ELectronics EXpress

LETTER

IEICE Electronics Express, Vol.15, No.7, 1–10

An energy-efficient parallel VLSI architecture for SVM classification

Yin Xu[1], **Zhijian Chen**[1a)], **Xiaoyan Xiang**[2], **and Jianyi Meng**[2]

[1] *Institute of VLSI Design, Zhejiang University, Hangzhou, China*

[2] *State Key Laboratory of ASIC and System, Fudan University, Shanghai, China*

a) *chenzj@vlsi.zju.edu.cn*

Abstract: This letter presents an energy-efficient VLSI architecture for SVM classification. Instead of accurate calculation, cost-reduced computing elements based on approximative techniques are designed to complete computation-intensive operations in the SVM-based classifier to save energy and resources. Besides, a partial parallel structure is applied to eliminate dimensional constraints for inputs of classifiers and balance between classification speed and energy consumption. We adopt 55-nm CMOS process to implement the proposed design. It occupies 0.0901 mm^2 area and consumes 15.9 mW at operating frequency of 100 MHz and from an operating voltage of 1 V. Experiment shows that the design provides an area reduction by 41.5% and a significant saving in energy efficiency by 61.8% compared with the baseline model.

Keywords: SVM, energy-efficient, cost-reduced element, partial parallel architecture

Classification: Integrated circuits

References

[1] T. Groleat, *et al.*: "Stretching the edges of svm traffic classification with fpga acceleration," IEEE Trans. Netw. Serv. Manag. **11** (2014) 278 (DOI: 10.1109/TNSM.2014.2346075).

[2] C. Kyrkou, *et al.*: "Embedded hardware-efficient real-time classification with cascade support vector machines," IEEE Trans. Neural Netw. Learn. Syst. **27** (2016) 99 (DOI: 10.1109/TNNLS.2015.2428738).

[3] M. A. B. Altaf, *et al.*: "A 16-channel patient-specific seizure onset and termination detection soc with impedance-adaptive transcranial electrical stimulator," IEEE J. Solid-State Circuits **50** (2015) 2728 (DOI: 10.1109/JSSC.2015.2482498).

[4] I. Kukenys and B. McCane: "Classifier cascades for support vector machines," Image and Vision Computing New Zealand (2008) (DOI: 10.1109/IVCNZ.2008.4762088).

[5] C. H. Chen: in *Handbook of Pattern Recognition and Computer Vision*, ed. C. H. Chen (World Scientific, Singapore, 2010) 481.

[6] R. Batuwita and V. Palade: "FSVM-CIL: Fuzzy support vector machines for class imbalance learning," IEEE Trans. Fuzzy Syst. **18** (2010) 558 (DOI: 10.1109/TFUZZ.2010.2042721).

[7] I.-C. Park and T.-H. Kim: "Multiplier-less and table-less linear approximation

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

for square and square-root," IEEE Int'l Conf. on Computer Design (2009) (DOI: 10.1109/ICCD.2009.5413129).
[8] M.-H. Sheu and S.-H. Lin: "Fast compensative design approach for the approximate squaring function," IEEE J. Solid-State Circuits **37** (2002) 95 (DOI: 10.1109/4.974551).
[9] A. Avramović, *et al.*: "An approximate logarithmic squaring circuit with error compensation for dsp applications," Microelectronics J. **45** (2014) 263 (DOI: 10.1016/j.mejo.2014.01.005).
[10] J. Alcala-Fdez, *et al.*: "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," J. of Multiple-Valued Logic and Soft Computing **17** (2010) 255 (DOI: 10.1007/s00500-008-0323-y).
[11] T. Fawcett: "Roc graphs: Notes and practical considerations for researchers," Pattern Recognition Letters **31** (2004) 1 (DOI: 10.1.1.10.9777).
[12] J. C. Wang, *et al.*: "VlSI design for SVM based speaker verification system," IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **23** (2015) 1355 (DOI: 10.1109/TVLSI.2014.2335112).
[13] N. Attaran, *et al.*: "A low-power multi-physiological monitoring processor for stress detection," IEEE Sensor (2016) (DOI: 10.1109/ICSENS.2016.7808776).

