Research Article | Vol. 32, No. 20 / 23 Sep 2024 / Optics Express 34264
Optics EXPRESS

Silicon integrated photonic-electronic neuron for noise-resilient deep learning

IOANNIS ROUMPOS,[1,2,*] LORENZO DE MARINIS,[3] STEFANOS KOVAIOS,[2,4] PETER SEIGO KINCAID,[3] EMILIO PAOLINI,[3] APOSTOLOS TSAKYRIDIS,[2,4] MILTIADIS MORALIS-PEGIOS,[2,4] MATHIAS BERCIANO,[5] FILIPPO FERRARO,[5] DIETER BODE,[5] SRINIVASAN ASHWYN SRINIVASAN,[5,6] MARIANNA PANTOUVAKI,[5,7] NICOLA ANDRIOLLI,[8] GIAMPIERO CONTESTABILE,[3] NIKOS PLEROS,[2,4] AND KONSTANTINOS VYRSOKINOS[1,2]

[1]*Department of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[2]*Center for Interdisciplinary Research & Innovation, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[3]*Scuola Superiore Sant'Anna, 56124 Pisa, Italy*
[4]*Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[5]*Imec, Kapeldreef 75, 3001 Leuven, Belgium*
[6]*Lightmatter Inc., 100 Summer Street, Boston, MA 02110, USA*
[7]*Microsoft Research Center, Cambridge, UK*
[8]*University of Pisa, 56122 Pisa, Italy*
[*]*ioroumpo@auth.gr*

Abstract: This paper presents an experimental demonstration of the photonic segment of a photonic-electronic multiply accumulate neuron (PEMAN) architecture, employing a silicon photonic chip with high-speed electro-absorption modulators for matrix-vector multiplications. The photonic integrated circuit has been evaluated through a noise-sensitive three-layer neural network (NN) with 1350 trainable parameters targeting heartbeat sound classification for health monitoring purposes. Its experimental validation revealed F1-scores of 85.9% and 81% at compute rates of 10 and 20 Gbaud, respectively, exploiting quantization- and noise-aware deep learning techniques and introducing a novel activation function slope stretching strategy for mitigating noise impairments. The enhanced noise-resilient properties of this novel training model are confirmed via simulations for varying noise levels, being in excellent agreement with the respective experimental data obtained at 10, 20, and 30 Gbaud symbol rates.

1. Introduction

Analog computing is emerging as a promising alternative for high-speed and power-efficient Deep Neural Networks (DNN), particularly in the computationally demanding segment of matrix multiplications [1]. Optical Neural Networks (ONN) constitute a valuable analog computing approach for efficient multiplication operations, leveraging the inherent parallelization and speed of light [2–4]. Moreover, the unique characteristics of ONNs, coupled with the advancing maturity of photonic integration, enable denser implementations of ONNs on silicon chips, known as integrated Photonic Neural Networks (PNNs). The integration of PNNs with already established electronic platforms can offer a hybrid photonic-electronic platform capable of harnessing the advantages of both analog electronic and optical analog computing paradigms.

So far, PNN demonstrations have showcased excellent performance in terms of computing speed and power consumption [5–26], leveraging different multiplexing techniques in order to extend the computational margins of the limited photonic hardware [27]. However, the spatial

#532306 https://doi.org/10.1364/OE.532306
Journal © 2024 Received 11 Jun 2024; revised 25 Jul 2024; accepted 28 Jul 2024; published 9 Sep 2024

Optics EXPRESS

division multiplexed (SDM) schemes [16,17] impose scalability challenges in large network implementations, since spatial expansion would lead to large circuit layouts with high insertion losses and challenging electrical routing. Similarly, the implementation of wavelength division multiplexed (WDM) layouts [13–15], [25] requires a number of optical channels that increases with network dimensions, facing severe limitations to comply with the large size of NN parameters. The only architectural framework that has been shown to be capable of supporting hundreds of trainable parameters relies on the utilization of time division multiplexing (TDM) concepts [28,29]. TDM leverages the time domain to scale up calculations, implementing the summation function through integration over time and as such significantly relaxing the demands for speed and power consumption at the Analog to Digital (ADC) stage. TDM schemes can be also synergized with additional multiplexing degrees like space and/or wavelength to boost circuit scalability and computational power, reducing at the same time latency [26,30,31].

Towards this direction, it has been recently introduced and numerically demonstrated a photonic-electronic multiply-accumulate neuron (PEMAN) architecture that performs TDM multiplications in the photonic domain and accumulation in the electronic domain via an analog integrator circuitry and applies a nonlinearity within its low-speed ADC [32]. The PEMAN is capable of executing all necessary operations of an artificial neuron, trading-off multiplication speed and energy consumption with accuracy. However, the photonic-electronic circuit is expected to be subject to various types of noise in case it gets physically deployed, including shot noise, thermal noise and data conversion quantization noise that limit the bit precision of the system. The physical constraints of analog photonic NN hardware have to be taken into account during the training process in order to maintain an effective network performance [23], an area that has been recently introduced as optics-informed Deep Learning [3]. In this realm, various methods have been so far proposed to mitigate the noise impact [18,24,33,34]. More specifically, the demonstration in [18] investigates the effects of non-deterministic noise sources of photonic hardware that were approximated via Additive Gaussian Noise Sources (AWGN), including laser Relative Intensity Noise (RIN), Johnson shot-noise and uniform quantization noise. The mean value and standard deviation of the AWGN were set to match the experimentally obtained noise characteristics of the photonic circuit. These values were then incorporated into the training process, resulting in higher classification accuracies of the MNIST dataset, compared to conventional training methods. On the other hand, the authors in [24,34] proposed a novel training method that incorporates only the limited frequency response of the deployed photonic components, without applying any quantization on the NN values. Finally, the authors in [33], proposed a mixed-precision quantization-aware training scheme that can adjust the bit resolution among the NN layers in order to reduce the inference execution time. Although these demonstrations show improvement on the performance of the photonic neural network, they primarily rely on quantization strategies applied only during the training phase, neglecting the application of quantization during the inference process which may further enhance the performance.

