EE 694 Seminar Report

Study of Low Power Analog Circuits for Machine Learning Applications

Submitted By

Sagar Prakash Zoting
203070064

On May 21, 2021

Under the guidance of

Prof. Pramod Murali

Department of Electrical Engineering
Indian Institute of Technology, Bombay
Academic Year 2020-2021

Contents

List of Figures

1. Introduction

The objective behind this project is to understand the low-power analog circuits used for machine learning applications.

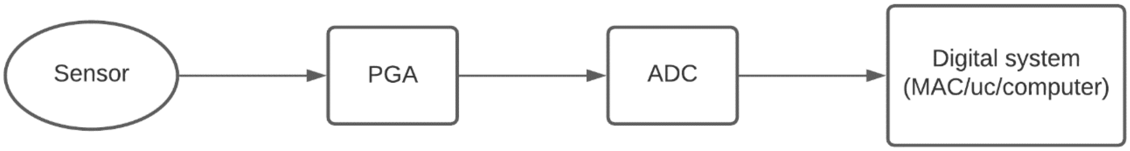

Figure 1: Classical Digital Classifier Model

Machine learning is used as a decision-making tool. Classification algorithms take specific predefined actions when a valid input is detected, defined as intelligent sensing. This algorithm is applied to large data sets obtained from cameras, microphones, and other sensors. The sensors convert physical signals like light and sound to analog electric signals. This analog signal is amplified in such a way that it can meet ADC's input needs. The ADC output can be processed digitally in a digital processor or computer based on the classification algorithm.

1. Sensors: The sensors detect the surrounding elements without distinguishing between input and noise. It is then converted to a digital signal. It serves as an input for the digital system that gives the decision. Temperature variations and noise sources influence the accuracy of the sensors, which affects the classification accuracy. The output voltage level of the sensor can be below the requirements of a digital system, so a programmable gain amplifier is necessary.

2. Amplifier: An ideal amplifier must amplify only the desired input and not the noise. However, in practical applications, unwanted signals with similar input characteristics are also amplified. Like sensors, amplifiers also have their non-idealities, such as offset currents and offset voltages that also affect accuracy. To design an amplifier that meets necessary specifications such as input impedance, output impedance, common-mode rejection ratio, power supply rejection ratio, frequency response, slew rate, open-loop gain, close loop gain, unity-gain bandwidth, power dissipation, and noise (thermal, flicker) need to be taken into considerations.

3. Analog-to-Digital converter: ADC performance depends on the resolution, accuracy, speed, bandwidth, noise level, and power consumption characterized by DNL, INL, offset error, gain error, SNDR, THD, and ENOB. Sources of errors in ADC are temperature, noise in the power supply, quantization, internal mismatch, and many others that affect the accuracy. A low-resolution ADC cannot distinguish between the input and similar pattern. In comparison, high-resolution ADC requires more power and high conversion time.

4. Digital classifier: Multiple higher-level languages (such as C and HDL) can process digital data. In digital classifiers, the computation requires more conversion time and memory as the dimensionality of input space increases. The scaling down of CMOS technology reduces the power of digital systems, but due to physical limitations, the scaling trend is slowing down. Supply voltages do not scale down as fast as feature size; therefore, leakage problems occur. Transistors in the subthreshold operating region consume a large portion of total power without contributing to the computation. In digital classifiers, the neural network evaluates and updates neuron status sequentially, which takes a long time when the interconnections of neurons are extensive. So, there are limits to parallel processing in digital classifiers.

Only special-purpose hardware can help exploit parallelism in the neural networks, which are cost-effective, small in area, and give an immediate decision with minimum precision, therefore fast and robust.

2. Analog Classifiers

Analog classifiers take decisions based on continuous quantities (voltage, current, and charge) directly from sensors. This classifier performs a particular task by extracting features from analog signals. Implementations are specific to applications, so the area is small.

Analog classifiers overcome the limits of digital VLSI technology by using a transistor as a computational element in weak inversion and strong inversion. The noise tolerance of the implementation is improved through machine learning algorithms.