1 Introduction

Machine learning is a popular topic in this data-intensive era, as regularities hidden behind big data can be found by using machine learning algorithms in various domains. Among these algorithms, support vector machine (SVM) has been widely used due to its computational efficiency and robustness [1]. However, it becomes difficult to design an energy-efficient SVM-based classification system in many real-world problems, as non-linear classifiers are needed to achieve high classification performance on complex datasets, which leads to producing more support vectors and increasing energy and resources consumption.

Therefore, schemes are proposed to solve the problems aforementioned. Some methods aim to obtain a similar but easy-implementing model by adjusting its parameters, such as parameters of classifiers are rounded off to the nearest power of two values and multipliers in the corresponding hardware design are replaced with shift operations to reduce both area and power [2]. However, it has a negative impact on the performance of classifiers in many applications. Some methods aim to find a simple and suitable classifier in specialized applications, such as two linear SVM-based classifiers instead of a non-linear SVM-based classifier are utilized to enhance both the sensitivity and specificity simultaneously in the patient-specific application [3]. There are also some methods aiming to improve the structure of classifiers, such as a cascaded classifier is put forward for applications where data distribution is between-class imbalanced [2, 4, 5]. In this situation, multiple SVM-based classifiers with various performance are arranged in order according to the computational complexity as well as accuracy. A large amount of data can be classified with less calculation in early stages, which results in significant energy saving compared with monolithic SVM classification [5]. Nevertheless, classification with complex calculation in later stages will be bottleneck if the complexity of the hardware architecture is not reduced.

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

In this letter, we propose a flexible hardware architecture with high-accuracy and low-energy features for the non-linear SVM-based classifier. A new cost-reduced computing component based on approximative techniques is designed to optimize the computation-intensive logic of the architecture by replacing these logic with our proposed one. Furthermore, to improve the flexibility and applicability of the architecture, a partial parallel structure with optimal architecture parameters is adopted in our design to eliminate the dimensional constraint of classifiers' inputs.

2 Overview of support vector machine

SVM is widely applied to many real-world classification applications [6] and it has become an extremely successful discriminative classifier for two-class problems, where it is common to label one class with minor samples as a positive and the other one with major samples as a negative. The goal of the SVM algorithm is to find a hyperplane which has the maximum margin from the two classes and can separate the data samples of the two classes efficiently. To get the classification function of SVMs, informative samples which determine the shape of the hyperplane need to be found and these samples are called support vectors (SVs). In addition, kernel functions and slack variables are applied to satisfy the requirements of solving non-linear classification problems. The most commonly used kernel functions include linear, polynomial, and radial-basis function (RBF). In this letter, RBF-based SVM is chosen in terms of its strong applicability and generality. The final classification decision function is shown as follows:

$$K(x_i, x_j) = exp(-\gamma \|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2) \tag{1}$$

$$y = \sum_{i=1}^{n_{sv}} y_i \alpha_i K(\boldsymbol{x_i}, \boldsymbol{x}) + b \tag{2}$$

where α_i is the Lagrange multiplier, y_i is the class label of a SV, $\mathbf{x}_i$ represents a SV, $\mathbf{x}$ is the input vector, $\mathrm{K}(\mathbf{x}_i, \mathbf{x})$ is the kernel function, and b is the bias.

3 Proposed hardware architecture

The main concern to implement an energy-efficient hardware architecture of the RBF-based SVM classifier is the low-energy realization of RBF. As we can see from Eq. (1), there are two difficulties to achieve the goal, which are square-intensive vector operations and hardware-unfriendly exponential function. In this letter, we design cost-reduced but appropriate components based on approximative techniques to overcome the difficulties aforementioned.

