error free output, the main block output and the estimator output, respectively. ANT exploits the difference in the error statistics of η and e to detect and compensate errors to obtain the final output $\hat{y}$ as follows:

$$\hat{y} = \begin{cases} y_a \text{ if } |y_a - y_e| \leq T_h \\ y_e \quad \text{otherwise} \end{cases} \tag{3}$$

where T_h is an application dependent threshold parameter to maximize the performance of ANT. The error rate p_η is defined as:

$$p_\eta = 1 - P_\eta(0) = Pr\{\eta \neq 0\}$$

where $P_\eta(\cdot)$ is the error probability mass function (PMF). The errors η are most conveniently obtained by applying voltage overscaling (VOS) where the supply voltage V_{dd} is scaled as follows:

$$V_{dd} = K_{vos} V_{dd-crit}$$

where K_{vos} is the voltage overscaling factor, and $V_{dd-crit}$ is the minimum voltage needed for error free operation in the main block. Note that for the ANT system to work properly, the estimator is not permitted to make large magnitude errors such as those arising from timing violations due to VOS. This helps maintain the difference in the error statistics at the output of the main block and the estimator as shown in Fig. 1(b).

B. EEG classification system using SVM

Portable health monitoring is an important class of applications that can benefit from the design of energy efficient machine learning kernels. It has been shown [15] that epileptic seizures can be efficiently detected by analyzing the EEG signal using an SVM kernel. The EEG seizure classification system [15] (see Fig. 2(a)) consists of a frequency selective filter bank to extract signal energy in the $0 - 20$ Hz range and a SVM classifier. The filter bank has passband of 3 Hz with a transition band of 1.5 Hz. Eight channels are employed to cover the entire frequency range [15].

SVM [16] is a popular supervised learning method for classification and regression. An SVM operates by first training a model (the training phase) followed by classification (the classification phase). During the training phase, feature vectors with labels are used to train a model. During the classification phase, the SVM produces a predictive label when provided with a new feature vector. The SVM training can be formulated as the following optimization problem to determine the maximum margin classifier [16] (see Fig. 2(b)):

$$\begin{aligned} &\min \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ &s.t. \\ &y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i - \mathrm{b}) \geq 1 - \xi_i \\ &\xi_i \geq 0 \end{aligned} \tag{4}$$

where C is the cost factor, ξ_i is the soft margin, $\mathbf{x}_i$ is a feature vector, y_i is the label corresponding to the feature vector $\mathbf{x}_i$, $\mathbf{w}$ is the weight vector and b is the bias. The trained model is represented as:

$$y = \mathbf{w}_o^{\mathbf{T}}\mathbf{x} - b \tag{5}$$

where $\mathbf{w}_o$ are the optimized weights. It can be shown that the optimum weights are represented as a linear combination of the feature vectors that lie on the margins (see Fig. 2(b)), i.e., support vectors:

$$\mathbf{w}_o = \sum_{i=1}^{N_s} \alpha_i y_i \mathbf{x}_{s,i}^T \tag{6}$$

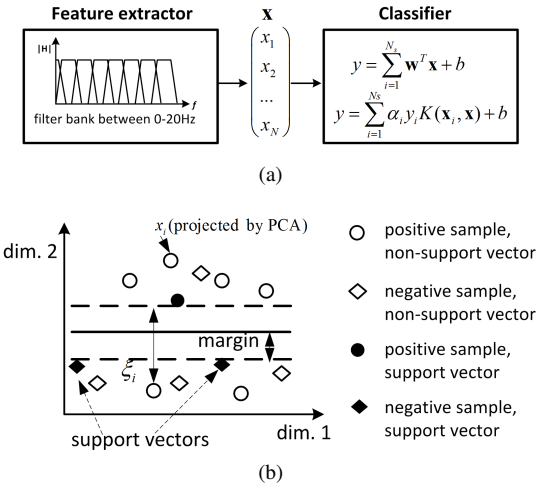
where N_s and $\mathbf{x}_{s,i}$ are the number of support vectors and ith support vector, respectively. The linear model can thus be represented by:

$$y = \sum_{i=1}^{N_s} \alpha_i y_i \mathbf{x}_{s,i}^T \mathbf{x} - b \tag{7}$$

The linear SVM in (7) can be easily extended into non-linear SVM by employing the kernel trick [16], resulting in:

$$y = \sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{x}_{s,i}, \mathbf{x}) - b \tag{8}$$

where $K(\mathbf{x}_{s,i}, \mathbf{x})$ is a kernel function. Popular kernel functions include polynomial, radial basis function (RBF), and others.



(a)

(b)

Fig. 2. EEG seizure classifier with SVM: a) system architecture, and b) the SVM principle.

III. Proposed ALG-ANT Technique

ALG-ANT reuses part of the **M**-block to generate an estimate of its error free output y_o. This is in contrast to conventional ANT technique (see Section II-A) which requires an explicit estimator. Such embedding of the estimator can be performed at the algorithmic level. We next derive two ALG-ANT techniques - one for the FIR filter kernel and another for the dot product kernel.