In this paper we report for the first time the photonic part of the PEMAN architecture as a silicon photonic chip and demonstrate its experimental performance in heartbeat sound classification tasks, exploiting an innovative optics-informed training model for enhancing noise resiliency. The silicon photonic chip employs high-speed SiGe electro-absorption modulators (EAMs) for both input and weight signal encoding, executing element-wise multiplications exclusively in the optical domain and completing the matrix multiplication via summation in the analog electronic domain. The silicon-integrated PEMAN layout has been experimentally evaluated by executing the computations required by a noise-sensitive three-layer NN with 1350 parameters trained for heartbeat sound classification as normal or abnormal. In order to enhance the noise tolerance of the NN, we propose a novel quantization strategy applied both during training and inference processes, including i) quantization at the output of the activation function and ii) stretching of

Optics EXPRESS

the slope of the activation function. After applying these strategies, the noise impact of the analog photonic hardware on the classification accuracy is minimized. Its experimental validation was carried out at data rates up to 30 Gbaud, revealing F1-scores up to 85.9% at 10 Gbaud and 81% at 20 Gbaud. Its robust performance was verified via simulation analysis for different noise levels, showing excellent agreement with the respective experimental results and highlighting that the combined use of the proposed quantization strategy and the activation function slope stretching allows network performance to degrade much slower with increasing noise levels compared to the respective digital NN.

2. Photonic electronic neurons with electro-absorption modulators

PEMAN is a neuromorphic analog processor that synergizes the advantages of both photonic and electronic components to efficiently perform all the necessary operations of an artificial neuron [32]. Figure 1(a) presents a schematic diagram of the PEMAN architecture, outlining both the required opto-electronic components and the employed TDM approach for implementing matrix-vector multiplications necessary to carry out the neuromorphic processing of an N-neuron layer, with each neuron fed by K inputs. In our proposed TDM approach, both the input vector and the weight matrix are transformed to serialized time vectors prior to being injected into the PEMAN accelerator. To this end, the weight matrix with a dimensionality of [N × K] and the input vector of [K × 1] are serialized to a w(t) and X(t) time vectors, with the latter generated by repeating the input vector N times. First a high-speed optical modulator imprints the input time vector X(t) on an incoming optical carrier. The resulting signal is split in two branches, each comprising an optical modulator responsible for imprinting the weight vector decomposed as a "positive" and "negative" vector (w_+ and w_-) so to drive the EAMs in a push-pull configuration. The output signals constituting the positive and negative products $X(t) \times w_+(t)$ and $X(t) \times w_-(t)$, are subsequently injected into a balanced photodiode that provides at its output the algebraic summation of the two products ($X(t) \times w_+(t)$ - $X(t) \times w_-(t)$). Finally, an electronic integrated circuit (EIC), incorporating an analog accumulator module, accumulates the incoming product values in each time slot, while a non-linear ADC applies the required nonlinearity and converts the signal to the electrical domain. More details considering the operating principles and the components of the EIC are described in [32].

In this work, a SiPho chip, fabricated using IMEC's SiP 300 mm wafer technology, was utilized for the experimental evaluation of the PEMAN architecture, with Fig. 1(b) illustrating a photo of the circuit covering a total area of 2.8 mm^2. The figure also highlights the constituent input (EAM X) and weight encoding (EAM w + and EAM w-) modulators, comprising 50 µm long Franz-Keldysh EAMs of 56 GHz bandwidth. Further information regarding the fabrication details can be found in [35]. The EAMs revealed an insertion loss (IL) of 5 dB at 1560 nm, while the extinction ratio (ER) ranges up to 7 dB when a static reversed biased voltage of 3 V is applied.

Figure 1(c) depicts the experimental setup used for evaluating the PEMAN architecture. A continuous-wave (CW) light beam at 1560 nm, corresponding to the optimum operating wavelength in terms of IL and achieved ER of the constituent EAMs, was coupled into the chip via a transverse electric (TE)-grating coupler. High-speed RF signals were generated by a 25 GHz bandwidth Keysight 8195a Arbitrary Waveform Generator (AWG) and subsequently amplified to $3V_{pp}$ using 67 GHz bandwidth RF amplifiers thereby providing amplitude modulated signals up to 30 Gbaud. All three EAMs were biased at -1.5 V to ensure reverse bias operation, considering that a $3V_{pp}$ RF signal was applied on the devices. The optical signals emerging at the chip's output were amplified by Erbium-Doped Fiber Amplifiers (EDFAs) to counteract the optical losses introduced by the photonic integrated circuit (PIC), prior to being filtered by optical bandpass filters (OBPF) to eliminate the EDFA's amplified spontaneous emission (ASE) noise. However, in the complete PEMAN, with co-designed and co-packaged photonic and electronic integrated circuits, the EDFA and the OBPF are not needed, that would otherwise hamper the