3.1 Squaring function approximation

The squaring function is a widely used fundamental arithmetic operation and its exact result can be obtained by using look-up table (LUT) or multiplier which are the two commonly-used methods. However, effects of circuit complexity become tremendous as the bit-width of inputs increases. Recently, squaring function approximation is presented without a significant influence on the performance of algorithms. Linear approximations with only simple operations such as shift,

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

concatenation, and addition are proposed [7]. A set of recursive boolean equations is put forward to approximate squaring functions [8]. [9] proposes an approximation based on a simple logarithmic interpolation. However, these existing methods are not suitable for the hardware implementation of SVM classifiers, as errors make the classification performance drop heavily. Therefore, we put forward a new compensative approximate approach with lower errors to represent a two's complement n-bit data. The data A and its corresponding squaring function can be represented as $A = -2^{n-1}x_{n-1} + \sum_{i=0}^{n-2} 2^i x_i$ and $A^2 = (A_s + s)^2 = A_s^2 + 2A_s s + s$, where s represents the sign of A, A_s represents the result of a bitwise exclusive OR on each bit of A and s and its representation is $A_s = \sum_{i=0}^{n-2} 2^i a_i$. The basic idea of the approach is to extract one basic part and two compensative parts from the representation of squaring functions. A general outputs of the approximate squarer are simplified as a set of Boolean equations for various bit-width inputs and they are shown as follows:

$$P_{basic} = 2^{2n-3}a_{n-2}a_{n-3} + \sum_{i=2k_1}^{2n-4} 2^i a_{k_1}(\bar{a}_{k_1-1} \oplus (a_{k_1-1}a_{k_1-2})) + \sum_{i=2k_1-1}^{2n-5} 2^i a_{k_1-2}(a_{k_1-1} \oplus a_{k_1}) + 2^2 a_1 \bar{a}_0 + a_0$$
$$k_1 = 2, 3, 4 \ldots \tag{3}$$

$$P_{com1} = 2^{2n-5}a_{n-2}a_0a_1 \cdots a_{n-5}s + \sum_{i=n}^{2n-6} 2^i a_{n-2}(a_{k_2+1} \oplus (a_0 \cdots a_{k_2}s)) + \sum_{i=4}^{n-1} 2^i a_{i-1}(a_0 \oplus s) + 2^3 a_2 s + 2^2 a_1 s + 2a_0 s + s$$
$$k_2 = 0, 1, 2, 3 \ldots \tag{4}$$

$$P_{com2} = 2^{2n-7}a_{n-3}a_0a_1 \cdots a_{n-6} + \sum_{i=n}^{2n-8} 2^i a_{n-3}(a_{k_3+1} \oplus (a_0 \cdots a_{k_3}s)) + \sum_{i=6}^{n-1} 2^i a_{i-2}(a_1 \oplus (a_0 s)) + 2^5 a_3 a_0 s$$
$$k_3 = 1, 2, 3 \ldots \tag{5}$$

$$A^2 \approx \tilde{A}^2 = P_{basic} + P_{com1} + P_{com2} \tag{6}$$

where n is greater than or equal to 7. When n is less than 5, only P_{basic} is used, P_{com1} and P_{com2} are used to approximate the squaring function otherwise.

3.2 Exponential approximation

The dedicated hardware implementation of exponential functions is required to satisfy the energy and resources constraints. In the literature, a number of works have been proposed. The Taylor series expansion is one of the oldest and most widely used methods. However, higher order factorial needs to be calculated when higher accuracy is wanted, which will result in more energy and resource consumption. An alternative method is to use LUT, but it is limited by the range of inputs and accuracy of results.

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

In this section, we apply two steps to get an approximation for the exponential function. Firstly, the exponential term is converted to a power of 2. Secondly, piecewise linear approximations are used to approximate a power of 2 where the input range is split into several segments and each of them is linearly approximated. The approximation works well as all inputs are negative values in SVM classification and detailed representations are shown as follows:

$$exp(x) = 2^{xlog_2 e} \approx 2^{1.5x} \tag{7}$$

$$2^y = 2^{floor(y)+delt} \approx 2^{floor(y)}(delt + 1) \tag{8}$$

where y is equal to $1.5x$, $floor(y)$ means the largest integer not greater than y, and $delt$ is the fractional part of y.