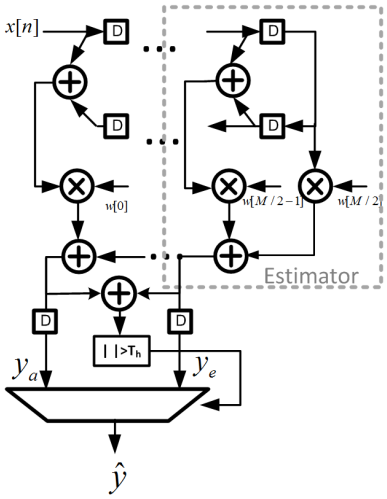


Fig. 3. ALG-ANT FIR filter structure.

A. ALG-ANT FIR Filter Kernel

The FIR filter is a commonly used kernel in signal processing and machine learning. The conventional FIR filter design method employs algorithms such as the weighted least square (WLS) method [17], which formulates the filter design as an optimization problem. Let $H(e^{j\omega})$ and $H_d(e^{j\omega})$ denote the designed and the ideal filter frequency responses, respectively, and $W(e^{j\omega})$ be a non-negative error weighting function. The WLS method minimizes the L_2 norm of the weighted difference between $H(e^{j\omega})$ and $H_d(e^{j\omega})$ as follows:

$$\min \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[W(e^{j\omega})H(e^{j\omega}) - W(e^{j\omega})H_d(e^{j\omega})\right]^2 d\omega \qquad (9)$$

Let $\mathbf{h} = [h[0]...h[M]]^T$, $\mathbf{d} = [d[0],...d[N-1]]^T$ be the pulse response of the $M+1$-tap filter $H(e^{j\omega})$ and the IDFT of $W(e^{j\omega})H_d(e^{jw})$, respectively, and let the N by $M+1$ matrix $\mathbf{W}$ be defined as $\mathbf{W}[n,l] = w[n-l]$ where $w[n]$ is the IDFT of $W(e^{j\omega})$. The optimization can be reduced to:

$$\min \|\mathbf{Wh} - \mathbf{d}\|^2 \qquad (10)$$

and has solution $\mathbf{h}^* = \mathbf{W}^{\dagger}\mathbf{d}$, where $\mathbf{W}^{\dagger}$ is the Moore-Penrose pseudo inverse.

In ALG-ANT, the optimization in (10) is modified to include architectural level constraints. In particular, we employ the filter architecture in Fig. 3 where the center $M+1-2K_f$ filter taps are employed to obtain the estimator output $y_e[n]$ (see Fig. 1(a)). Here K_f is a design parameter that determines the estimator length. The rationale for using the center taps of an FIR filter to obtain an estimate of its final output $y_o[n]$ is that for linear phase FIR filter, the center taps of the filter can provide a good estimate of the filter response [18]. Doing so embeds the estimator completely into the main block. To achieve this, we reformulate the objective function in (10) as follows:

$$\min(1-\gamma)\|\mathbf{Wh} - \mathbf{d}\|^2 + \gamma\left\|\tilde{\mathbf{W}}\mathbf{h}\right\|^2 \qquad (11)$$

where $\tilde{\mathbf{W}} = \begin{bmatrix} I_{K_f\times K_f} & 0_{K_f\times\hat{K}_f} & 0_{K_f\times K_f} \\ 0_{\hat{K}_f\times K_f} & 0_{\hat{K}_f\times\hat{K}_f} & 0_{\hat{K}_f\times K_f} \\ 0_{K_f\times K_f} & 0_{K_f\times\hat{K}_f} & I_{K_f\times K_f} \end{bmatrix}$, $\hat{K}_f = M + 1 - 2K_f$ and the parameter γ ($0 \le \gamma < 1$) is used to control the relative strength of the two optimization terms. Doing so constrains the magnitude of the outer taps of the filter. The optimization in (11) can be solved by setting the derivative of the loss function in (11) to zero, resulting in the following filter:

$$\mathbf{h}^* = ((1-\gamma)\mathbf{W}^T\mathbf{W} + \gamma\tilde{\mathbf{W}}^T\tilde{\mathbf{W}})^{-1}((1-\gamma)\mathbf{W}^T\mathbf{d}) \qquad (12)$$

where $\mathbf{h}^*$ is the optimum ALG-ANT filter coefficients. In practice, γ and K_f are design parameters that can be employed to trade off the two optimization terms in (11). A large γ will weigh more on the estimator design, leading to an more accurate estimator. However, a large γ tends to decrease the performance of the main block since the resulting filter deviates from the ideal filter $\mathbf{d}$. A small K_f (thus larger estimator length) will lead to a more accurate estimator because more coefficients can be employed, but a small value of K_f will limit the amount by which VOS can be applied before the estimator begins to exhibit large magnitude timing violations. Thus, in this design, as expected, the accuracy of the estimator and the main block trade-off with each other, and so does the extent of VOS that can be applied. In section IV, K_f is determined by the error rate p_η (thus K_{vos}) and γ is optimized via a grid search.