Optics EXPRESS

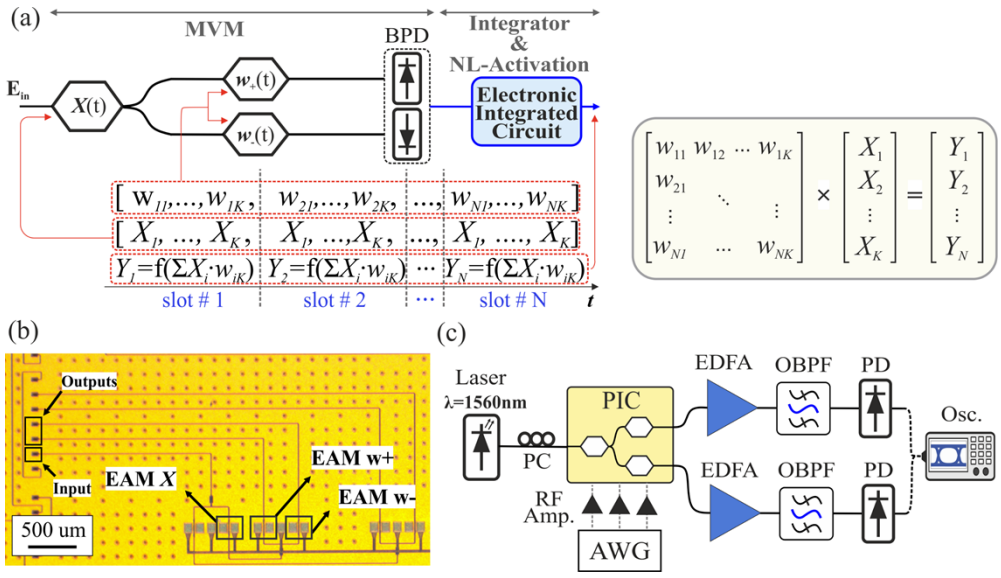

Fig. 1. (a) PEMAN architecture along with a schematic explanation of the matrix vector multiplication that is performed on different time slots, for both positive and negative weights, on the device. (b) Microscope photo of the silicon photonic chip indicating the EAMs of the circuit. (c) Experimental setup for the PEMAN evaluation.

device power efficiency and compactness. The resulting signals were injected into two different 70 GHz bandwidth photodiodes and acquired by two separate digital oscilloscope channels due to the lack of an on-chip balanced photodiode. Finally, the time synchronization and the algebraic summation of the two waveforms, along with the accumulation and the non-linear activation function, were carried out digitally using software routines.

The integrated PEMAN circuit was evaluated over the NN model illustrated in Fig. 2(a). The NN was trained to classify heartbeat sound samples in normal and abnormal classes, with the abnormal class correlated with the presence of a heart disease, such as Murmur or Extra-systole [37,38]. Prior to being injected into the NN, the heartbeat sound signals were first processed in a feature extraction module, responsible for extracting 52 features from the input time-series. The NN followed a fully-connected topology and comprised an input, a hidden, and an output layer with 52, 25 and 2 neurons, respectively. The ReLU activation function with a stretched slope factor of 4 [39] was utilized at the hidden layer (Layer 1), while the output layer employed a softmax function, followed by a comparison function that classified the heartbeat sample as abnormal when the value of the first neuron was higher than or equal to that of the second neuron, and as normal otherwise. It should be noted that both the input values and the weights of the neural network were quantized into 4 levels, corresponding to an equivalent bit resolution of 2 bits within the range of [-1,1] and [0,1] for weights and inputs, respectively.

This approach was motivated by the inherent noisy profile of analog photonic accelerators, that reduces the effective bit resolution of the system, constraining it within a range of 1 to 6 bits [22,32]. Quantized photonic neural networks have been proposed as a system-agnostic approach in mitigating the signal degradation originating from analog noise, either involving quantization during only the training phase of the NN [33] or quantization both during the training and inference phase [36]. The latter method is tailored to the quantization typically enforced through the constituent digital-to-analog (DAC) and analog-to-digital (ADC) converters in opto-electronic layouts and, as such, was the method employed in our work. In order to