An Analog signal allows a single wire to carry multi-bit information, which improves the power and area efficiency. Mathematical expressions are implemented with the help of the transistor based on conservation of charge for basic manipulation, Kirchhoff's current law for summation, the capacitance node for integration of current.

The analog system benefits from the scaling of technology [4] that improves transconductance in weak inversion, which improves efficiency. Furthermore, due to the reduction in parasitic wiring capacitance, computation throughput is improved. An efficient computation system constructed as slow but massively parallel form using computing elements biased in weak inversion (or subthreshold).

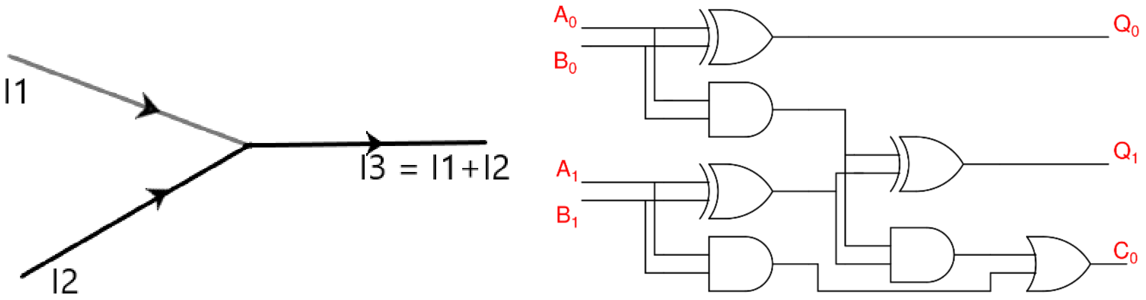

Figure 2: Addition of Two Currents vs. Addition 2-bit [4]

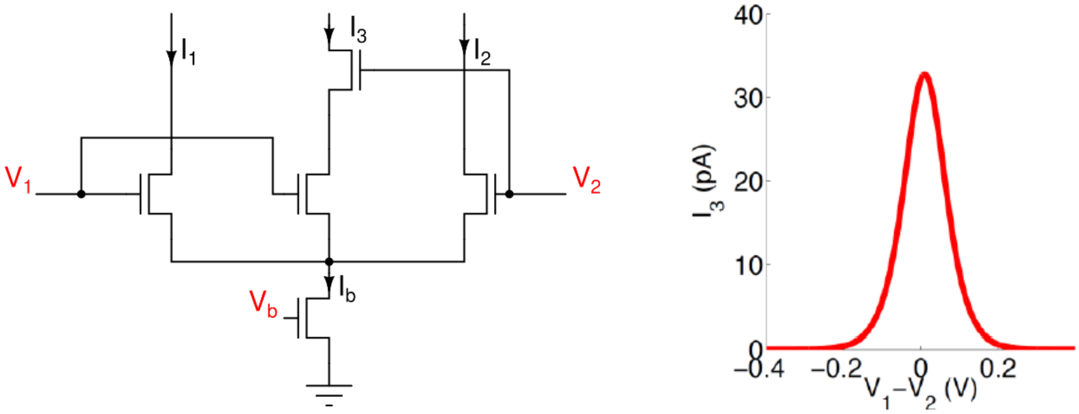

Figure 3: Implementation of Derivative of tanh Function [4]

As shown above, the derivative of tanh function [4], similarly, implementation of many other expressions in the analog domain is possible.

Errors in analog have a much lower magnitude, cause slight performance degradation, and machine learning algorithms can compensate for static errors. Moreover, the leakage is no longer a problem because the subthreshold channel current in the analog circuit can be used to operate instead of being wasted in the digital system.

[1] Several millijoules of energy per classification consumed by a digital system. Implementation of the low energy subsystem in parallel with a digital classifier can selectively activate the entire system by reducing energy consumption when the valid input arrives.

In the analog system, the output of the sensor comprises desirable and undesirable signals. In addition, every step of the implementation contributes some noise. Therefore, for precise classification, signal conditioning is required, making it more complex.