B. ALG-ANT Dot Product Kernel

We next derive ALG-ANT for the dot product kernel, another widely used kernel in machine learning. The dot product kernel is employed in the linear SVM (see Fig. 2(a)), which provides good classification performance and results in a particularly simple architecture [19]. In the dot product kernel (see Fig. 4), the input vector $\mathbf{x}$ and the weight vector $\mathbf{w}$ are multiplied element-wise and the resulting products are added up.

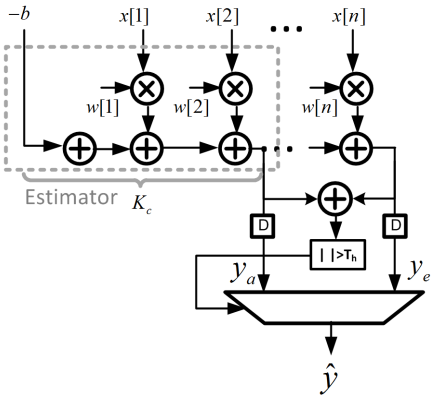


Fig. 4. ALG-ANT linear SVM: the dot product result is unaltered when the order of the multiply-accumulates (MACs) is varied.

One observation in (5) is that it is only the final dot product that contributes to the classification result, not the order in which computation is done. This suggests that we can implement the dot product kernel via dimension reordering (DR) which will reorder the

dimensions of the inputs $\mathbf{x}$ and $\mathbf{w}$ in the classification engine and use more important weights first during the dot product evaluation.

The reordered weight vector $\hat{\mathbf{w}}$ can be calculated via a simple sorting operation:

$$\hat{\mathbf{w}} = [\hat{w}_1, \hat{w}_2, ...\hat{w}_n]$$

where $|\hat{w}_i| \geq |\hat{w}_j|$ for $i < j$. In other words, we reorder the calculation of the dot product according to the importance of weights w_i in these dimensions. The resulting incremental refinement architecture enables us to employ the intermediate stage output as the estimator output $y_e[n]$, as shown in Fig. 4. As the estimator length K_c increases, the classification results will improve but the extent to which VOS can be applied will reduce. This trade-off is explored in the next section.

IV. Simulation Results

To evaluate the performance of ALG-ANT technique, we apply ALG-ANT FIR filter and dot product kernel to the FE and CE of the SVM EEG seizure classification system. This section presents the design trade-offs in the proposed ALG-ANT technique, and the simulation results in a $45\,\mathrm{nm}$ CMOS process when ALG-ANT is applied to the EEG classification system.

A. Evaluation Methodology

Figure 5(a) shows the evaluation methodology (similar to [13], [20]) employed to quantify system-level performance metrics and to estimate system-level energy consumption that integrates circuit, architecture, and system level design variables. The methodology consists of two parts: 1) *system-level error injection*, and 2) *system-level energy estimation*. Comparison of the proposed ALG-ANT with conventional approach is done in a commercial $45\,\mathrm{nm}$ CMOS process.

System-level error injection is done as follows:

1) characterize delay vs. V_{dd} of basic gates such as AND and XOR using HSPICE for $0.2\,\mathrm{V} \leq V_{dd} \leq 1.2\,\mathrm{V}$.
2) develop structural Verilog HDL models of key kernels needed in the EEG classification system using the basic gates characterized in Step 1. These kernels are a $12\,\mathrm{b}$ input, $8\,\mathrm{b}$ coefficient, and $16\,\mathrm{b}$ output, 44-tap FIR filter bank (used in the FE) and a $8\,\mathrm{b}$ input, $8\,\mathrm{b}$ coefficient, and $19\,\mathrm{b}$ output, 120-dimension dot-product kernel (used in the SVM CE).
3) HDL simulations of these kernels were conducted at different voltages by including the appropriate voltage-specific delay numbers obtained in Step 1 into the HDL model. The clock frequency is fixed at $76\,\mathrm{MHz}$ (error free frequency at $V_{dd} = 1.2\,\mathrm{V}$) for all voltages to emulate VOS. The error probability mass functions (PMFs) of these kernels and error rates p_η are obtained for different supply voltages (and thus voltage overscaling factor K_{vos}).
4) System performance is evaluated by injecting errors into a fixed point MATLAB-model of the EEG classification system. The errors are obtained by sampling the error PMFs obtained in Step 3.

Figure 5(b) shows the error PMF of the 44-tap low pass FIR filter used in the FE at $V_{dd} = 0.9\,\mathrm{V}$ ($f_{clk} = 76\,\mathrm{MHz}$) which corresponds to a $K_{vos} = 0.75$ and an error rate $p_\eta = 0.27$. Figure 5(c) shows that the error rate p_η increases from 10^{-4} to 0.98 as the voltage scales down from $1.2\,\mathrm{V}$ to $0.7\,\mathrm{V}$.