Optics EXPRESS

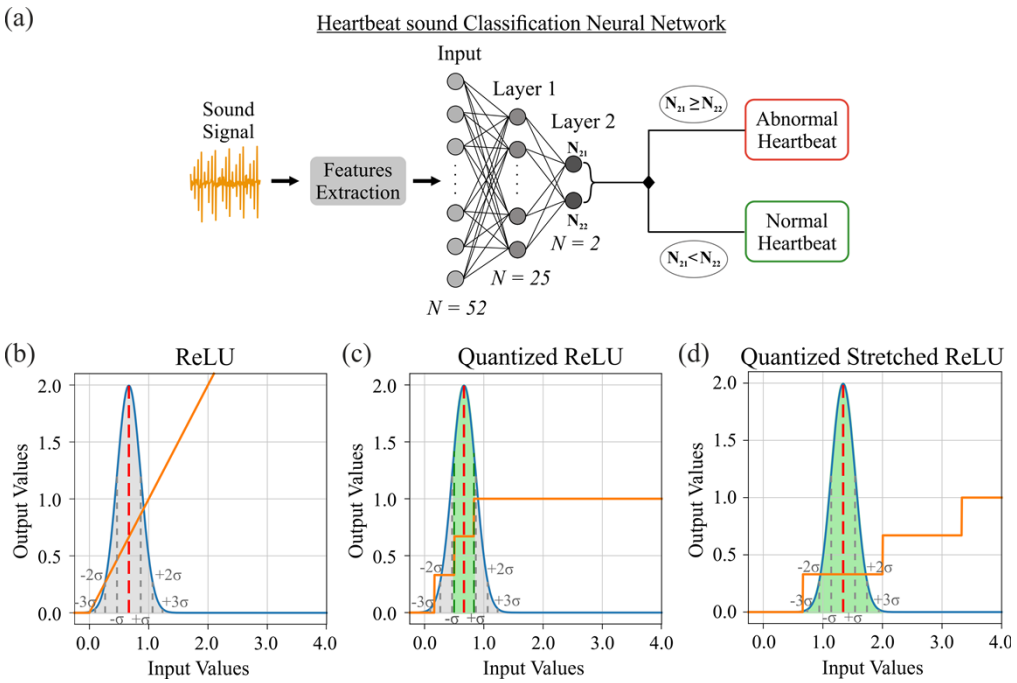

Fig. 2. (a) Topology of the heartbeat sound classification neural network. Different cases describing the effect of noise mitigation on an input value of 0.66 with σ of 0.2, to emulate the AWGN of the system. (b) Noise distribution (blue curve) and the ReLU function (orange curve). (c) Noise distribution and quantized ReLU. (d) Noise distribution and quantized stretched ReLU.

illustrate how quantized photonic neural networks can improve NN performance, Figs. 2(b)-(d) depict the input/output correlation of an activation function module for three different activation functions i.e. i) typical ReLU, ii) Quantized ReLU and iii) Quantized Stretched ReLU. The input is approximated by a gaussian distribution with a mean value of $\mu = 0.66$ for i) and ii), and $\mu = 1.32$ for iii), corresponding to the targeted input values, and a standard deviation of $\sigma = 0.2$, which emulates the presence of AWGN in the opto-electronic neuron. The output of a single photonic neuron with the typical ReLU applied is defined as:

$$y = f\left(\int X(t) \cdot w(t)\, dt + \mathcal{N}(\mu, \sigma^2)\right), \tag{1}$$

where $X(t)$ and $w(t)$ represent the input and weight time sequences, respectively, $\mathcal{N}(\mu, \sigma^2)$ is the additive noise and f is the ReLU function that defined as:

$$f(x) = \begin{cases} 0, & x \le 0 \\ x, & x > 0 \end{cases}. \tag{2}$$

Figure 2(b) illustrates that when the ReLU activation function is applied, the additive noise is identical pre- and post- activation function resulting in incorrect output values. A strategy for addressing the noise induced discrepancy is the quantization of the values after the ReLU activation function, denoted as the Quantized ReLU case. The neuron output is then defined as:

$$y = Q\left[f\left(\int X(t) \cdot w(t)\, dt + \mathcal{N}(\mu, \sigma^2)\right)\right], \tag{3}$$

Optics EXPRESS

and specifically for the 2-bit resolution case as:

$$Q(x) = \begin{cases} 0, & x \le 0.16 \\ 0.33, & 0.16 < x \le 0.50 \\ 0.66, & 0.50 < x \le 0.83 \\ 1, & x > 0.83 \end{cases}. \tag{4}$$

Figure 2(c) illustrates the noise suppression qualities of the quantized ReLU configuration, with the output reaching the correct value for an error interval of the input value of approximately ±1σ. To further enhance the noise resilience of the network, a method based on stretching the quantization intervals can be applied [39], resulting in the Quantized Stretched ReLU configuration. The neuron output in this case is defined as:

$$y = Q\left[\frac{f\left(\int X(t)\cdot w(t)\,dt + \mathcal{N}(\mu,\sigma^2)\right)}{a}\right], \tag{5}$$

where the parameter a refers to the stretching factor of the ReLU slope. Figure 2(d) illustrates the input output correlation of the Quantized Stretched ReLU configuration for an input value of $\mu = 1.32$ with $\sigma = 0.2$, when using a stretched interval with a factor $a = 4$. As can be observed, the output reaches the correct value for a wider input error interval, illustrated by the area highlighted in green, that corresponds to approximately to ±3σ of the error distribution, increasing the noise resilience of the NN. It should be mentioned that even though this approach results in a shifting of the output values at lower levels, the NN accuracy is not affected as the monotonicity of the output sequence remains intact. A simulation analysis considering different noise levels in the neural network inference, for the three aforementioned cases, is presented in the next section.