3. Examples of Analog Classifiers

A. A Low-Energy Machine-Learning Classifier Based on Clocked Comparator for Direct Inference on Analog Sensors [1]

In this paper, analog pixel voltage is input, and the classifier gives the 10-way digit classification. Clocked comparator structure replaces amplifier, ADC, MAC (multiplication and accumulation control unit). Machine learning algorithm based on boosted linear classifiers that overcome non-idealities of analog circuits and allow low accuracy of classifier weights. Linear boosting algorithms take different weak weight vectors for different patterns and use them to make a strong classifier model. With the proposed system, classification accuracy is achieved with only a 4-bit resolution. On the other hand, an ideal digital ADC/MAC system requires at least 10 bits. It improvises 12-29 times energy saving over digital implementation.

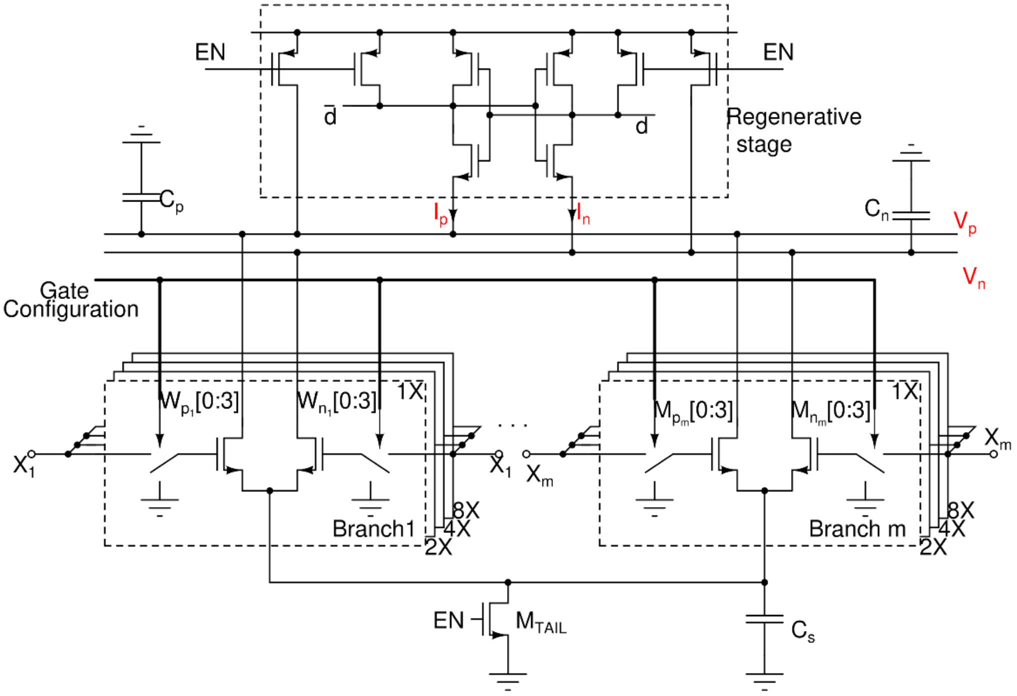

Figure 4: Analog Classifier Implementation for Digit Recognition [1]

Working: Clocked comparator structure implements a linear classifier which takes the inner product of input feature vector $\vec{x}$ to be classified and weight vector $\vec{w}$ (derived from training). The above circuits consist of M branches driven by analog input channels. Image is downscaled to 48 features, therefore M = 48 in this case. Each branch consists of 2 sets of NFETs having a binary-scale width (i.e., $\times 1$, $\times 2$, $\times 4$, $\times 8$) whose gate voltages are digitally configurable to the ground or the analog sensor input. By treating the current from two sets of NFETs as a differential signal, the total branch current corresponds to the signed multiplication between the analog sensor input and weight representing gate configuration. NFETs of all the branches connected to $\left(\mathrm{V_p}/\mathrm{V_n}\right)$ Positive and negative summing nodes for current accumulation. Initially, $\mathrm{V_p}/\mathrm{V_n}$ are pre-charged. When EN is high causes accumulated branch currents to discharge the node capacitances $\mathrm{C_p}/\mathrm{C_n}$ which triggers comparison by regenerative stage. The gate configuration is loaded via shift register after classifier training. When EN is low pre-charged to VDD. When EN is high, the source node $\mathrm{V_s}$ is pulled down by tail NFET by aggregate branch currents I_p/I_n. NFETs are biased in velocity saturation results in linear relations between branch currents and input voltage. Currents from all 48 branches adding together, the structure achieves the 48 MAC operations.