System-level energy estimation is done as follows:

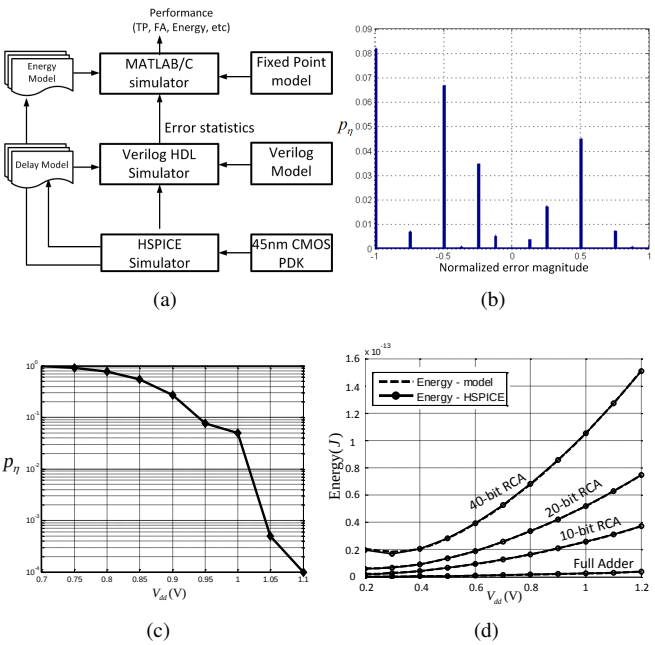


Fig. 5. Evaluation methodology: a) simulation setup, and b) error PMF at $V_{dd} = 0.9\,\mathrm{V}$, c) error rate p_η vs. V_{dd} for the 44-tap low pass filter employed in the FE, the CHB-MIT EEG data set [14] is employed as the input, and d) comparison of the energy model and HSPICE simulations in a $45\,\mathrm{nm}$ CMOS process.

1) Obtain a full adder (FA) count N_{FA} of the kernel being analyzed.
2) Conduct a one-time characterization of the energy consumption of a FA incorporating both dynamic and leakage energies as follows [20]:

$$E_{FA} = C_{FA}V_{dd}^2 + V_{dd}I_{leak}(V_{dd})\frac{1}{f_{clk}} \tag{13}$$

with

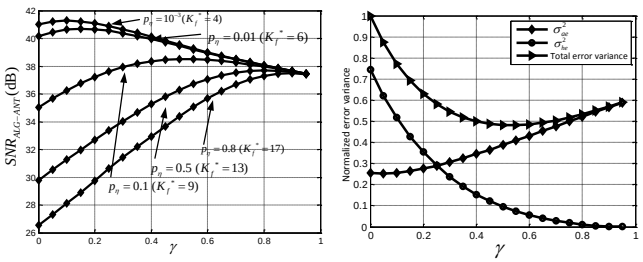
$$I_{leak}(V_{dd}) = \mu C_{ox}\frac{W}{L}(m-1)V_T^2 e^{\frac{-V_t}{mV_T}} e^{\frac{-\eta_d V_{dd}}{mV_T}}(1 - e^{\frac{-V_{dd}}{V_T}}) \tag{14}$$

where C_{FA} is the effective load capacitance of the FA and is extracted from HSPICE, V_{dd} is the supply voltage, V_t, V_T, μ, C_{ox}, and η_d are the threshold voltage, the thermal voltage, the carrier mobility, the gate capacitance per unit W/L, and the drain induced barrier lowering (DIBL) coefficient, respectively, obtained from the process files, and m is a constant related to the sub-threshold slope factor and is a fitting parameter.
3) The energy estimate of the kernel is obtained as $E_{op} = N_{FA}E_{FA}$.

Figure 5(d) shows the modeling results of the FA and ripple carry adder (RCA) of various bit width to show the accuracy and scalability of the energy model. The energy model is within 5% (for $0.2\,\mathrm{V} \leq V_{dd} \leq 1.2\,\mathrm{V}$) of circuit simulation results.