3. Results and discussion

The performance of the integrated PEMAN neuron during NN inference was assessed for 20 different heartbeat samples, each comprising 52 extracted features, and for three different compute rates, 10, 20 and 30 Gbaud. In Fig. 3(a) and (b) the experimental and software derived output of Layer 1 for the first 5 input samples are overlaid, at computing speeds of 10 and 20 Gbaud, with mean squared error values (MSE) of 0.014 and 0.022, respectively. The divergence between the experimentally acquired and ideal software traces for all 20 input samples at Layer 1 was quantified also by calculating an error vector corresponding to their difference and fitting a zero mean gaussian distribution to the resulting values. The 30 Gbaud signals are not shown due to the higher discrepancy between the software and experimental waveforms. Figure 3 (c) illustrates the derived standard deviation values for operating compute rates of 10, 20 and 30 Gbaud, resulting in values of $\sigma_{10} = 0.09$, $\sigma_{20} = 0.10$ and $\sigma_{30} = 0.17$, respectively. The significant increase in the noise profile at 30 Gbaud is attributed to the limited bandwidth characteristics of the hardware setup mainly attributed in the electrical part. Figure 3 (d) and (e) illustrate the software and experimental output values of the 2 neurons of Layer 2 for all 20 different input samples again at compute rates of 10 and 20 Gbaud, respectively. The experimentally derived values are indicated by the cross ('x') scatter points and the dotted lines, while the software derived values are denoted with the circle ('o') and continuous lines. A close inspection of the output values for the two compute rates reveals a slight degradation in the matching of the software and experimentally derived values when moving from 10 to 20 Gbaud. This discrepancy is reflected in the achieved F1 score values of the neural network, which are 85.9% and 81.0%, for 10 and 20 Gbaud respectively, while the baseline software F1 score is 94.8%. The confusion

Optics EXPRESS

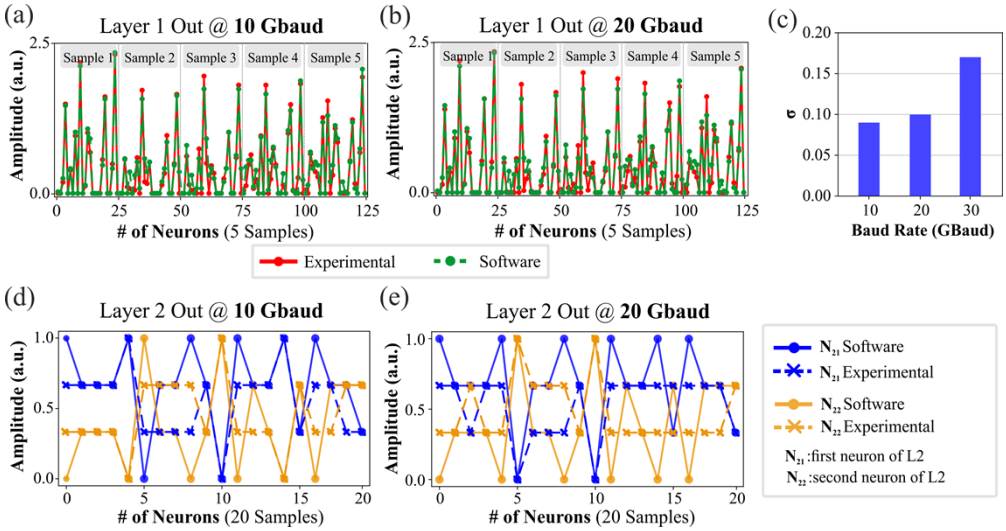

Fig. 3. Experimental and software traces for the (a) 10 Gbaud and (b) 20 Gbaud NN encompassing the first 5 input samples after the application of the stretched ReLU function in layer 1. (c) Bar plot of the noise standard deviation values, calculated on the total 20 input samples traces of layer 1 after the activation function, for the different compute rates. (d) Experimental and software output traces of the two neurons at layer 2 after softmax, encompassing all input samples, for 10 Gbaud and (e) 20 Gbaud.

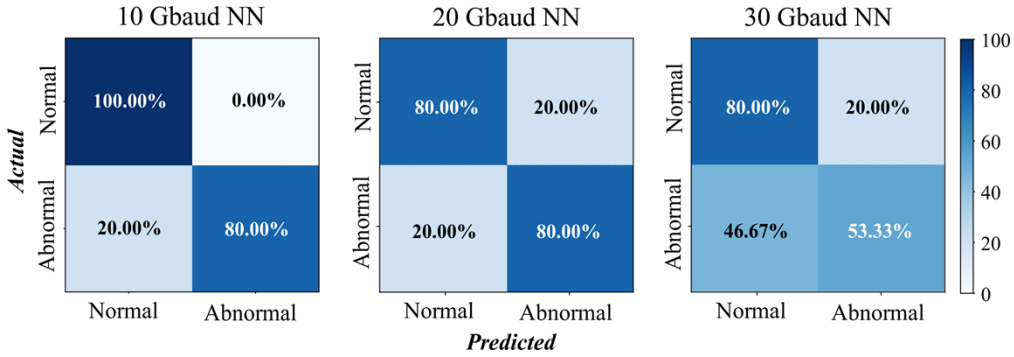

Fig. 4. Confusion matrices for compute rates of 10, 20, and 30 Gbaud.

matrices for the different compute rates, stemming from the experimental validation of the NN, are depicted in Fig. 4.