Energy consumed by given design

$$E_{CLASS} = C_P V_{DD}^2 + C_N V_{DD}^2 + C_s(V_{DD} - V_S)V_{DD} \ldots (1)$$

In the proposed paper, a 130-nm CMOS circuit is explained. Image of $28 \times 28 = 784$ pixels is resized and down-sampled to give 48-pixel features as raw input voltage classified in the 10-way digit. It yields an accuracy of 90% with clocked comparators consumes the only CV^2. The system consumes 543 pJ per classification at a rate of 1.3M images per second, representing 33× lower energy than an ADC/digital-MAC system which is very efficient.

B. Analog Voice Activity Detection System for Domestic Environment [2]

The circuit is based on an energy-efficient analog implementation of switched-capacitors with continuous-time non-linear operations. The proposed design consists of a programmable gain amplifier, a squarer, an integrator, averaging circuit, and a threshold update circuit. VAD can be applied to speech recognition, speaker verification, speech enhancement, a voice operation switch, and voice-over-internet protocol.

The VAD function is implemented as an additional low-power block, connected in parallel to the main signal processing chain. When the VAD unit detects the presence of the voice, it turns on the high-performance signal processing chain. This way, the power of the high-performance chain is consumed only when necessary (i.e., when voice is present). Otherwise, only the VAD block is active with low power consumption.

In this document, an energy-based computation model is implemented, which requires low complexity. It compares the energy of the incoming signal against a threshold value. This method is based on the hypothesis that voiced speech segments' energy is higher than unvoiced segments, and voiced speech segments have most of their energy at low frequencies.

The time-variant energy of the incoming audio signal is under the assumption that speech is more nonstationary than ambient noise. The device is fabricated in 180-nm CMOS technology and occupies an area of 0.14 mm^2. Classification accuracy in a domestic environment in the presence of loud ambient noise as large as 99.5% is achieved. The proposed VAD circuit consumes an average current of 633 nA from a 1.2V power supply (760 nW).

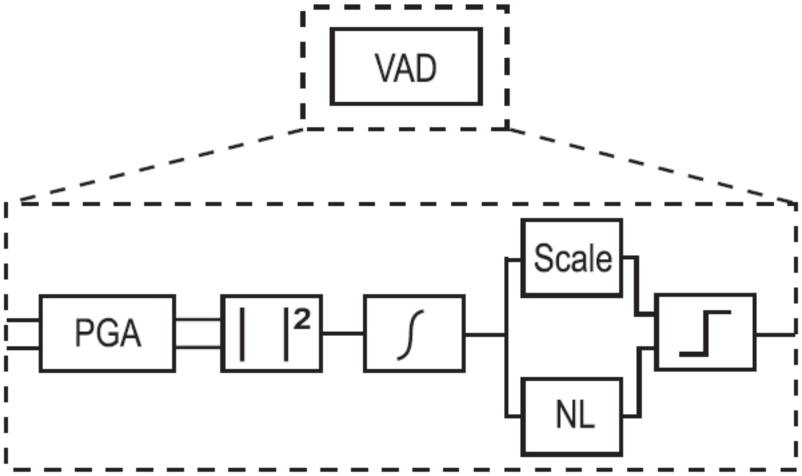

Figure 5: Implementation of VAD Module [2]