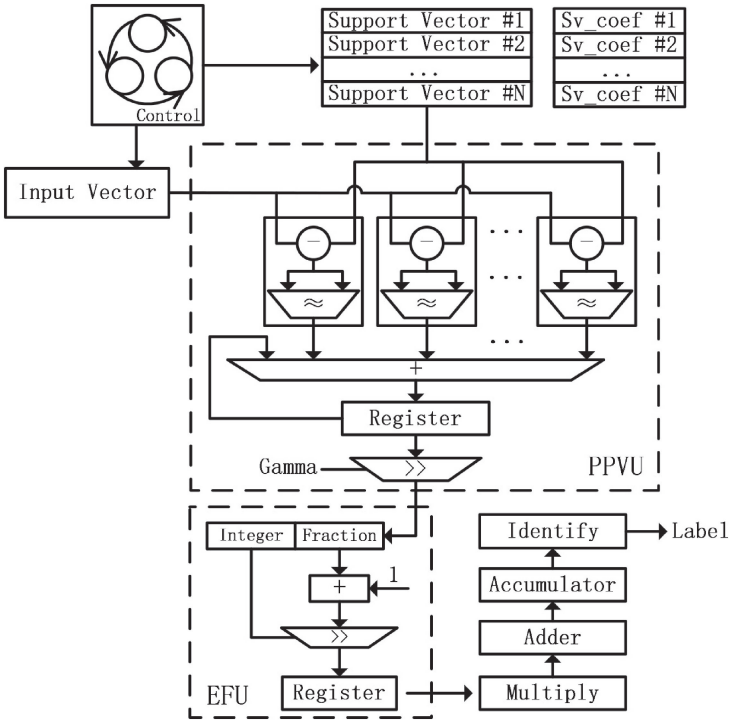

Fig. 1. The feature-based partial parallel architecture of SVM-based classifier.

3.3 Approximation-based hardware architecture

A feature-based partial parallel architecture (PPA) is presented in Fig. 1 to complete the computation between a testing sample and a SV. The main computational units include a partial parallel vector unit (PPVU) and an exponential function unit (EFU), which are described in details below:

1) PPVU: this unit is responsible for calculating $-\|\mathbf{x}-\mathbf{y}\|^2 \times \gamma$, where $\boldsymbol{x} = (x1, x2, \ldots, xn)$ represents a n-dimension testing vector, and $\boldsymbol{y} = (y1, y2, \ldots, yn)$ denotes a support vector. Norm squares are computed by approximate square units instead of multipliers, which helps to save energy and resources. An adder tree is applied to add these square values up in the end. In terms of various dimensional datasets, a partial parallel architecture is used to eliminate the effect of data dimension. Therefore, the final sum of norm squares can be

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

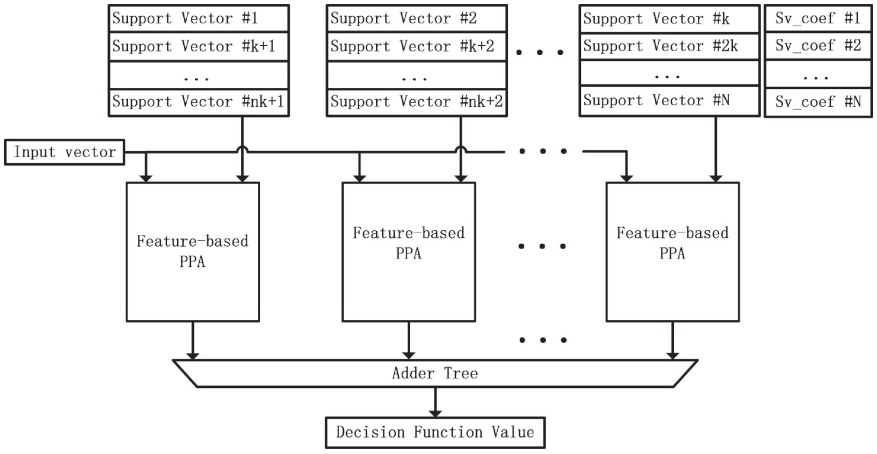

Fig. 2. The sample-based parallel architecture of SVM-based classifier.

got after several iterations and the number of iteration can be calculated using the equation shown in Eq. (9), where N_f and n_p represent the dimension of inputs and the feature-based parallelism of the proposed architecture, respectively. As the parameter γ is a power of two, remaining operations can be completed with a shifter and an adder to obtain the target value.

$$iteration\_number = \left\lceil \frac{N_f}{n_p} \right\rceil \tag{9}$$

2) EFU: this unit is responsible for calculating $exp(x)$. As is aforementioned, the exponential function is transferred and can be represented as a new approximation. In order to implement the exponential approximation, a small area of hardware is occupied as only an adder and a shifter will be used.