Finally, we have calculated the energy and footprint efficiency of the device using state of the art DACs and RF amplifiers, as well as the values reported in [32] for the electrical backend of PEMAN. The energy efficiency can be broken down into the energy efficiency of the optical and the electrical components, as follows:

$$e = e_{el} + e_{opt} \tag{6}$$

The optical term is defined as:

$$e_{opt} = \frac{P_{Laser} + P_{EAM-X} + 2P_{EAM-w}}{CR} \tag{7}$$

Optics EXPRESS

where P_{Laser}, P_{EAM-X} , P_{EAM-w} are the power consumption of the laser, the input EAM and the weight EAM, respectively, following the principles defined in [40], and CR denotes the compute rate. The electrical energy efficiency can be calculated as:

$$e_{el} = \frac{3P_{DAC} + 3P_{RFamp} + P_{TIA+INT+ADC}}{CR} \tag{8}$$

where P_{DAC} is referred to the power consumption of the DAC, P_{RFamp} to the power consumption of the amplifiers required to drive the EAMs and $P_{TIA+INT+ADC}$ referred to the power required for the electrical backend of the PEMAN architecture. Considering, the following values, $P_{Laser} = 10\ mW$ assuming a wall-plug efficiency of 20%, $P_{EAM} = 0.6\ mW$, $P_{DAC} = 144\ mW$ [41], $P_{RFamp} = 61\ mW$ [42] and $P_{TIA+INT+ADC} = 13\ mW$ [32], the resulting energy per operation for the different compute rates are $e_{10} = 64.0\ pJ/MAC$, $e_{20} = 32.0\ pJ/\ MAC$ and $e_{30} = 21.3\ pJ/MAC$ for 10, 20 and 30 Gbaud, respectively. Regarding the footprint efficiency of this chip, it is considerably higher because the circuit design has not been optimized. Based on the tested circuit the three EAMs with RF contacts occupy an area of 0.122 mm^2 while the grating couplers cover an area of 0.015 mm^2 considering a pitch of 127 µm. The resulting footprint efficiencies, calculated as in [43], are 66.7, 133.3, and 200 GMACs/mm^2 for the computation speeds of 10, 20 and 30 Gbaud, respectively. Considering a practical routing, an optimized cell for this design would need a height of 300 µm and a width of 500 µm, resulting in an area of 0.15 mm^2 that would result in footprint efficiencies of 37.0, 74.0 and 111.1 GMACs/mm^2.

Following the experimental evaluation of the integrated PEMAN neuron, a simulation analysis was carried out in order to examine the performance of the opto-electronic neural network for different noise profiles and NN configurations. In this analysis, all the noise sources of the electro-optic hardware, such as thermal noise and shot noise, are embedded in a single additive stochastic noise component. As such, the system's noise is modelled as AWGN $\mathcal{N}(0, \sigma^2)$, with zero mean and a standard deviation of σ, expressed as a fraction of the signal power, and added to the resulted product waveform ($X \times w$) during NN inference. The inference performance of the NN for different noise profiles was assessed by gradually increasing the system's noise and measuring the achieved F1 score. Three different NN configurations were examined, i.e., (a) "Typical" floating point (FP) NN configuration, where the input and weight values are represented by floating point values while employing the ReLU activation function, (b) Quantized-NN configuration where the input and weight values are quantized to 2 bits and the ReLU activation function is employed, and (c) Stretched ReLU Quantized NN configuration where the input and weight values are quantized to 2 bits, and a stretched slope ReLU, with a stretching factor a = 4 is employed. Figure 5 illustrates the achieved F1 scores of all three NN configurations, when the noise standard deviation σ is swept across the range [0, 0.5]. The simulation results indicate that the FP NN is the most vulnerable to additive noise, with its performance significantly deteriorating even from low-noise levels, i.e., with σ = 0.1. The Q-NN exhibits better noise tolerance, achieving F1 scores >65% for noise values up to σ = 0.1. Finally, the SR-QNN configuration further enhances the noise tolerance of the network, retaining f1 scores higher than 50% even for significant noise values.

The correlation of the experimentally achieved F1 scores with the experimentally measured noise profiles at the three targeted compute rates of 10, 20 and 30 Gbaud, are also illustrated in Fig. 5. As can be observed, they are in good agreement with the simulation results as they lie within the error window of the blue curve, which corresponds to the experimental configuration. The noise standard deviation associated with 30 Gbaud is much higher than the other two, but it could be reduced if the setup were optimized for higher rates.

Optics EXPRESS

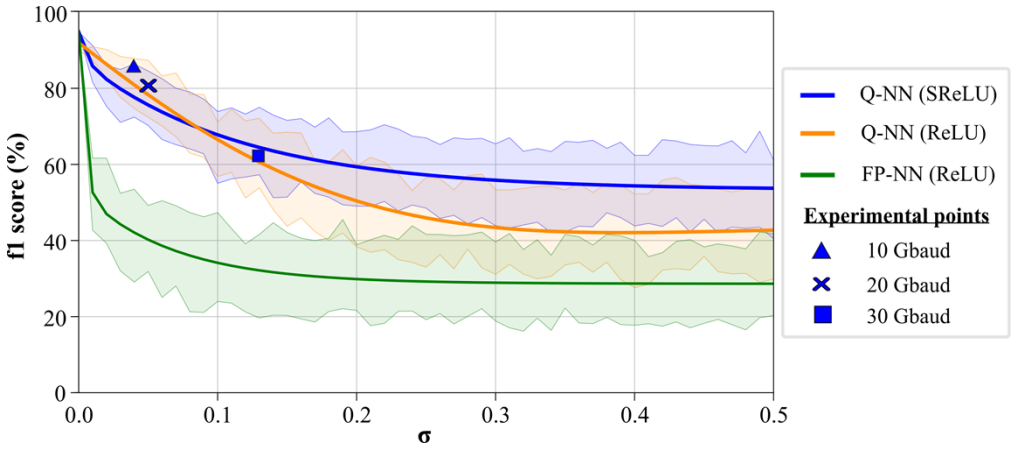

Fig. 5. Simulation results with the presence of noise in the products, for an NN configuration with floating point values and ReLU activation (green), with quantized values and ReLU activation (orange) and with quantized values and stretched slope ReLU activation. The shadowed areas indicate the error window of each curve that obtained after 100 iterations for each σ value. In the graph the experimental points measured from the product waveforms are also presented for 10, 20, and 30 Gbaud NN.