B. Design Optimization

The FE and CE in an ALG-ANT based system have a number of design parameters that need to be selected for optimal system performance. For the FE, (11) indicates that K_f determines the estimator complexity. Hence, K_f places a lower bound on the supply voltage because the estimator needs to be free of timing violations. Similarly, γ indicates how closely the main block approximates the



(a) (b)

Fig. 6. ALG-ANT applied to the filter design problem: a) $SNR_{ALG-ANT}$ vs. γ for various p_η, and b) approximation error σ_{ae}^2, hardware error σ_{he}^2 and total error $(\sigma_{ae}^2 + \sigma_{he}^2)$ vs. γ at $p_\eta = 0.1$, the CHB-MIT EEG data set [14] is employed as input, error variances are normalized w.r.t. total error variance at $\gamma = 0$.

ideal frequency response. Thus, the accuracy of the estimator and the main block trade-off with each other, which suggests that an optimum value for γ and K_f, i.e., γ^* and K_f^*, exists. To explore the trade off between main block and estimator performance, the application level ALG-ANT filter SNR is defined. Let y_o, $\hat{y}$ denote the error free main filter output and ALG-ANT filter output, and let y_d denote the error free ideal filter (with coefficient $\mathbf{d}$) output. The ALG-ANT filter SNR is defined as

$$SNR_{ALG-ANT} = 10log_{10}(\frac{\sigma_{y_o}^2}{\sigma_{ae}^2 + \sigma_{he}^2}) \tag{15}$$

where $\sigma_{ae}^2 = E(y_o - y_d)^2$ is the variance of approximation error and $\sigma_{he}^2 = E(y_o - \hat{y})^2$ is the variance of hardware error.

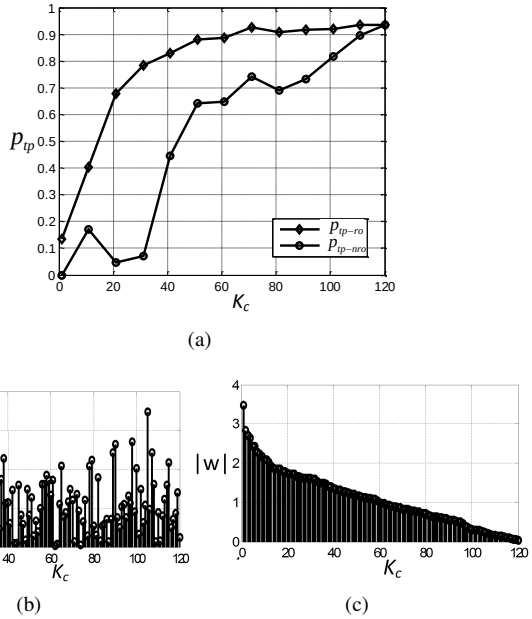
In order to determine these SNR-optimum values, K_f^* is first determined by choosing the maximum estimator length at a given supply voltage V_{dd}, and hence error rate p_η (thus K_{vos}). In particular, K_f^* increases with p_η as shown in Fig. 6(a). Next, γ^* is obtained via sweeping its value and observing the $SNR_{ALG-ANT}$. Figure 6(a) shows that when p_η is low, i.e. K_f^* is small, γ^* is small because the approximation error σ_{ae}^2 dominates. On the other hand, when p_η is high, i.e., K_f^* is large, $\gamma^*$is large because the hardware error σ_{he}^2 dominates, and the optimization procedure will strive for a more accurate estimator, as shown in (11). Figure 6(b) shows this trade-off for a specific value of p_η, where it can be seen that as γ increases, σ_{ae}^2 increases because the overall filter no longer minimizes the difference between ideal filter and main block; at the same time, σ_{he}^2 decreases because the estimator gives better approximations.

For the linear SVM, DR is applied. We employ the CHB-MIT EEG data set [14] to train the SVM and use leave-one-out cross validations to evaluate the classifier performance. The system performance metric employed is the true positive (TP) rate p_{tp} and false positive/alarm (FP) rate p_{fp}, defined as:

$$p_{tp} = \frac{TP}{TP+FN}$$
$$p_{fp} = \frac{FP}{FP+TN}$$

where TP, FN, FP, and TN are the number of true positives, false negatives, false positives, and true negatives, respectively. A good classifier achieves high values of p_{tp} (> 0.9) at a small constant false alarm rate p_{fp} (< 0.01).

Figure 7 studies the impact of DR in the SVM classifier in an error free condition and a $p_{fp} \leq 0.01$. It indicates that the TP rate in the



(a) (b) (c)

Fig. 7. Comparison of classification results with and without DR, feature vectors extracted with FE are employed as input: a) p_{tp} (with $p_{fp} \leq 0.01$) of the SVM classifier vs. estimator length K_c where the estimator is directly obtained by using the first K_c taps of the dot product kernel; the results with DR are denoted as p_{tp-ro}, while the results of directly using the reduced dimension classifier is denoted as p_{tp-nro}, and b) the weights without DR, and c) the weights with DR.

absence of DR (p_{tp-nro}) increases non-monotonically with K_c (the estimator complexity). In particular, $p_{tp-nro} \leq 0.5$ for $K_c \leq 45$, and $p_{tp-nro} \geq 0.9$ only when $K_c \geq 112$. In contrast, when DR is employed the TP rate p_{tp-ro} increases monotonically with K_c, and $p_{tp-ro} \geq 0.9$ when $K_c \geq 64$, which is 43% smaller than when DR is not used. Note that DR needs to be performed only once during the training and thus does not incur overhead during classification. Figure 7(b) shows that without DR, the large magnitude weights are scattered across the dimensions, leading to poor classification results unless the value of K_c is sufficiently large. DR uses the important weights first (see Fig. 7(c)), and thus can produce acceptable results with much smaller values of K_c.