Equation energy $\mathrm{E(i)} = \frac{1}{\mathrm{T_{INT}}} \int_{\mathrm{(i-1)T_{INT}}}^{\mathrm{iT_{INT}}} |\mathrm{y(t)}|^2 \, \mathrm{dt} \ldots\ldots (2)$

y(t) is an audio signal divided into a definite time frame for processing.
Noise level (NL) estimation in each frame

$$for\ E(i) > NL(i-1) \rightarrow NL(i) = \beta_1 NL(i-1) + (1-\beta_1)E(i) \ldots\ldots (3)$$

$$for\ E(i) \leq NL(i-1) \rightarrow\ NL(i) = \beta_2 NL(i-1) + (1-\beta_2)E(i) \ldots\ldots (4)$$

$$0.95 \leq\ \beta_1, \beta_2 \leq 0.995$$

Voice presence detected by evaluating $SNR(i) = \frac{(E(i)-NL(i))}{NL(i)} \ldots\ldots (5)$

If $\mathrm{SNR(i)} > \mathrm{TH_{SP}}$ represents voice detected. So, the VAD module generates an activation signal. Equation 5 reformed as below.

$$E_{SC}(i) = \gamma . E(i) > NL(i)\ \ where\ \gamma = \frac{1}{1 + TH_{SP}}$$

The microphone signal passed through a bandpass filter and amplified by the PGA, squared, and integrated for the desired time frame to obtain the signal energy E(i), used to update the noise level NL(i) (NL block) and to produce the scaled energy value $\mathrm{E_{SC}(i)}$. The achieved values of NL(i) and $\mathrm{E_{SC}(i)}$ are compared to generate the VAD signal eventually.

C. A Modular Current-Mode Classifier Circuit for Template Matching Application [3]

The proposed article intends to classify an unknown object into a class with similar things. For pattern-recognition problems, a classifier is established to build a transformation between features and classes. Many well-known classification algorithms can be used, such as the K-nearest neighbor (KNN) algorithm, the c-means algorithm, and the nearest prototype (NP) algorithm. The classifier uses the concept of Euclidean distance to measure similarities between input patterns.

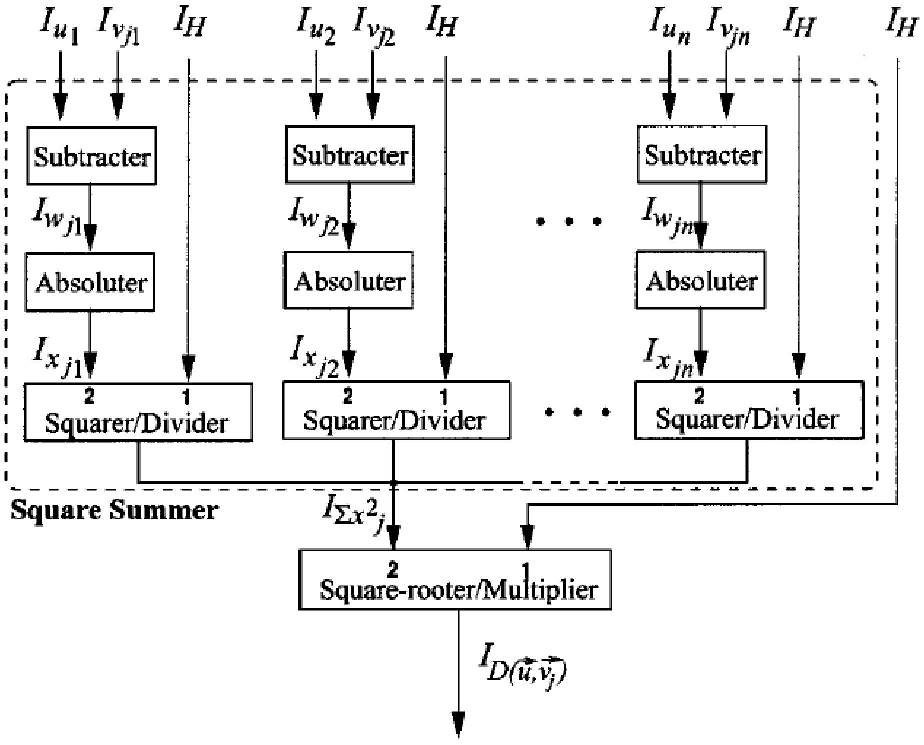

Figure 6: Block Diagram of Template Matching [3]

All circuits designed by biasing MOS transistors in the saturation region for a higher signal-to-noise ratio (SNR). These circuits offer the advantages of low supply voltage, low power consumption, high dynamic range, and high modularity.