4. Conclusions

This paper presents an experimental demonstration of the photonic segment of the PEMAN architecture, utilizing a silicon photonic chip with high-speed electro-absorption modulators for matrix-vector multiplications. The evaluation, conducted on the photonic integrated circuit, where a three-layer NN for health monitoring was implemented, yielded promising results. Specifically, with a NN layout comprising 1350 trainable parameters, aimed at classifying heartbeat sounds, the network achieved notable F1-scores of 85.9% and 81% at compute rates of 10 and 20 Gbaud, respectively, while the performance at 30 Gbaud was degraded due to limited bandwidth of the hardware setup. Additionally, strategies such as quantization of the input and weight values and ReLU function slope stretching were implemented to address noise impairments stemming from the hardware constraints. Simulation analysis underscored the pivotal role of the quantization strategy in maintaining network performance amidst additive noise during NN inference. These findings highlight the potential of leveraging photonic architectures for efficient and robust neural network implementations.

Funding. HORIZON EUROPE Digital, Industry and Space (101092766).

Acknowledgments. This work was supported by the European Commission through the HORIZON projects SIPHO-G (101017194) and ALLEGRO (101092766) and partially supported by the Italian Ministry of Foreign Affairs and International Cooperation, grant number IN22GR06.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, *et al.*, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**(2), 102–114 (2021).
2. M. A. Nahmias, T. F. de Lima, A. N. Tait, *et al.*, "Photonic Multiply-Accumulate Operations for Neural Networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–18 (2020).
3. A. Tsakyridis, M. Moralis-Pegios, G. Giamougiannis, *et al.*, "Photonic neural networks and optics-informed deep learning fundamentals," APL Photonics **9**(1), 011102 (2024).

Optics EXPRESS

4. G. Dabos, D. V Bellas, R. Stabile, *et al.*, "Neuromorphic photonic technologies and architectures: scaling opportunities and performance frontiers Invited," Opt. Mater. Express **12**(6), 2343–2367 (2022).
5. G. Giamougiannis, A. Tsakyridis, M. Moralis-Pegios, *et al.*, "Universal Linear Optics Revisited: New Perspectives for Neuromorphic Computing With Silicon Photonics," IEEE J. Sel. Top. Quantum Electron. **29**(2: Optical Computing), 1–16 (2023).
6. A. Tsakyridis, G. Giamougiannis, M. Moralis-Pegios, *et al.*, "Universal Linear Optics for Ultra-Fast Neuromorphic Silicon Photonics Towards Fj/MAC and TMAC/sec/mm^2 Engines," IEEE J. Sel. Top. Quantum Electron. **28**, 1–15 (2022).
7. N. Farmakidis, B. Dong, and H. Bhaskaran, "Integrated photonic neuromorphic computing: opportunities and challenges," Nat. Rev. Electr. Eng. **1**(6), 358–373 (2024).
8. N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication," IEEE J. Sel. Top. Quantum Electron. **29**, 1–11 (2023).
9. W. Zhou, B. Dong, N. Farmakidis, *et al.*, "In-memory photonic dot-product engine with electrically programmable weight banks," Nat. Commun. **14**(1), 2887 (2023).
10. F. Shokraneh, S. Geoffroy-Gagnon, M. S. Nezami, *et al.*, "A Single Layer Neural Network Implemented by a 4×4 MZI-Based Optical Processor," IEEE Photonics J. **11**(6), 1–12 (2019).
11. C. Huang, S. Fujisawa, T. F. de Lima, *et al.*, "A silicon photonic–electronic neural network for fibre nonlinearity compensation," Nat. Electron. **4**(11), 837–844 (2021).
12. S. Bandyopadhyay, A. Sludds, S. Krastanov, *et al.*, "Single chip photonic deep neural network with accelerated training," (2022).
13. J. Feldmann, N. Youngblood, M. Karpov, *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**(7840), 52–58 (2021).
14. A. N. Tait, T. F. de Lima, E. Zhou, *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**(1), 7430 (2017).
15. B. Shi, N. Calabretta, and R. Stabile, "Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–11 (2020).
16. Y. Shen, N. C. Harris, S. Skirlo, *et al.*, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).
17. H. Zhang, M. Gu, X. D. Jiang, *et al.*, "An optical neural chip for implementing complex-valued neural network," Nat. Commun. **12**(1), 457 (2021).
18. G. Mourgias-Alexandris, M. Moralis-Pegios, A. Tsakyridis, *et al.*, "Noise-resilient and high-speed deep learning with coherent silicon photonics," Nat. Commun. **13**(1), 5572 (2022).
19. F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," Nature **606**(7914), 501–506 (2022).
20. H. Zhou, J. Dong, J. Cheng, *et al.*, "Photonic matrix multiplication lights up photonic accelerator and beyond," Light: Sci. Appl. **11**(1), 30 (2022).
21. G. Giamougiannis, A. Tsakyridis, M. Moralis-Pegios, *et al.*, "Neuromorphic silicon photonics with 50 GHz tiled matrix multiplication for deep-learning applications," Adv. Photonics **5**(01), 016004 (2023).
22. G. Giamougiannis, A. Tsakyridis, M. Moralis-Pegios, *et al.*, "Analog nanophotonic computing going practical: silicon photonic deep learning engines for tiled optical matrix multiplication with dynamic precision," (2023).
23. I. Roumpos, L. De Marinis, M. Kirtas, *et al.*, "High-performance end-to-end deep learning IM/DD link using optics-informed neural networks," Opt. Express **31**(12), 20068–20079 (2023).
24. M. Moralis-Pegios, G. Mourgias-Alexandris, A. Tsakyridis, *et al.*, "Neuromorphic Silicon Photonics and Hardware-Aware Deep Learning for High-Speed Inference," J. Lightwave Technol. **40**(10), 3243–3254 (2022).
25. A. Totovic, G. Giamougiannis, A. Tsakyridis, *et al.*, "Programmable photonic neural networks combining WDM with coherent linear optics," Sci. Rep. **12**(1), 5605 (2022).
26. A. Tsakyridis, G. Giamougiannis, G. Mourgias-Alexandris, *et al.*, "Silicon Photonic Neuromorphic Computing with 16 GHz Input Data and Weight Update Line Rates," in *Conference on Lasers and Electro-Optics* (2022), pp. 1–2.
27. Y. Bai, X. Xu, M. Tan, *et al.*, "Photonic multiplexing techniques for neuromorphic computing," Nanophotonics **12**(5), 795–817 (2023).
28. R. Hamerly, L. Bernstein, A. Sludds, *et al.*, "Large-Scale Optical Neural Networks Based on Photoelectric Multiplication," Phys. Rev. X. **9**(2), 21032 (2019).
29. A. Sludds, S. Bandyopadhyay, Z. Chen, *et al.*, "Delocalized photonic deep learning on the internet's edge," Science **378**(6617), 270–276 (2022).
30. C. Pappas, T. Moschos, M. Moralis-Pegios, *et al.*, "A TeraFLOP Photonic Matrix Multiplier using Time-Space-Wavelength Multiplexed AWGR-based Architectures," in *Optical Fiber Communications Conference and Exhibition* (2024), pp. 1–3.
31. S. Kovaios, I. Roumpos, A. Tsakyridis, *et al.*, "Sub-pJ/MAC Silicon Photonic GeMM for Optical Neural Networks using a Time-Space Multiplexed Coherent Xbar," in *Optical Fiber Communication Conference* (Optica Publishing Group, 2024), paper M4C.3.
32. L. De Marinis, A. Catania, P. Castoldi, *et al.*, "A Codesigned Integrated Photonic Electronic Neuron," IEEE J. Quantum Electron. **58**(5), 1–10 (2022).