ef

C. ALG-ANT System Performance

The system architecture of the SVM EEG classification system is shown in Fig. 8 where the feature extractor employs the design parameters from [14], [20], with an input of 12 b for the filter bank and 8 b bit for the SVM classifier. Three architectures are considered: 1) the conventional classifier (denoted as CONV), the classifier with retraining [14] (denoted as RETRAIN), and the classifier with ALG-ANT (denoted as ALG-ANT). In the retraining method [14], the classifier is trained with features extracted in the presence of VOS errors. Unlike in the retraining method [14] where CE needs to be error free, ALG-ANT can tolerate errors in both FE and CE. Therefore, two setups are considered in our experiment: 1) errors in FE only and 2) errors in both FE and CE. The max value of error rate p_η for which $p_{tp} > 0.9$ and $p_{fp} < 0.01$ is referred to as the error tolerance metric $p_{\eta-max}$ of the architecture. In the first setup, $p_{\eta-max}$ is the error rate in the FE, and in the second setup, $p_{\eta-max}$ is the maximum of the error rate in FE and CE.

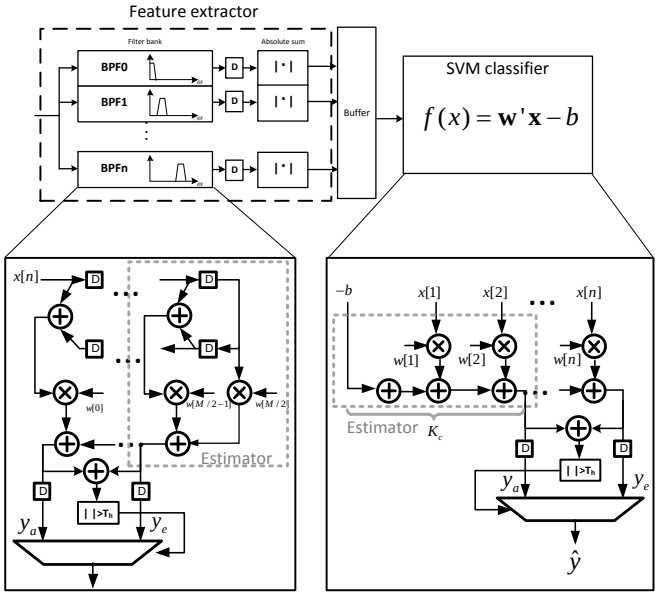


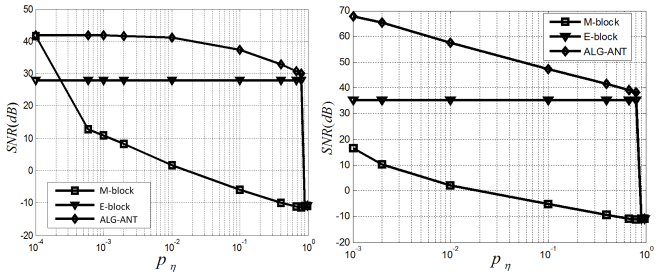
Fig. 8. ALG-ANT based SVM EEG classification system architecture.

Figure 9(a) shows improvement in the application level SNR achieved by ALG-ANT at the FE output. The application level SNR of the **M**-block and the **E**-block is defined as:

$$SNR_a = 10log1_{10}(\frac{\sigma_{y_o}^2}{\sigma_{ae}^2 + \sigma_\eta^2})$$
$$SNR_e = 10log_{10}(\frac{\sigma_{y_o}^2}{\sigma_{ae}^2 + \sigma_e^2})$$