The classification speed of SVM-based classifiers is dominated by the number and dimension of SVs and only the influence of dimension has been eased previously. Therefore, a sample-based parallel architecture is designed based on the feature-based architecture to ease the effect of SVs' number and it is shown in Fig. 2. The architecture can handle several SVs simultaneously, which can further improve the classification speed of SVM-based classifiers obviously.

4 Experiment

In this section, the main contents include evaluating the effect of approximative techniques on the classification performance, exploring optimal architecture parameters of the proposed design, and presenting corresponding hardware implementation details at last.

4.1 Experiment for the approximate model

To evaluate the classification performance, twenty datasets from KEEL dataset repository [10] are selected. These datasets are chosen according to the dimension of the input vector and imbalanced ratio. Some datasets are multi-class, therefore, they will be transformed into two classes for the need of the research. Table I shows the detailed characteristics of the datasets, which contains the total number of

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

Table I. Description of selected real-world datasets.

Datasets	#Dim	#Samples	#IR	Datasets	#Dim	#Samples	#IR
Iris-setosa-vs-other	4	150	2	New-thyroid2	5	215	5.14
Led7digit02456789vs1	7	443	10.97	Yeast1vs7	7	459	14.3
Wisconsin	9	683	1.86	Contraceptive2vs13	9	1473	3.42
Heart	13	270	1.25	Wine2vs13	13	178	1.51
Cleveland0vs4	13	177	12.62	Australian	14	690	1.25
Zoo12vs34567	16	101	1.53	Wdbc	30	569	1.68
Ionosphere	33	351	1.79	Satimage2vs45	36	2036	1.90
Texture9vs12	40	1000	1	Kaddcup-land-vs-satan	41	1610	75.67
spectfheart	44	267	3.85	sonar	60	208	1.14
Optdigit23vs45	64	1255	1	Movement-libras1-7vs8-15	90	360	1.14

Table II. Bit widths of variables used in the classification architecture.

Variable	Integer Part	Fraction Part	Signed
Vector Component	4	12	Y
Coefficient	20	12	Y
Bias	12	12	Y
γ	0	16	N
Sample Number	16	0	N
Dimension	8	0	N
Decision Function Value	20	12	Y

datasets (#Sample), the dimension of inputs (#Dim), and the negative-to-positive imbalanced ratio (#IR).

We adopt geometric-mean (G-mean), area under the receiver operating characteristics curve (AUC) [11], and sensitivity as the metrics to assess the results of classification model obtained from the SVM algorithm. Besides, to be consistent with hardware implementation, all parameters of the model are quantized according to Table II. These quantization parameters are chosen in terms of the influence of quantization on the classification accuracy and the quantized model is evaluated on the various datasets. Fig. 3 shows the classification error results of the classifier between with and without approximative techniques according to metrics aforementioned. It is obvious that the classification performance is more or less affected by approximative techniques. However, the maximum error rate occurs according to G-mean and it is only 4.4%. Besides, performance of some datasets is even improved because the fluctuation of classification results caused by approximation is affected by the dimension and number of SVs and hence the orientation of performance varies as the two factors vary.