Optics EXPRESS

33. M. Kirtas, N. Passalis, A. Oikonomou, *et al.*, "Mixed-precision quantization-aware training for photonic neural networks," Neural Comput. Appl. **35**(29), 21361–21379 (2023).
34. G. Mourgias-Alexandris, M. Moralis-Pegios, A. Tsakyridis, *et al.*, "Channel response-aware photonic neural network accelerators for high-speed inference through bandwidth-limited optics," Opt. Express **30**(7), 10664–10671 (2022).
35. M. Pantouvaki, S. A. Srinivasan, Y. Ban, *et al.*, "Active Components for 50 Gb/s NRZ-OOK Optical Interconnects in a Silicon Photonics Platform," J. Lightwave Technol. **35**(4), 631–638 (2017).
36. E. Paolini, L. De Marinis, M. Cococcioni, *et al.*, "Photonic-aware neural networks," Neural Comput. Appl. **34**(18), 15589–15601 (2022).
37. Heartbeat sound, [online] Available: https://www.kaggle.com/datasets/abdallahaboelkhair/heartbeatsound?select=Heartbeat_Sound.
38. E. Paolini, L. De Marinis, G. Contestabile, *et al.*, "Validation of Photonic Neural Networks in Health Scenarios," in *Proc. International Conference on Photonics in Switching and Computing* (2023).
39. E. Paolini, L. De Marinis, L. Valcarenghi, *et al.*, "Activation Stretching for Tackling Noise in Photonic Aware Neural Networks," in *Optical Fiber Communication Conference* (Optica Publishing Group, 2024), paper Th2A.13.
40. S. Kovaios, I. Roumpos, M. Moralis-Pegios, *et al.*, "Scaling photonic neural networks: A silicon photonic GeMM leveraging a Time-Space multiplexed Xbar," J. Lightwave Technol. **1**, 1–9 (2024).
41. A. Nazemi, K. Hu, B. Catli, *et al.*, "3.4 A 36Gb/s PAM4 transmitter using an 8b 18GS/S DAC in 28 nm CMOS," in *International Solid-State Circuits Conference* (IEEE, 2015), pp. 1–3.
42. H. Ramon, J. Lambrecht, J. Verbist, *et al.*, "70 Gb/s 0.87 pJ/bit GeSi EAM Driver in 55 nm SiGe BiCMOS," in *European Conference on Optical Communication* (2018), pp. 1–3.
43. A. R. Totović, G. Dabos, N. Passalis, *et al.*, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," IEEE J. Sel. Top. Quantum Electron. **26**(5), 1–15 (2020).