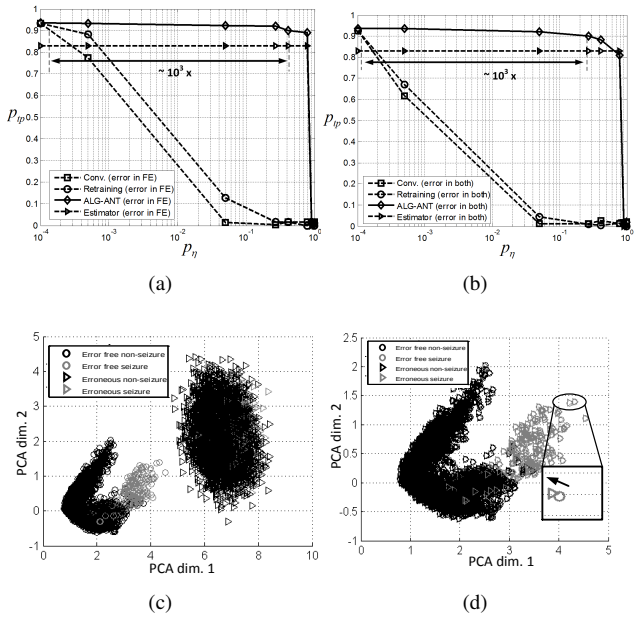
where σ_η^2 and σ_e^2 are the variances of the **M**-block error η and **E**-block error e, respectively, and σ_{ae}^2 is the approximation error (see (15)). As shown in Fig. 9, SNR_a drops sharply from $41\,\text{dB}$ to less than $-10\,\text{dB}$ as the error rate p_η increases, while SNR_e remains constant at $27.8\,\text{dB}$ for $p_\eta \leq 0.78$. In contrast, $SNR_{ALG-ANT}$ (see (15)) is higher than both the **M**-block and the **E**-block, and approaches the **E**-block SNR as p_η increases. To capture the effect of hardware errors on system performance, we also define hardware SNR for the **M**-block, **E**-block and ALG-ANT system as $\hat{SNR}_a = 10log_{10}(\frac{\sigma_{y_o}^2}{\sigma_\eta^2})$, $\hat{SNR}_e = 10log_{10}(\frac{\sigma_{y_o}^2}{\sigma_e^2})$ and $\hat{SNR}_{ALG-ANT} = 10log_{10}(\frac{\sigma_{y_o}^2}{\sigma_{he}^2})$ where σ_{he}^2 is defined in (15). Figure 9(b) shows that $\hat{SNR}_a$ drops from $18\,\text{dB}$ to less than $-10\,\text{dB}$, and $\hat{SNR}_e$ remains constant at $36\,\text{dB}$ for $p_\eta \leq 0.78$, while $\hat{SNR}_{ALG-ANT}$ is greater than both the $\hat{SNR}_a$ and $\hat{SNR}_e$. $\hat{SNR}_{ALG-ANT}$ approaches the $\hat{SNR}_e$ as p_η increases.

Figure 10(a) shows that when errors are in FE only, p_{tp} for the conventional system drops sharply and $p_{\eta-max}$ is as low as 1.5×10^{-4}. Retraining does slightly better as the classifier is retrained to adapt to the error affected features. However, the error tolerance $p_{\eta-max}$ is below 10^{-3}. This is most likely due to the fact that unlike stuck-at-faults studied in [14], timing errors due to VOS are dynamic and depend on the state of circuit. In contrast, when ALG-ANT is applied, p_{tp} has a graceful degradation as p_η increases and $p_{\eta-max}$ is improved to 0.41. The performance of the conventional ANT system was found to be similar to the ALG-ANT system, and thus is not shown. Figure 10(a) also shows that the p_{tp} is always lower than



(a) (b)

Fig. 9. SNR at the output of the FE: a) application SNR, and b) hardware SNR.



(a) (b)

(c) (d)

Fig. 10. Simulation results: a) p_{tp} of conventional, retraining, and ALG-ANT classifier with $p_{fp} \leq 0.01$ when errors are in feature extractor only, b) p_{tp} of conventional, retraining, and ALG-ANT system with $p_{fp} \leq 0.01$ when errors are in both the feature extractor and the classifier, c) PCA results of error free and erroneous features for conventional classifier, and d) PCA results of error free and erroneous features for ALG-ANT classifier.

0.9 when only estimator is employed. Similarly, when errors present in both FE and CE (see Fig. 10(b)), ALG-ANT classifier achieves $p_{\eta-max} = 0.19$. These are both 3-orders-of-magnitude greater than the existing systems.

Principle component analysis (PCA) is performed on the feature vectors to understand the reason why conventional system fails and ALG-ANT is able to maintain good performance. Figure 10(c) shows that in the conventional system, circuit errors have two effects on the feature vectors: 1) errors make it harder to separate the positive and negative samples, and 2) the entire feature space is shifted. The SVM fails to correctly perform classification without knowledge of the error statistics. Figure 10(d) shows the large magnitude error is compensated and converted to small residual errors when ALG-ANT is applied, which will cause a very small shift in the feature space. As a result, the SVM classifier can still perform correct classification.

Table I compares the error tolerance $p_{\eta-max}$, feature extraction energy/feature (E_F), and classification energy/decision (E_C) of three classifiers. When VOS is applied for energy savings, compared with

TABLE I
PERFORMANCE AND ENERGY COMPARISON

	Errors in FE only		Errors in both FE and CE		
	$p_{\eta-max}$	Energy Savings in FE	$p_{\eta-max}$	Energy Savings in FE	Energy Savings in CE
Conventional	1.5×10^{-4}	NA	10^{-4}	NA	NA
Retraining	3×10^{-4}	5.1%	10^{-4}	0	0
ALG-ANT	0.41	44.3%	0.19	37.1%	36.9%

the conventional classifier, the ALG-ANT classifier is able to achieve 44.3% energy savings when errors are in FE only. When both FE and CE are in error, the ALG-ANT classifier is able to achieve 37.1% and 36.9% energy savings in the FE and CE, respectively. The energy savings are due to: 1) the elimination of explicit estimator, and 2) the scaling of supply voltage.