Expression for Euclidean distance

$$D(\vec{\mathrm{u}},\vec{\mathrm{v}_\mathrm{j}}) = \sqrt{(\mathrm{u}_1 - \mathrm{v}_{\mathrm{j}1})^2 + (\mathrm{u}_2 - \mathrm{v}_{\mathrm{j}2})^2 + \cdots + (\mathrm{u}_\mathrm{n} - \mathrm{v}_{\mathrm{jn}})^2}$$

Expression for Euclidean distance in terms of MOSFET current in saturation

$$I_{D(\bar{u},\bar{v_j})} = \sqrt{\left(I_{u_1} - I_{v_{j1}}\right)^2 + \left(I_{u_2} - I_{v_{j2}}\right)^2 + \cdots + \left(I_{u_n} - I_{v_{jn}}\right)^2}$$

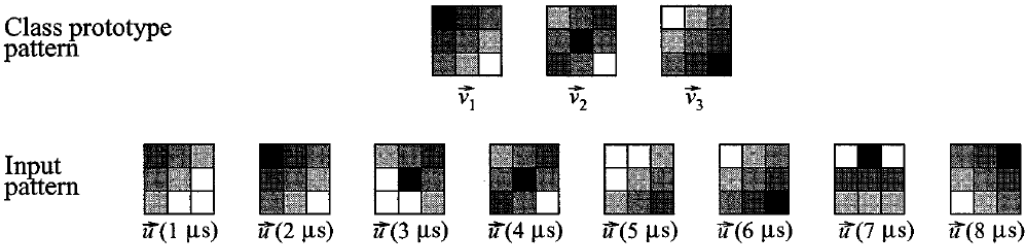

Figure 7: Class Prototype and Test Pattern for Template Matching [3]

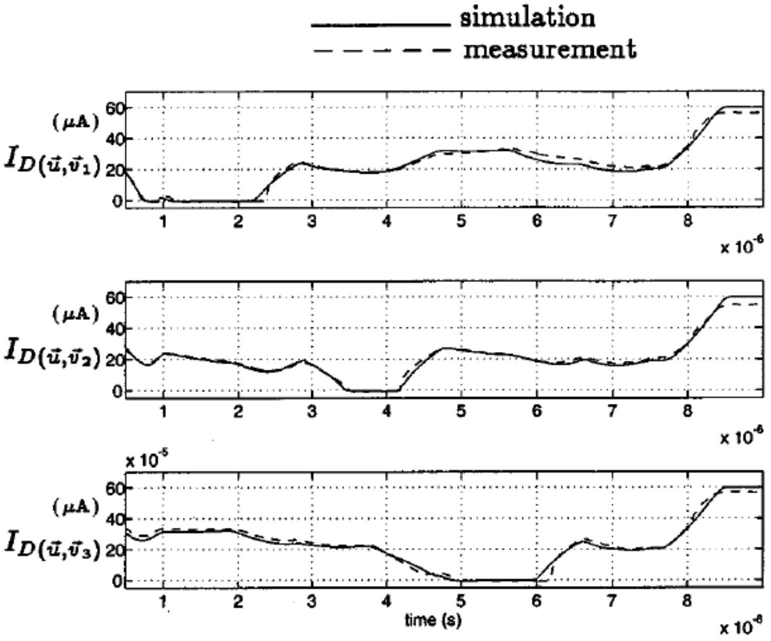

Figure 8: Experimental Results of Pattern Matching [3]

The proposed system is implemented with a 0.6µm CMOS process. The current-mode approach of the circuit allows it to operate with 14.95-mW power consumption. The dynamic range of 100 μA can also be reached with a nonlinearity below 1%. The measured rise time/fall time of this circuit is about 530/30 ns. Some modifications are required to minimize nonideal effects in the MOS transistors. Relying on the parallel processing characteristic and the mentioned merits, this circuit is suited to real-world low-power and low supply voltage applications.