4.2 Architecture parameter exploration

Different parallel degrees have influence on hardware implementation costs which include latency, energy and resource consumption. Therefore, architectures with various parallel degrees are implemented and experiment is explored to evaluate the influence. Besides, power delay product (PDP) is used to represent the energy consumption. Fig. 4 shows results of the feature-based parallel architecture based on different degrees of feature parallelism. In terms of results, we choose the degree

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

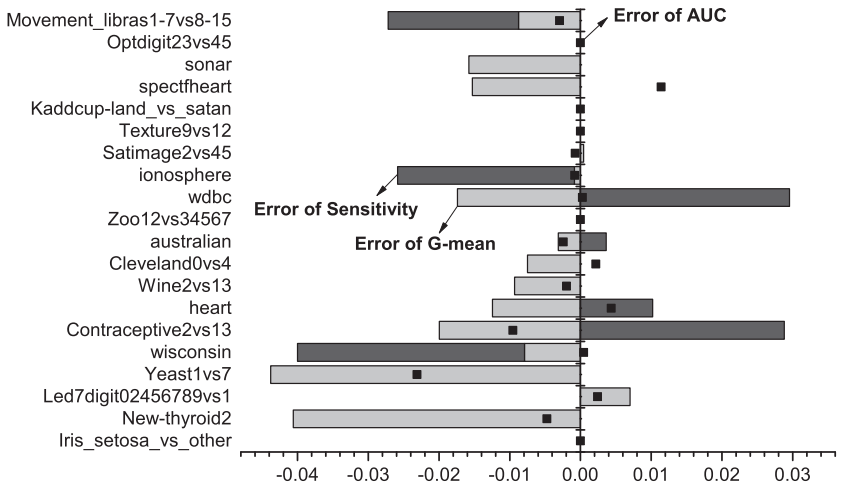

Fig. 3. The relative error of the classifier between with and without approximative techniques in terms of three metrics.

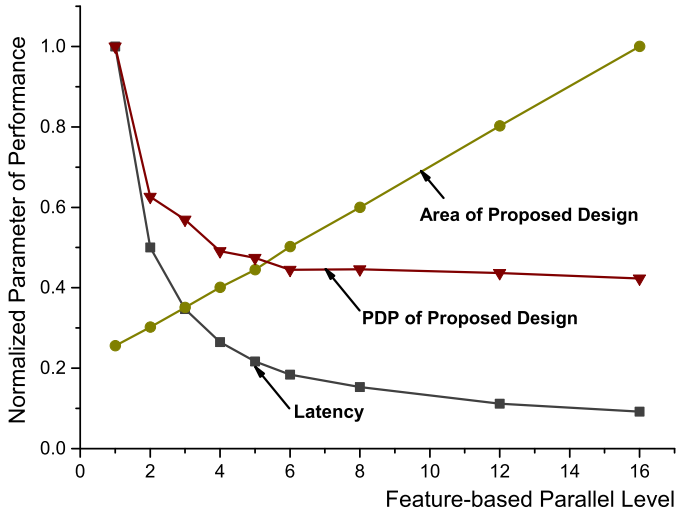

Fig. 4. Results of the parallel architecture based on various feature-based parallelism.

of feature parallelism equal to 6 as the energy consumption reaches the optimal value with minimal resources and nearly stable latency. To achieve the improved parallel architecture, Fig. 5 shows variation trends of its hardware implementation. According to the trends, we choose the improved architecture with sample-based parallel degree equal to 4 as the energy consumption is relatively low and the latency is reduced at the same time. Once parameters of the parallelism degree are determined, our energy-efficient architecture is determined as well.

The proposed architecture is designed in the Verilog HDL and synthesized using a commercial 55 nm CMOS standard cell library. We follow the typical ASIC design flow to perform the synthesis, floor-plan, place, and routing. Parasitic extraction is done after the layout generation. Fig. 6 shows the layout and implementation details of our design. Besides, Table III provides a comparison of implementation details between the proposed architecture and architectures in earlier published papers. The baseline design in the table is implemented with multipliers in PPVU and three-order Taylor expansion combined with region