V. CONCLUSIONS

In this paper, we propose ALG-ANT, where the estimator is embedded into the main block via algorithm level transforms, resulting in a low overhead error-resilient architecture. The effectiveness of the proposed ALG-ANT technique is demonstrated through the design of a SVM EEG seizure classification system where simulation results in a commercial $45\,\mathrm{nm}$ CMOS show that ALG-ANT achieves up to 0.41 error tolerance (3-orders-of-magnitude greater than conventional approaches) and up to 44.3% energy savings over conventional systems.

ALG-ANT takes advantage of the fact that many algorithms can be reformulated into an incremental refinement form and thus can employ an intermediate result as the estimator output. Though this approach is algorithm specific, we have demonstrated that key signal processing and machine learning kernels can be reformulated into an ALG-ANT structure. It is of great interest to integrate ALG-ANT and other SEC techniques into a CAD flow to enable the design of Shannon-inspired error-resilient systems.

ACKNOWLEDGMENT

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

REFERENCES

[1] S. Das, D. Blaauw, D. Bull, K. Flautner, and R. Aitken, "Addressing design margins through error-tolerant circuits," in *Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE*, July 2009, pp. 11–12.

[2] R. Hegde and N. Shanbhag, "A voltage overscaled low-power digital filter IC," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 2, pp. 388–391, Feb 2004.

[3] N. Shanbhag, R. Abdallah, R. Kumar, and D. Jones, "Stochastic computation," in *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, June 2010, pp. 859–864.

[4] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Design Automation Conference, 2003. Proceedings*, June 2003, pp. 338–342.

[5] R. Bahar, J. Mundy, and J. Chen, "A probabilistic-based design methodology for nanoscale computation," in *Computer Aided Design, 2003. ICCAD-2003. International Conference on*, Nov 2003, pp. 480–486.

[6] D. Ernst, N. S. Kim, and et al., "Razor: a low-power pipeline based on circuit-level timing speculation," in *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, Dec 2003, pp. 7–18.

[7] D. Blaauw, S. Kalaiselvan, K. Lai, W.-H. Ma, S. Pant, C. Tokunaga, S. Das, and D. Bull, "Razor ii: In situ error detection and correction for pvt and ser tolerance," in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, Feb 2008, pp. 400–622.

[8] J. Tschanz, K. Bowman, C. Wilkerson, S.-L. Lu, and T. Karnik, "Resilient circuits; enabling energy-efficient performance and reliability," in *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, Nov 2009, pp. 71–73.

[9] C.-H. Chen, D. Blaauw, D. Sylvester, and Z. Zhang, "Design and evaluation of confidence-driven error-resilient systems," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 22, no. 8, pp. 1727–1737, Aug 2014.

[10] R. Hegde and N. Shanbhag, "Energy-efficient signal processing via algorithmic noise-tolerance," in *Low Power Electronics and Design, 1999. Proceedings. 1999 International Symposium on*, Aug 1999, pp. 30–35.

[11] B. Shim, S. Sridhara, and N. Shanbhag, "Reliable low-power digital signal processing via reduced precision redundancy," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, no. 5, pp. 497–510, May 2004.

[12] R. Abdallah and N. Shanbhag, "An energy-efficient ecg processor in 45-nm cmos using statistical error compensation," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 11, pp. 2882–2893, Nov 2013.

[13] J. Choi, E. Kim, R. Rutenbar, and N. Shanbhag, "Error resilient mrf message passing architecture for stereo matching," in *Signal Processing Systems (SiPS), 2013 IEEE Workshop on*, Oct 2013, pp. 348–353.

[14] N. Verma, K. H. Lee, K. J. Jang, and A. Shoeb, "Enabling system-level platform resilience through embedded data-driven inference capabilities in electronic devices," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 5285–5288.

[15] J. Yoo, L. Yan, D. El-Damak, M. Altaf, A. Shoeb, and A. Chandrakasan, "An 8-channel scalable eeg acquisition soc with patient-specific seizure classification and recording processor," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 1, pp. 214–228, Jan 2013.

[16] V. Vapnik, "An overview of statistical learning theory," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 988–999, Sep 1999.

[17] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing (2Nd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[18] J. Ludwig, S. Nawab, and A. Chandrakasan, "Low-power digital filtering using approximate processing," *Solid-State Circuits, IEEE Journal of*, vol. 31, no. 3, pp. 395–400, Mar 1996.

[19] K. Lee, S.-Y. Kung, and N. Verma, "Low-energy formulations of support vector machine kernel functions for biomedical sensor applications," *Journal of Signal Processing Systems*, vol. 69, no. 3, pp. 339–349, 2012.

[20] S. Zhang and N. Shanbhag, "Embedded error compensation for energy efficient dsp systems," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, Dec 2014, pp. 30–34.