4. Limitation of Analog Classifiers

A classifier is built on CMOS technology. Analog classifiers are composed of a sensor, PGA, classifier circuit. Secondary effects [3] are observed on CMOS transistors. These effects degrade the classifier's performance if not addressed.

The increase of the source-to-substrate voltage ($\mathrm{V_{SB}}$) in a MOS transistor thicken the depletion region, so the threshold voltage(V_t) is increase, known as the body effect. The speed of the circuit slowed down due to the large well-to-substrate capacitances. The channel-length modulation effect(λ) causes the drain current to depend on the drain-to-source voltage ($\mathrm{V_{DS}}$). The finite output resistance ($\mathrm{r_{ds}}$) also causes the loading effect when cascading other stages.

In MOS circuits, variations in process, voltage, and temperature (PVT) can affect the bias, currents, and load distribution resulting in a wrong decision. One of the goals of analog classifiers is to reduce the area of implementation. But as technology scaled-down, the noise sources start dominating, which can be eliminated by adding more circuiting resulting increase in area.

The slew rate affects the classifier's performance. In practice, the devices selected for system implementation have frequency limitations. If the input frequency exceeds the slew rate of the devices shows non-linear behavior. In analog systems, precision is directly proportionate to cost. Therefore, analog computation is cheaper at low values of accuracy but more expensive at high accuracy.

In any circuit implementation, expected performance deals tradeoffs between several parameters, i.e., if the device area required small, there is a tradeoff between power consumption and bandwidth, bandwidth and intrinsic noise (thermal and flicker), noise and Signal-to-Noise Ratio (SNR), SNR and the dynamic signal range, and so on. Parameters such as parasitic components, threshold voltage, scaling, offset voltages, and harmonic distortion affect classifier performance. These parameters highly depended on the device size ($\propto 1/WL$). Local variations like sheet resistance, channel dopant concentration, mobility, and gate oxide thickness decrease with the device's size. Flicker noise ($1/f$) is not a process parameter, as in the case of the threshold voltage, but it depends on $1/WL$, i.e., flicker noise also decreases as the device size increase. Conversely, Adjusting the layout of the transistor with the bias helps to minimize the matching problems.

Due to downscaling and parasitic capacitance, there are non-idealities like charge injection errors and clocked feedthrough errors that affect classification accuracy. The circuit limits the dynamic range of analog classifiers' voltage/current headroom and noise level. Both are difficult to address and problematic for multiplication operations because it requires more dynamic range than other operations.

[1] In the post-layout simulation, charge injection error observed in clocked comparator-based classifiers causes FETs to transition from sub-threshold to above threshold, which tends to increase the voltage V_P/V_N. Therefore, the circuit is sensitive to these transients, which can affect the accuracy of the decision. As iteration increases, FET transitions increase, causing instability and decision errors. Matching the gate configuration of V_P/V_N can reduce the charge-injection error.

5. References

[1] Zhuo Wang, Naveen Varma, IEEE TCAS-1 VOL. 64 no. 11, November 2017 *"A Low-Energy Machine-Learning Classifier Based on Clocked Comparators for Direct Inference on Analog Sensors."*

[2] Marco Croce, Brian Friend, Francesco Nesta, IEEE Journal of Solid-State Circuits, VOL. 56 NO.3 March 2021, *"A 760-nm, 180-nm CMOS Fully Analog Voice Activity Detection System for Domestic Environment."*

[3] Bin-Da Liu, Chuen-Yau Chen and Ju-Ying Tsao. IEEE TCAS-II, Analog, and Digital Signal Processing, VOL. 47 NO. 2, February 2000. *"A Modular Current-Mode Classifier Circuit for Template Matching Application."*

[4] Junjie Lu, University of Tennessee, Knoxville, Doctoral Dissertation, May-2014 *"An Analog VLSI Deep Machine Learning Implementation"*