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

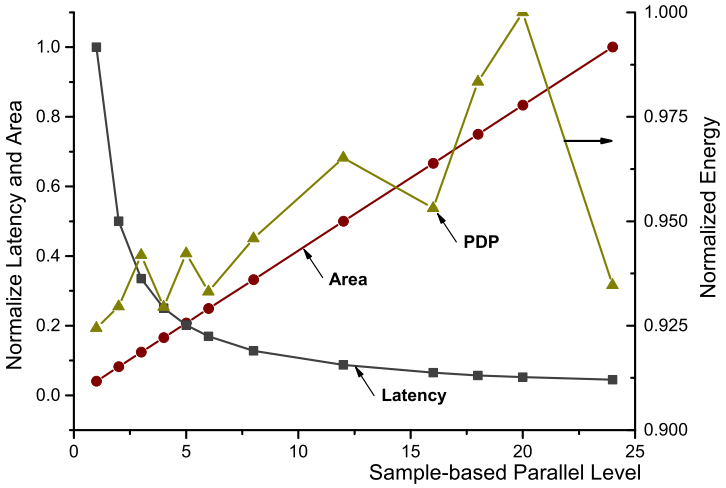

Fig. 5. Results of the parallel architecture based on various sample-based parallelism.

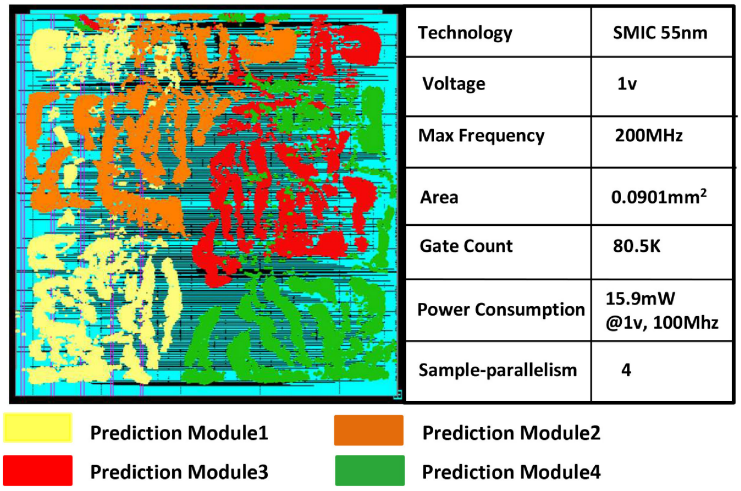

Fig. 6. Layout of proposed design and implementation details.

Table III. Comparison of implementation details between the proposed architecture and architectures in earlier published papers.

Performance	Wang et al. [12]	Attaran et al. [13]	Baseline Design	The proposed design
Technology	90 nm	130 nm	55 nm	55 nm
Max Frequency	100 MHz	125 MHz	150 MHz	200 MHz
Power Consumption	8.12 mW@1v, 100 MHz	20.2 mW@1v, 125 MHz	41.6 mW@1v, 100 MHz	15.9 mW@1v, 100 MHz
Gate Count	1731 k	-	137.5 k	80.5 k
Feature Parallel	4	4	6	6
Improved Parallel	1	1	4	4

constriction in EFU. In terms of comparison results, our proposed architecture can save resources by 41.5% and reduce energy consumption by 61.8% compared with the baseline design. Although the power consumption of designs in [12] is less than ours, our architecture has higher parallel degree and can complete the classification with less latency.

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018

5 Conclusion

This letter presents an energy-efficient architecture of the RBF-based SVM classifier. A partial parallel structure is adopted to eliminate the dimensional constraints of inputs and improve the flexibility of the design. Cost-reduced components based on approximative techniques are designed and the architecture is optimized by replacing the computation-intensive logic with presented components. We adopt a commercial 55-nm technology to synthesize the proposed design and the entire layout are generated for post-layout analysis. It occupies 0.0901 mm^2 area and consumes 15.9 mW at the operating frequency of 1 MHz and from an operating voltage 1 V. Experiments show that the design can save area by 41.5% and reduce energy consumption by 61.8% compared with the design without approximative techniques.

Acknowledgments

This work has been supported by the Fundamental Research Funds for the Central Universities (grant NO. 2015QNA4018), and the State Key Laboratory of ASIC and System (grant No. 2015KF009).

DOI: 10.1587/elex.15.20180099
Received January 31, 2018
Accepted February 26, 2018
Publicized March 14, 2018
Copyedited April 10, 2018